Safe, Efficient, and Socially Compliant Automated Driving in Mixed Traffic
Sensing, Anomaly Detection, Planning and Control

Dong, Yongqi

DOI
10.4233/uuid:12a9aff5-0cb6-46ab-a2d1-ae927d564913

Publication date
2025

Document Version
Final published version

Citation (APA)
Dong, Y. (2025). *Safe, Efficient, and Socially Compliant Automated Driving in Mixed Traffic: Sensing, Anomaly Detection, Planning and Control*. [Dissertation (TU Delft), Delft University of Technology]. TRAIL Research School. https://doi.org/10.4233/uuid:12a9aff5-0cb6-46ab-a2d1-ae927d564913

Safe, Efficient, and Socially Compliant Automated Driving in Mixed Traffic

Sensing, Anomaly Detection, Planning and Control

Yongqi Dong

Safe, Efficient, and Socially Compliant Automated Driving in Mixed Traffic:

Sensing, Anomaly Detection, Planning and Control

Yongqi Dong

Delft University of Technology

Safe, Efficient, and Socially Compliant Automated Driving in Mixed Traffic: *Sensing, Anomaly Detection, Planning and Control*

Dissertation

for the purpose of obtaining the degree of doctor

at Delft University of Technology,

by the authority of the Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen,

Chair of the Board for Doctorates,

to be defended publicly on

Monday 12 May 2025 at 17:30 o'clock

by

Yongqi DONG

Master of Control Science and Engineering,

Tsinghua University, China,

born in Shangdong, China

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof.dr.ir. B. van Arem	Delft University of Technology, *promotor*
Dr.ir. H. Farah	Delft University of Technology, *promotor*
Independent members:	
Prof.dr.eng. M.A. Sotelo Vázquez	University of Alcalá, Spain
Prof.dr. S. Sacone	University of Genova, Italy
Prof.dr. M. Wang	Dresden University of Technology, Germany
Prof.dr. D. Gavrila	Delft University of Technology
Prof.dr.ir. S.P. Hoogendoorn	Delft University of Technology

This research is funded by Applied and Technical Sciences (TTW), a subdomain of the Dutch Institute for Scientific Research (NWO), through the Project Safe and Efficient Operation of Automated and Human-Driven Vehicles in Mixed Traffic (SAMEN) under contract 17187.

TRAIL Thesis Series no. T2025/6, the Netherlands Research School TRAIL

TRAIL
P.O. Box 5017
2600 GA Delft
The Netherlands
E-mail: info@rsTRAIL.nl

ISBN: 978-90-5584-361-9

Printed in the Netherlands

To all those whose paths have intersected with mine throughout my life journey.

Preface

Many years later, as I was about to get my final degree, memories flooded back – my mother's beaming smile as I earned full marks in elementary school, the silhouette of my father carrying me on his motorcycle for more than two hours to my high school, the miniature Four Wheel Drive Brother Model Car that filled my childhood, and the Christmas day flight to the Netherlands...

With my interdisciplinary education and research background, I am grateful that I delved into the domain of automated mobility domain and pursued a Ph.D. degree in this field.

My Ph.D. journey commenced amidst the backdrop of a global pandemic, necessitating nearly two years of remote work and study from home. Despite the initial challenges, I quickly adapted and established a productive rhythm for research and study within the familiar confines of my home environment. Yet, amidst this solitary pursuit, I remained deeply connected to the vibrant group of the Traffic and Transportation Safety (TTS) lab and the Transport and Planning Department at Delft University of Technology. As I draw close to the completion of my thesis, I find myself already yearning for the camaraderie and shared pursuit of knowledge that defined my time within these esteemed groups.

First and foremost, I would like to express my utmost gratitude to my promotors, Dr.ir. Haneen Farah and Prof.dr.ir. Bart van Arem. Their unwavering guidance, insightful suggestions, and continuous support have been invaluable throughout every stage of my academic journey. From shaping research proposals to conducting research and manuscript preparation, their expertise and constructive feedback have been instrumental in shaping this thesis. Moreover, their mentorship extended beyond academia, providing invaluable advice that contributed to my personal growth, resilience, and well-being. It has been an honour and a privilege to share this wonderful Ph.D. journey with such esteemed supervisors.

As my daily supervisor, Haneen has consistently demonstrated her availability and support, offering invaluable guidance and assistance during our regular meetings, typically on a weekly

basis. Her ability to break down complex problems and provide practical solutions has been instrumental in refining my research approach and enhancing the quality of my work. Additionally, Haneen generously shared her expertise in project supervision, collaboration strategies, and time management, which greatly benefited my professional development, equipping me with valuable skills and insights that extend beyond the scope of my research project.

As my promoter, Bart consistently offered guidance and valuable insights from a higher-level perspective, enriching my understanding of the research landscape and providing thoughtful suggestions for advancement during our joint monthly meetings. His expertise extended beyond the confines of our specific research domain, allowing him to offer valuable advice on various aspects of networking, academic writing, modelling techniques, and the selection of journals and conferences for paper submissions. Moreover, Bart generously shared a wealth of resources from both within and outside our department and university, further enhancing the depth and breadth of my research endeavours.

I am profoundly grateful for the proactive approach adopted by both my daily supervisor and promotor in tackling questions and challenges that extend beyond their areas of expertise. Their willingness and readiness to seek external assistance and supervision reflect their unwavering commitment to the success of my research endeavours. Additionally, their robust support for my five-month international research collaboration with the University of California, Berkeley (UC Berkeley) highlights their steadfast dedication to nurturing opportunities for my professional development and fostering collaborative ventures on a global scale.

I extend my heartfelt gratitude to my doctoral committee members – Prof.dr.eng. M.A. Sotelo Vázquez, Prof.dr. S. Sacone, Prof.dr. M. Wang, Prof.dr. D. Gavrila, and Prof.dr.ir. S.P. Hoogendoorn, for their diligent review of the thesis draft manuscript and their invaluable comments aimed at enhancing the quality of this thesis. Their expertise and constructive feedback have been instrumental in refining the content and ensuring its scholarly rigour.

My sincere appreciation goes to the *Safe and efficient operation of AutoMated and human drivEN vehicles in mixed traffic* (SAMEN) project group and user committee members for their invaluable contributions and collaborative spirit throughout this journey. A special thank you to Haneen for her exceptional leadership and dedication in spearheading this project. I am grateful for the stimulating interactions with the consortium partners, particularly Maarten Sierhuis (Nissan), Rik Nuyttens (3M), Maria Oskina (RHDHV), Harm-Jan Mostert (Province North Holland), Gerdien Klunder (TNO), Marco van Burgsteden (CROW), and Jennifer Faber (NWO). Your insights and expertise have greatly enriched my understanding and perspective on the subject matter. To my fellow SAMEN researchers Nagarjun Reddy, Yiyun Wang, Narayana Raju, and Wouter Schakel, as well as the master's and bachelor's students under my supervision, Shiva Nischal Lingam, Sandeep Patil, Lanxin Zhang, Mathijs den Otter, Eline van der Kooij, Yuteng Zhang, Henan Yuan, Tobias Datema, Vincent Wassenaar, Joris Van de Weg, Cahit Tolga Kopar, Harim Suleman, and Sanny Toonen, I am thankful for the valuable knowledge and experiences you have shared. Special thanks to Prof. Masayoshi Tomizuka, Dr. Wei Zhan, and Dr. Chen Tang at the Mechanical Systems Control Lab, UC Berkeley. Our collaborative efforts have been instrumental in advancing the collective understanding of automated driving systems. I will fondly remember our celebrations after user committee meetings and other gatherings – they will be missed dearly.

I am immensely grateful to my friendly and outstanding colleagues at the TTS lab, the Department of Transport and Planning, and the TRAIL research school. A special mention goes to my office mates, Sina, Siri, Samkie (Siqi), Johan, Vincent, Paul, Solmaz, Ivan, Omid, Willem-Jan, and Laxman, whose camaraderie and support have been invaluable throughout my Ph.D. journey. Johan, thanks for your assistance in proofreading and revising the thesis summary in Dutch. I also extend my gratitude to the social facilitators, Eilif, Konstantinos, Bing, Saman, and Lucas, whose contributions have enriched our social and entertainment activities. Additionally, I would like to thank Prof.dr. Bert van Wee, Prof.dr.ir. Joost de Winter, Prof.dr.ir. Hans Hellendoorn, Dr. Jan van Gemert, Saeed, Irene, Conchita, Xiaolin, Guopeng, Yiru, Xue, Callum, Weiming, Samir, Yufei, Kexin, Zili, and Edwin for their valuable insights, camaraderie, and technical support. The research collaborations, Ph.D. forums, TRAIL congresses, coffee breaks, and cultural exchanges have truly made my Ph.D. experience vibrant and fulfilling.

I am also indebted to all my teachers and mentors who have played pivotal roles in shaping my intellectual growth and personal development. Their guidance, wisdom, and encouragement, particularly during my formative years, have instilled in me the confidence and resilience needed to navigate the challenges of academia and beyond. I am profoundly grateful for their dedication to nurturing the minds of future generations and hope to pay their kindness forward in my own endeavours. I am also grateful to those who have engaged in thought-provoking discussions and debates with me, challenging me to grow and evolve. Their constructive criticism and differing perspectives have served as important reminders of my imperfections and limitations, pushing me to strive for continuous improvement and self-awareness.

Special thanks are extended to my Taichi masters, Yongsheng Zhang and Weimin Luo, for imparting invaluable lessons on finding inner peace and the practice of meditation. Their teachings and guidance have not only enriched my spiritual well-being but have also contributed to my overall physical health and resilience. I am deeply grateful for their mentorship and the profound impact on my life.

I express my deepest gratitude to my parents and my family for their unwavering support throughout my academic journey. Especially for my father, over the span of more than 15 years, I can still vividly recall the silhouette of him carrying me on his motorcycle for more than two hours, traversing towns to get me to my high school. My family's enduring encouragement and belief in me have been the driving force behind my pursuit of knowledge and academic success.

Lastly, I extend my heartfelt appreciation to all individuals who have crossed paths with me, supporting me in ways both visible and invisible throughout my academic and life journey. Your encouragement, assistance, and belief in my abilities have been instrumental in reaching this milestone. I am truly fortunate to have such a supportive network of family, friends, mentors, and colleagues. Thank you all for being part of my journey and for your invaluable contributions to my personal and academic growth.

I also thank myself for the countless overnight hard work, for never giving up, and for the unwavering commitment and determination to persevere on this challenging academic path.

Yongqi Dong

Aachen, December, 2024

Content

Summary

Background

The steady development of automated vehicles (AVs) promises significant benefits in terms of traffic safety and efficiency. However, the transition to full automation AVs and their deployment on the road will be gradual, leading to a phase of mixed-traffic conditions where AVs at various levels coexist with human-driven vehicles (HDVs). This transition poses unprecedented hurdles, requiring a deeper understanding of the emerging challenges for AVs in sensing and perceiving road environments, as well as in the novel interactions between AVs and HDVs. Furthermore, the social compliance of AVs and the optimisation of their deployment strategies need to be considered as well.

Contents of this thesis

This thesis addresses the multifaceted challenges associated with AVs' development and deployment in mixed-traffic environments. The main objective of this thesis is to enhance the capabilities of AVs, enabling them with a wider Operational Design Domain (ODD), and thus facilitate the implementation of safe, efficient, and socially compliant automated driving in mixed traffic. Referring to the modular design of AV systems, three key perspectives, i.e., sensing and perception, anomaly detection, as well as planning and control, are tackled in this thesis. To be specific:

Chapters 2-4 focus on enhancing sensing and perception capabilities through the development of hybrid spatial-temporal deep learning models and self-supervised pretraining methods. Lane detection is chosen as the focus of these chapters since it is vital for current vehicle localisation and positioning, and it is also the foundation of various automated driving features. The main findings of these chapters are summarised as follows.

Chapter 2 presents a pioneering hybrid spatial-temporal sequence-to-one deep learning architecture tailored for vision-based lane detection tasks. By integrating the spatial convolutional neural network (SCNN) with spatial-temporal Recurrent Neural Network (RNN) modules, this architecture effectively captures correlations and dependencies among continuous image frames. Through extensive experimentation on various driving scenes, including challenging scenarios, the proposed model variants exhibit superior performance over existing state-of-the-art models. Notably, even the lighter model variants demonstrate remarkable accuracy, outperforming their counterparts while maintaining lower computational complexity.

Building upon the foundation laid in Chapter 2, Chapter 3 focuses on refining vision-based sensing and perception through the development of customised spatial-temporal attention mechanisms. These mechanisms, including temporal attention, spatial-temporal attention, and spatial-temporal attention with fully connected layers, are meticulously designed to optimise the utilisation of spatial-temporal correlations across different regions of interest within the consecutive image frames. Leveraging linear Long Short Term Memory (LSTM) neural networks in conjunction with the proposed attention blocks, this chapter demonstrates the feasibility of lightweight and computationally efficient solutions for sequential deep neural networks (DNNs). Through rigorous experiments, ablation studies, and comparative analysis across diverse datasets, the effectiveness of the proposed attention mechanisms in enhancing lane detection performance is convincingly established.

In Chapter 4, the exploration of enhancing vision-based sensing and perception capabilities continues with the introduction of a self-supervised pretraining method employing masked sequential autoencoders (MSAE). This innovative approach leverages both labelled and unlabelled data to improve detection accuracy and expedite the training process of DNN models dedicated to lane detection tasks. Additionally, a customised Focal Loss based PolyLoss is introduced to further enhance the detection accuracy. Through comprehensive experimentation and comparative analysis, the efficacy of the proposed pretraining method and loss function is demonstrated, showcasing substantial improvements in lane detection performance across diverse driving scenarios. Specifically, the utilisation of MSAE-based pretraining and the adoption of the customised PolyLoss result in superior performance metrics, underscoring the pivotal role of self-supervised learning techniques and tailored loss functions in fortifying the robustness and efficiency of vision-based sensing and perception systems in AVs.

These chapters address the challenges of vision-based lane detection, crucial for AV navigation and safety.

Chapters 5-6 delve into anomaly detection, investigating techniques for identifying abnormal lane rendering in digital map applications and detecting anomalies in driving behaviour.

Chapter 5 introduces an innovative approach to anomaly detection in lane rendering images of digital map applications, utilising Transformer-based models with self-supervised pretraining and customised fine-tuning. By transforming anomaly detection into a classification problem, the chapter proposes a four-phase pipeline that includes data pre-processing, self-supervised pre-training with masked image modelling (MiM), customised fine-tuning using cross-entropy-based loss, and post-processing. Experimental results demonstrate the pipeline's effectiveness, with significant improvements in detection accuracy and reduced training time achieved

through self-supervised pre-training. Ablation studies regarding tackling the problem with different numbers of classes further validate the pipeline's performance enhancements, particularly in addressing data imbalance. This approach not only enhances anomaly detection accuracy but also contributes to reducing labour costs associated with manual labelling and anomaly detection efforts, offering significant societal benefits.

Additionally, Chapter 6 explores the critical task of detecting abnormal driving behaviour, addressing the need for more feasible and efficient approaches by leveraging semi-supervised ML methods. Utilising large-scale real-world driving data, the study develops a semi-supervised ML model based on the Hierarchical Extreme Learning Machine (HELM). This approach utilises partly labelled data and introduces Surrogate Measures of Safety (SMoS) (specifically the event-based safety indicators of Two-Dimensional Time-To-Collision (2D-TTC)) as the pivotal input features to enhance performance. Results demonstrate the effectiveness of the proposed semi-supervised ML model, showcasing superior performance compared to baseline methods. The integration of SMoS significantly improves detection accuracy, highlighting its significant role in enhancing model performance. By leveraging unlabelled data for training and only a small sample of labelled data for fine-tuning, the proposed semi-supervised approach achieves competitive performance while reducing dependency on fully labelled datasets, making it suitable for real-world applications.

To sum up, the exploration of semi-supervised and self-supervised ML methods presents promising avenues in anomaly detection. The pioneering research presented in this thesis represents a significant stride towards leveraging data-driven ML-based anomaly detection methodologies to enhance the safety of driving.

Chapters 7-9 shift the focus to planning and control strategies for AVs, presenting a comprehensive examination of decision-making frameworks and control algorithms. These chapters introduce a conceptual framework aimed at fostering socially compliant driving behaviour and propose a range of model-based and learning-based approaches.

Chapter 7 lays the groundwork by introducing a conceptual framework that emphasises socially compliant automated driving. This framework encompasses various social components such as cultural nuances, norms, and driving styles. A key innovation is the introduction of bidirectional behavioural adaptation, highlighting the dynamic interactions between AVs and human drivers. Furthermore, the framework advocates for the incorporation of a spatial-temporal memory module to enable continuous refinement of driving strategies, thereby promoting adaptability and safety in diverse traffic scenarios. Validation through an online expert survey lends credence to the framework's efficacy. This conceptual framework lays a solid foundation for learning-based and model-based approaches for implementing planning and control algorithms for automated driving.

In the learning-based approach explored in Chapter 8, Deep Reinforcement Learning (DRL) takes centre stage, with a focus on integrating safety, efficiency, comfort level, and energy consumption considerations into the learning framework. Multiple DRL algorithms are evaluated across diverse driving manoeuvres, particularly roundabout driving, highlighting the importance of real-world requirements in reward function design and simulation-based training. Among the compared DRL algorithms, Trust Region Policy Optimisation (TRPO) emerges as

leading in safety and efficiency, while Proximal Policy Optimisation (PPO) excels in comfort during roundabout driving. Moreover, the extension of the training environment to encompass various driving scenarios showcases the adaptability of DRL models to train a uniform driving model for real traffic environments, signalling promising avenues for future research.

Regarding the model-based approach, Chapter 9 introduces the DRF-SVO-MPCC algorithm, aimed at enhancing AVs' understandability and predictability to human drivers, particularly during interactions with HDVs when driving through the roundabouts, as this challenging manoeuvre involves large curvature and tackles both longitudinal and lateral control. This algorithm integrates the perceived Driving Risk Field (DRF), Social Value Orientation (SVO), and Model Predictive Contouring Control (MPCC), enabling AVs to navigate social scenarios with sensitivity to the benefits of surrounding HDVs. Simulation experiments, conducted on various roundabout scenarios, underscore the algorithm's superiority in trajectory tracking and adaptability to different driving styles, ensuring safety and social compliance. The findings illuminate the potential of the DRF-SVO-MPCC algorithm in fostering harmonious interactions between AVs and HDVs, setting a precedent for socially aware automated driving systems.

Overall, this thesis represents a solid endeavour to advance the planning and control capabilities of AVs in mixed-traffic environments. Through the development of novel conceptual frameworks and innovative model-based and learning-based algorithmic solutions, it lays the groundwork for the realisation of safe, efficient, socially compliant, and adaptable automated driving, contributing to safer and more harmonious transportation systems.

Conclusion and perspectives

In summary, this thesis contributes to advancing the knowledge of how to improve automated driving systems in the realms of sensing and perception, anomaly detection, as well as planning and control. By integrating theoretical frameworks, methodological innovations, and data-driven empirical evaluations, notable progress has been achieved in fostering the development of safe, efficient, and socially compliant automated driving within mixed-traffic environments.

Despite the considerable progress made, several directions for future research have been identified. These include the imperative for more expansive high-quality datasets, exploration of domain adaptation techniques for both sensing and anomaly detection tasks, as well as the seamless integration of model-based and learning-based methodologies for planning and control. Additionally, transitioning towards a unified driving model and effectively addressing the complexities of multi-agent interactions in intricate urban settings remain pivotal areas for further exploration. Furthermore, interdisciplinary collaboration will be instrumental in harnessing the full potential of automated vehicles to revolutionise transportation systems.

Samenvatting

Achtergrond

De gestage ontwikkeling van geautomatiseerde voertuigen (AV's) belooft significante voordelen op het gebied van verkeersveiligheid en efficiëntie. Echter, de overgang naar full automation AV's en hun inzet op de weg zal geleidelijk verlopen, wat zal leiden tot een fase van gemengde verkeersomstandigheden waarin AV's op verschillende niveaus samen rijden met door mensen bestuurde voertuigen (HDV's). Deze overgang brengt ongekende uitdagingen met zich mee, waarvoor een dieper begrip nodig is van de opkomende uitdagingen voor AV's bij het waarnemen en begrijpen van wegomgevingen, evenals bij de nieuwe interacties tussen AV's en HDV's. Bovendien moeten ook de sociale gedragsconfirmatie van AV's en de optimalisatie van hun inzetstrategieën worden overwogen.

Inhoud van dit proefschrift

Dit proefschrift behandelt de veelzijdige uitdagingen die gepaard gaan met de ontwikkeling en implementatie van AV's in gemengde verkeersomgevingen. Het hoofddoel van dit proefschrift is om de mogelijkheden van AV's te verbeteren door hen te voorzien van een breder Operationeel Ontwerp Domein (ODD) en zo de implementatie van veilig, efficiënt en sociaal aanvaardbaar geautomatiseerd rijden (AD) in gemengd verkeer mogelijk te maken. Met betrekking tot het modulaire ontwerp van AV-systemen worden drie belangrijke perspectieven behandeld in deze scriptie, namelijk waarneming en perceptie, anomaliedetectie, evenals planning en controle. Om specifiek te zijn:

Hoofdstukken 2-4 richten zich op het verbeteren van de mogelijkheden voor waarneming en perceptie door de ontwikkeling van hybride ruimtelijk-temporele deep-learning modellen en self-supervised pretraining technieken. Rijstrookdetectie wordt gekozen als de focus vanwege haar belang voor de huidige voertuiglokalisatie en -positionering, en het is ook de basis van

verschillende geautomatiseerde rijfuncties. De belangrijkste bevindingen van de hoofdstukken worden hieronder samengevat.

Hoofdstuk 2 introduceert een baanbrekend hybride ruimtelijk-temporele sequentie-naar-een deep learning algoritme dat is afgestemd op visuele detectietaken van rijstroken. Door de ruimtelijke convolutie-neurale netwerk (SCNN) te integreren met ruimtelijk-temporele recurrente neurale netwerk (RNN)-modules, vangt deze architectuur effectief correlaties en afhankelijkheden tussen continue beeldframes op. Door uitgebreide experimenten op verschillende rijscenario's, inclusief uitdagende scenario's, vertonen de voorgestelde modelvarianten superieure prestaties ten opzichte van bestaande state-of-the-art modellen. Met name de lichtere modelvarianten tonen opmerkelijke nauwkeurigheid, waarmee ze hun tegenhangers overtreffen terwijl ze een lagere computationele complexiteit behouden.

Verder bouwend op de basis gelegd in hoofdstuk 2, richt hoofdstuk 3 zich op het verfijnen van visuele waarneming en perceptie door de ontwikkeling van aangepaste ruimtelijk-temporele aandachtsmechanismen. Deze mechanismen, waaronder tijdelijke aandacht, ruimtelijk-temporele aandacht en ruimtelijk-temporele aandacht met volledig verbonden lagen, zijn zorgvuldig ontworpen om het gebruik van ruimtelijk-temporele correlaties over verschillende interessegebieden binnen de opeenvolgende beeldframes te optimaliseren. Door lineaire Long Short Term Memory (LSTM) neurale netwerken te gebruiken in combinatie met de voorgestelde aandachtsblokken, demonstreert het hoofdstuk de haalbaarheid van eenvoudige en rekenkundig efficiënte oplossingen voor sequentiële diepe neurale netwerken (DNN's). Door rigoureuze experimenten, ablatiestudies en vergelijkende analyses over verschillende datasets, wordt overtuigend de effectiviteit van de voorgestelde aandachtsmechanismen bij het verbeteren van de prestaties van rijstrookdetectie vastgesteld.

In Hoofdstuk 4 wordt het onderzoek naar het verbeteren van de mogelijkheden voor visuele waarneming en perceptie voortgezet met de introductie van een self-supervised pretraining techniek met behulp van gemaskeerde sequentiële auto-encoders (MSAE). Deze innovatieve aanpak maakt gebruik van zowel gelabelde als ongelabelde gegevens om de detectienauwkeurigheid te verbeteren en het trainingsproces van DNN-modellen voor rijstrookdetectietaken te versnellen. Daarnaast wordt een aangepaste Focal Loss op basis van PolyLoss geïntroduceerd om de detectienauwkeurigheid verder te verbeteren. Door uitgebreide experimenten en vergelijkende analyses wordt de doeltreffendheid van de self-supervised pretraining methode en verliesfunctie aangetoond, waarbij aanzienlijke verbeteringen in de prestaties van rijstrookdetectie over verschillende rijscenario's worden getoond. Specifiek, het gebruik van MSAE-gebaseerde pretraining en de adoptie van de aangepaste PolyLoss resulteren in superieure prestatiemetingen, waarbij de cruciale rol van zelf-begeleide leertechnieken en op maat gemaakte verliesfuncties in het versterken van de robuustheid en efficiëntie van visuele waarneming en perceptiesystemen in AV's wordt benadrukt.

Deze hoofdstukken behandelen de uitdagingen van visuele rijstrookdetectie, die cruciaal zijn voor AV-navigatie en veiligheid.

Hoofdstukken 5-6 richten zich op anomaliedetectie, waarbij technieken worden onderzocht voor het identificeren van abnormale rijstrookweergave in digitale kaarttoepassingen en het detecteren van afwijkingen in rijgedrag.

Hoofdstuk 5 introduceert een innovatieve benadering voor anomaliedetectie in afbeeldingen van rijstrookweergave in digitale kaarttoepassingen, waarbij Transformer-gebaseerde modellen worden gebruikt met self-begeleid pretraining en aangepaste fine-tuning. Door anomaliedetectie om te zetten in een classificatieprobleem, stelt het hoofdstuk een vierfasen-pipeline voor die datavoorbewerking, self-supervised pretraining met masked image modelling (MiM), aangepaste fine-tuning met behulp van cross-entropy gebaseerd verlies, en postprocessing omvat. Experimentele resultaten tonen de effectiviteit van de pipeline aan, met aanzienlijke verbeteringen in detectienauwkeurigheid en verminderde trainingsduur door self-supervised pretraining. Ablatiestudies met betrekking tot de aanpak van het probleem met verschillende aantallen klassen, valideren verder de prestatieverbeteringen van de pipeline, met name bij het aanpakken van onevenwichtigheden in de data. Deze benadering verbetert niet alleen de nauwkeurigheid van anomaliedetectie, maar draagt ook bij aan het verminderen van de arbeidskosten die gepaard gaan met handmatige labeling en anomaliedetectie-inspanningen, wat aanzienlijke maatschappelijke voordelen biedt.

Daarnaast onderzoekt hoofdstuk 6 de cruciale taak van het detecteren van abnormaal rijgedrag, waarbij wordt ingegaan op de behoefte aan meer haalbare en efficiënte benaderingen door gebruik te maken van semi-begeleide machine learning (ML)-methoden. Door gebruik te maken van grootschalige, gegevens uit echte rijomstandigheden, ontwikkelt de studie een semi-begeleid ML-model op basis van Hierarchical Extreme Learning Machine (HELM). Deze aanpak maakt gebruik van gedeeltelijk gelabelde gegevens en introduceert Surrogate Measures of Safety (SMoS), specifiek de gebeurtenis-gebaseerde veiligheidindicatoren van Two-Dimensional Time-To-Collision (2D-TTC), als de belangrijkste invoerfuncties om de prestaties te verbeteren. Resultaten tonen de effectiviteit van het voorgestelde semi-begeleide ML-model, waarbij superieure prestaties worden getoond in vergelijking met basismethoden. De integratie van SMoS verbetert de detectienauwkeurigheid aanzienlijk, waarbij zijn belangrijke rol bij het verbeteren van de modelprestaties wordt benadrukt. Door gebruik te maken van ongelabelde gegevens voor training en slechts een kleine steekproef gelabelde gegevens voor fine-tuning, bereikt de voorgestelde semi-begeleide aanpak concurrerende prestaties terwijl de afhankelijkheid van volledig gelabelde datasets wordt verminderd, waardoor het geschikt is voor real-world toepassingen.

Samengevat biedt de verkenning van semi-begeleide en zelf-begeleide ML-methoden veelbelovende mogelijkheden in anomaliedetectie. Het baanbrekende onderzoek gepresenteerd in dit proefschrift vertegenwoordigt een significante stap voorwaarts in het benutten van op data gedreven, ML-gebaseerde anomaliedetectiemethoden om de veiligheid van rijgedrag te verbeteren.

Hoofdstukken 7-9 verleggen de focus naar planning en besturingsstrategieën voor AV's, waarbij een uitgebreid onderzoek wordt gepresenteerd naar besluitvormingskaders en besturingsalgoritmen. Deze hoofdstukken introduceren een conceptueel kader dat is gericht op het bevorderen van sociaal conform rijgedrag en stellen een reeks op model en leren gebaseerde benaderingen voor.

Hoofdstuk 7 legt het fundament door een conceptueel kader te introduceren dat het belang van sociaal conform geautomatiseerde rijden benadrukt. Dit kader omvat verschillende sociale componenten zoals culturele nuances, normen en rijstijlen. Een belangrijke innovatie is de

introductie van bidirectionele gedragsaanpassing, waarbij de dynamische interacties tussen AV's en menselijke bestuurders worden benadrukt. Bovendien pleit het kader voor de integratie van een ruimtelijk-temporale geheugenmodule om de continue verfijning van rijstrategieën mogelijk te maken, waardoor aanpasbaarheid en veiligheid in diverse verkeersscenario's worden bevorderd. Validatie via een online enquête onder experts zorgt voor geloofwaardigheid van de effectiviteit van het kader. Dit conceptuele kader legt een solide basis voor op model en leren gebaseerde benaderingen.

In de op leren gebaseerde aanpak die wordt verkend in hoofdstuk 8, staat Deep Reinforcement Learning (DRL) centraal, met de nadruk op het integreren van veiligheid, efficiëntie, comfortniveau en energieverbruik in het leerkader. Meerdere DRL-algoritmen worden geëvalueerd voor diverse rijmanoeuvres, met name rotonde rijden, waarbij de nadruk wordt gelegd op de vereisten van de echte wereld in ontwerp van beloningsfunctie en simulatiegebaseerde training. Onder de vergeleken DRL-algoritmen komt Trust Region Policy Optimisation (TRPO) naar voren als een koploper in veiligheid en efficiëntie, terwijl Proximal Policy Optimisation (PPO) uitblinkt in comfort tijdens rotonde rijden. Bovendien laat de uitbreiding van de trainingsomgeving om verschillende rijscenario's te omvatten de aanpasbaarheid van DRL-modellen zien om een uniform rijmodel te trainen voor echte verkeersomgevingen, waarbij veelbelovende richtingen worden aangegeven voor toekomstig onderzoek.

Met betrekking tot de op model gebaseerde benadering introduceert hoofdstuk 9 het DRF-SVO-MPCC-algoritme, gericht op het verbeteren van de begrijpelijkheid en voorspelbaarheid van AV's voor menselijke bestuurders, met name tijdens interacties met HDV's bij het rijden over rotondes. Dit algoritme integreert het waargenomen rijrisicoveld (DRF), sociale waarderingsoriëntatie (SVO) en modelpredictieve contourbesturingscontrole (MPCC), waardoor AV's sociale scenario's kunnen navigeren met gevoeligheid voor het welzijn van omliggende HDV's. Simulatie-experimenten, uitgevoerd op verschillende rotonde-scenario's, onderstrepen de superioriteit van het algoritme in trajectvolgen en aanpasbaarheid aan verschillende rijstijlen, waarbij veiligheid en sociale conformiteit worden gegarandeerd. De bevindingen belichten het potentieel van het DRF-SVO-MPCC-algoritme om harmonieuze interacties tussen AV's en HDV's te bevorderen, waarbij een precedent wordt geschapen voor sociaal bewuste geautomatiseerde rijsystemen.

Al met al vertegenwoordigt dit proefschrift een uitgebreide inspanning om de planning- en controlecapaciteiten van AV's in gemengde verkeersomgevingen te bevorderen. Door de ontwikkeling van nieuwe conceptuele kaders en innovatieve op model en leren gebaseerde algoritmische oplossingen legt het de basis voor het realiseren van veilig, efficiënt, sociaal aanvaardbaar en aanpasbaar geautomatiseerd rijden, wat bijdraagt aan veiligere en harmonieuzere transportsystemen.

Conclusie en perspectieven

Samengevat heeft deze scriptie aanzienlijke vooruitgang geboekt in geautomatiseerde rijsystemen op het gebied van waarneming en perceptie, anomaliedetectie, evenals planning en controle. Door theoretische kaders, methodologische innovaties en op gegevens gebaseerde empirische evaluaties te integreren, is er belangrijke vooruitgang geboekt bij het bevorderen

van de ontwikkeling van veilig, efficiënt en sociaal aanvaardbaar geautomatiseerd rijden binnen gemengde verkeersomgevingen.

Ondanks de aanzienlijke vooruitgang zijn verschillende richtingen voor toekomstig onderzoek geïdentificeerd. Deze omvatten de noodzaak van meer uitgebreide datasets van hoge kwaliteit, verkenning van technieken voor domeinaanpassing voor zowel waarneming als anomaliedetectietaken, evenals de naadloze integratie van op model en leren gebaseerde methodologieën voor planning en controle. Bovendien blijft de overgang naar een uniform rijmodel en het effectief aanpakken van de complexiteiten van multi-agentinteracties in ingewikkelde stedelijke omgevingen cruciale gebieden voor verder onderzoek. Bovendien zal interdisciplinaire samenwerking instrumenteel zijn bij het benutten van het volledige potentieel van geautomatiseerde voertuigen om transportsystemen te revolutioneren.

1 Introduction

Fully automated vehicles (AVs) are expected to be beneficial to traffic safety and efficiency (Talebpour & Mahmassani, 2016; Yaqoob et al., 2020). Although steady development of high levels of AVs is witnessed, their deployment will not occur instantaneously. A transition period will ensue, during which AVs with various automation levels will co-exist and share the road with non-connected and non-automated road users, e.g., human-driven vehicles (HDVs), leading to mixed-traffic conditions.

Consequently, this new reality of mixed traffic will lead to unprecedented road and traffic conditions, accompanied by novel types of interactions among vehicles at different levels of automation, which could have significant implications for both traffic safety and efficiency (Fagnant & Kockelman, 2015; Fraedrich et al., 2015). Therefore, it is imperative to enhance our understanding of how AVs can be programmed to effectively sense their environment and respond appropriately with predictable behaviour across diverse driving contexts, particularly in challenging scenarios. In addition, addressing the challenge of making AVs socially compliant, i.e., understood and accepted by HDVs, together with optimising their deployment while considering economic, environmental, and societal impacts, are critical knowledge gaps that necessitate interdisciplinary research and collaboration. Addressing these challenges will yield valuable insights into the design and enhancement of both physical and digital road infrastructure, the development of AVs' sensing, perception, and driving control algorithms, as well as the deployment of AVs within mixed-traffic contexts. It will also facilitate effective simulation and analysis of the impacts on traffic safety and efficiency resulting from the deployment of AVs. These insights are then essential for reliable traffic operation and management in mixed-traffic settings. Additionally, they play a critical role in formulating policies that govern the seamless incorporation of AVs into existing transportation systems. By systematically addressing these critical gaps, we can pave the way for the successful integration of AVs into mixed-traffic environments, ultimately contributing to safer, more efficient, and sustainable transportation systems.

Generally, the development of automated vehicles comprises modules of *sensing and perception*, *localisation and mapping*, *decision-making and planning*, *control and action*, as well as *safety, redundancy, and anomaly monitoring*. A typical abstracted architecture of the modular design is depicted in **Figure 1-1**. Sensing and perception typically entail the utilisation of multiple sensors, such as cameras, LiDAR, and radar, for object detection, tracking, and monitoring the vehicle's state. Sensor fusion techniques are commonly employed to integrate data and sensing outputs from various sensors, thereby enhancing perception accuracy. The *sensing and perception* module is interconnected with and contributes to the *localisation and mapping* module. Localisation determines the vehicle's precise position and orientation relative to its surroundings, while mapping creates and updates detailed maps of the environment for navigation and planning purposes. Both the *sensing and perception* module and the *localisation and mapping* module are linked to the *decision-making and planning* module, which encompasses path planning, behaviour planning, and decision-making. Path planning generates optimal trajectories considering vehicle dynamics, traffic rules, and environmental constraints. Behaviour planning determines the vehicle's behaviour and actions based on the surrounding context and traffic conditions. Decision-making involves making real-time decisions such as lane changes, overtaking, intersection navigation, and obstacle avoidance. The *control and action* module executes vehicle dynamics control using actuators for acceleration, braking, steering, and other vehicle operations. This module aims to maintain desired vehicle parameters (e.g., speed and heading) and ensure the stability of the vehicle's motion. Finally, the *safety, redundancy, and anomaly monitoring* module oversees all the aforementioned modules in the automated vehicle system. Its functions include implementing fail-safe mechanisms for system integrity, continuously monitoring vehicle state to detect and mitigate anomalies and failures, and enabling emergency response actions such as braking and steering in critical situations. Note that the *safety, redundancy, and anomaly monitoring* module is a customised component not commonly employed in existing module designs for AVs. It is tailored specifically for the research tasks outlined in this thesis, which will be elaborated upon in subsequent sections.

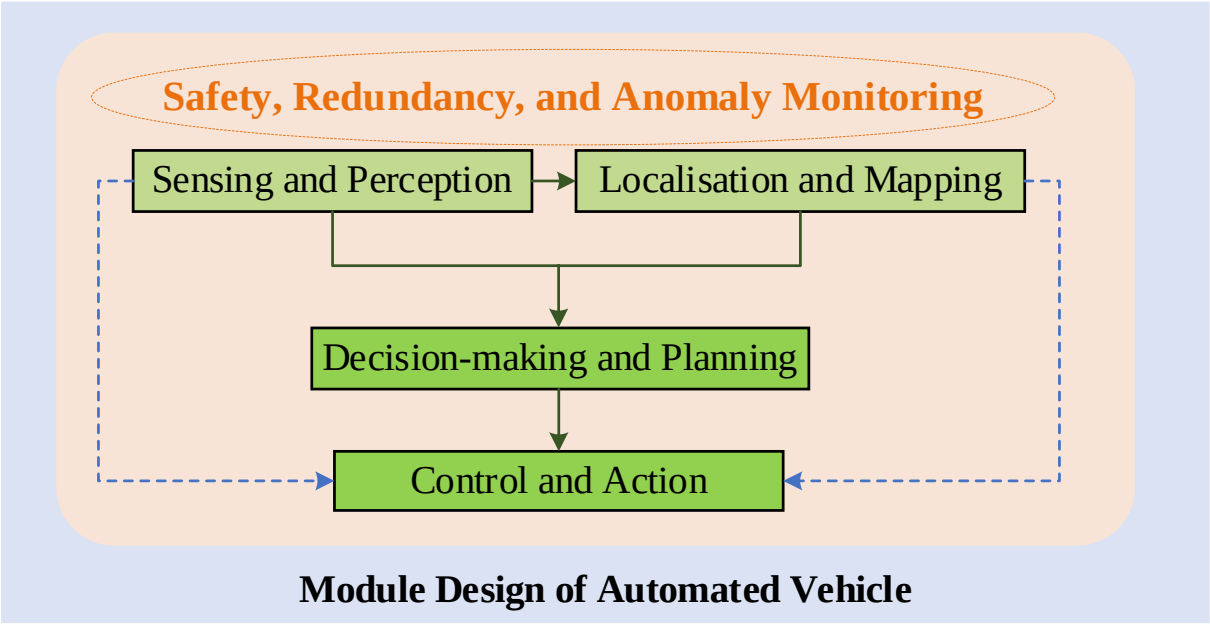


Figure 1-1. A system architecture abstraction of automated vehicle's module design

Corresponding to the aforementioned challenging knowledge gaps, the primary objective of this thesis is to enhance the capabilities of AVs, enabling them with a wider Operational Design Domain (ODD), and thus facilitate the implementation of safe, efficient, and socially compliant automated driving (AD) in mixed-traffic environments. The ODD delineates the specific conditions under which an automated driving system is supposed to function properly (SAE International, 2021a). Aligned with the customised module design illustrated in **Figure 1-1**, this thesis endeavours to consider all the modules and tackle the multifaceted challenges by adopting a comprehensive approach focusing on three fundamental pillars: sensing and perception, anomaly detection, together with planning and control. Sensing and perception pertain to how AVs interact with road infrastructure (e.g., lane-marking and obstacle detection and tracking), whereas planning and control address the interactions between AVs and other vehicles on the road (e.g., navigating through a roundabout with surrounding HDVs). Anomaly detection, as a crucial component, typically targets identifying edge cases, system anomalies, and abnormal driving behaviours.

To be specific, for sensing and perception, this thesis focuses on vision-based lane detection since it is vital for current vehicle localisation positioning itself within the lanes, and it is also the foundation of Advanced Driver Assistance Systems (ADAS), such as., Lane Keeping Assistance and Lane Departure Warning systems (Andrade et al., 2019; Bar Hillel et al., 2014; W. Chen et al., 2020; Liang et al., 2020; Xing et al., 2018). For planning and control, this thesis proposes a conceptual framework emphasising socially compliant decision-making and develops both model-based and learning-based approaches. Driving through roundabouts is selected as the primary focus due to its inherent challenges, including navigating large curvature, interacting with multiple participants, and addressing both longitudinal and lateral control aspects. For anomaly detection, driven and influenced by the availability of data, this thesis concentrates on two types of anomalies, i.e., (1) detection of abnormal lane rendering images within navigation map apps and (2) identification of abnormal driving behaviour in naturalistic driving scenarios. For (1), abnormal lane-level rendered map background images in digital map applications can lead to ambiguity in human drivers' understanding and adversely influence their decision-making when using navigation services, potentially resulting in critical unsafe situations. Therefore, it is crucial to accurately detect abnormal lane rendering map images to mitigate such safety risks. For (2), abnormal driving behaviour brings great uncertainty to traffic and may lead to accidents, posing danger to both the driver and the public (Jia et al., 2020; Sar et al., 2023). Accurate identification and detection of abnormal driving are vital to alert surrounding vehicles and ensure traffic safety. Furthermore, detecting and removing abnormal driving behaviour from naturalistic driving data would be a prerequisite step for training a human-like driving model for AVs with imitation learning.

This thesis introduction is organised as follows. Firstly, the research background and scope are introduced in *Section 1.1*. Then, in *Section 1.2*, the research gaps are identified, and corresponding research objectives and research questions are raised. *Section 1.3* outlines the research approach and research methods employed in this thesis. Next, *Section 1.4* highlights the scientific and practical contributions. Finally, *Section 1.5* presents the outline of this thesis.

1.1 Research background and scope

According to the World Health Organization (2023), each year, road traffic accidents cause nearly 1.19 million fatalities, and millions more suffer serious injuries. Furthermore, these traffic accidents cost most nations around 3% of their Gross Domestic Product (GDP)[1] (Toroyan et al., 2013). AVs, which are designed and programmed to be capable of driving themselves or performing specific necessary functions (e.g., car following, lane keeping) without being controlled or monitored by an individual for at least part of a journey, are promised to increase road safety and efficiency (Greenblatt & Shaheen, 2015; Jamson et al., 2011; Talebpour & Mahmassani, 2016; Yaqoob et al., 2020). These vehicles are gradually being introduced and deployed into everyday life. However, the transition from human-driven to fully automated vehicles will not occur overnight. The Society of Automotive Engineers (SAE) delineates six levels of driving automation (SAE International, 2021a), ranging from *No Driving Automation (Level 0)* to *Full Driving Automation (Level 5)*, as illustrated in **Figure 1-2**. Among the six levels, Level 0 is without any automation; Level 1 and Level 2 are with partial automation, wherein drivers remain responsible for driving, even when assisted by automated features and the drivers' feet are off the pedals. At Level 1 and Level 2, drivers must continually supervise the automated support features, such as steering or brake/acceleration assistance. The distinction between Level 1 and Level 2 lies in the scope of support provided. At Level 1, only one aspect of control, either steering **or** brake/acceleration, can be supported, whereas at Level 2, both steering **and** brake/acceleration, encompassing longitudinal and lateral control, can be simultaneously supported. At Level 3, Level 4, and Level 5, the AD features are responsible for all the dynamic driving tasks when they are engaged, where the drivers are not driving even if they are seated in the driver's seat. However, differences exist among these three levels. At Level 3, known as conditional automation, the drivers are still required to intervene and take control when prompted by the AD features. In other words, drivers must be on standby to resume control when requested by the system. While at Level 4 and Level 5, the AD features will never make such requests. Additionally, for Level 4, the AD features can only drive the vehicle under specific conditions defined by the ODD. In contrast, Level 5 allows the AD features to operate the vehicle under all conditions. This thesis scope considers AVs across Levels 1 to 4 of automation.

Currently, the gradual deployment of different levels of AVs has resulted in a transitional phase characterised by mixed traffic, where vehicles with varying levels of automation co-exist and share the road with HDVs. This transition brings about unprecedented challenges for AVs in sensing and perceiving the road environment, as well as novel interactions between AVs and HDVs affecting the traffic conditions. These challenges and novel interactions may give rise to uncertainties and issues that affect both road safety and efficiency (Fagnant & Kockelman, 2015; Fraedrich et al., 2015). Moreover, the integration of AVs into existing traffic systems necessitates a thorough understanding of their operational parameters and capabilities.

These challenges and uncertainties underscore the importance of expanding the ODD of AVs. The ODD refers to the specific operating conditions in which the AD system is designed to function properly. **Figure 1-3** illustrates several examples of ODD relative to different driving

[1] Road traffic injuries, World Health Organization, https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

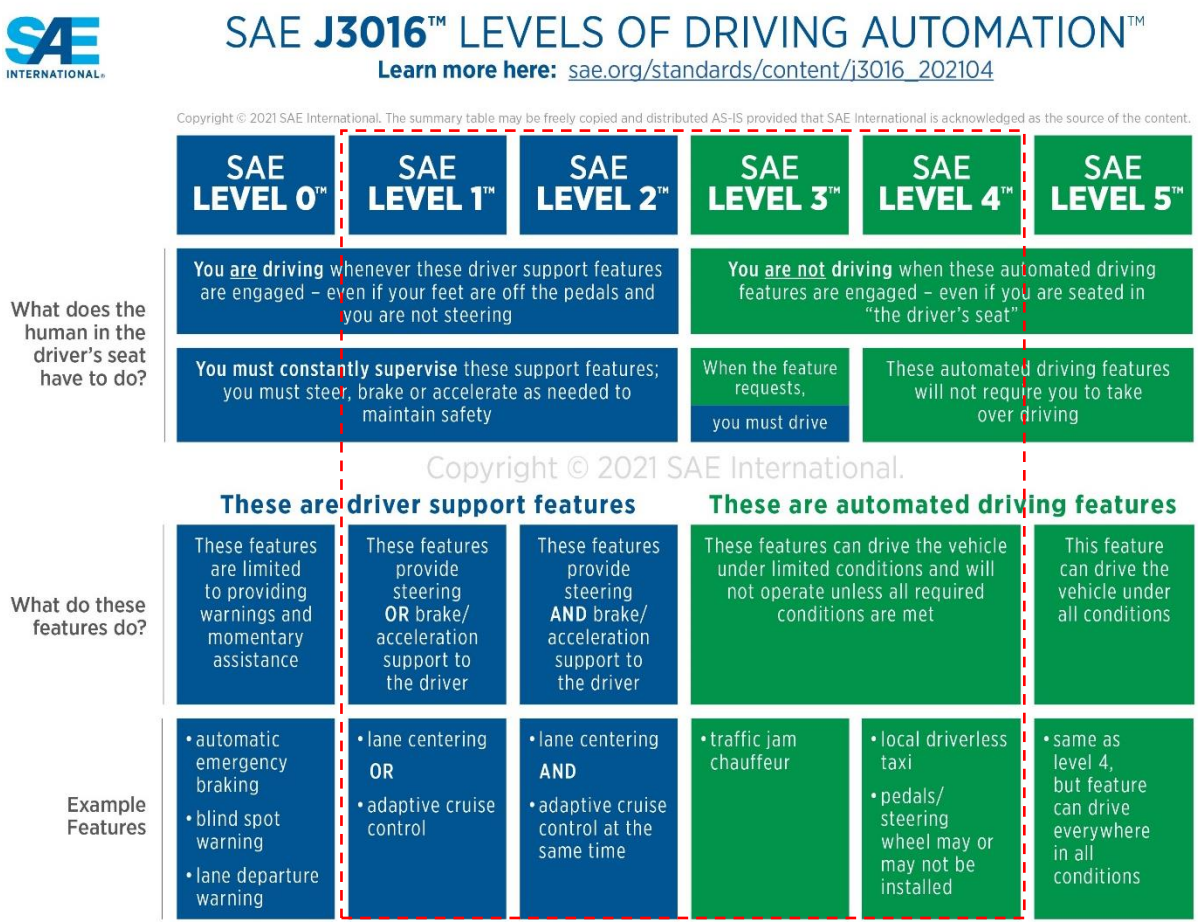


Figure 1-2. Visual chart for "Levels of Driving Automation" (SAE International, 2021b)

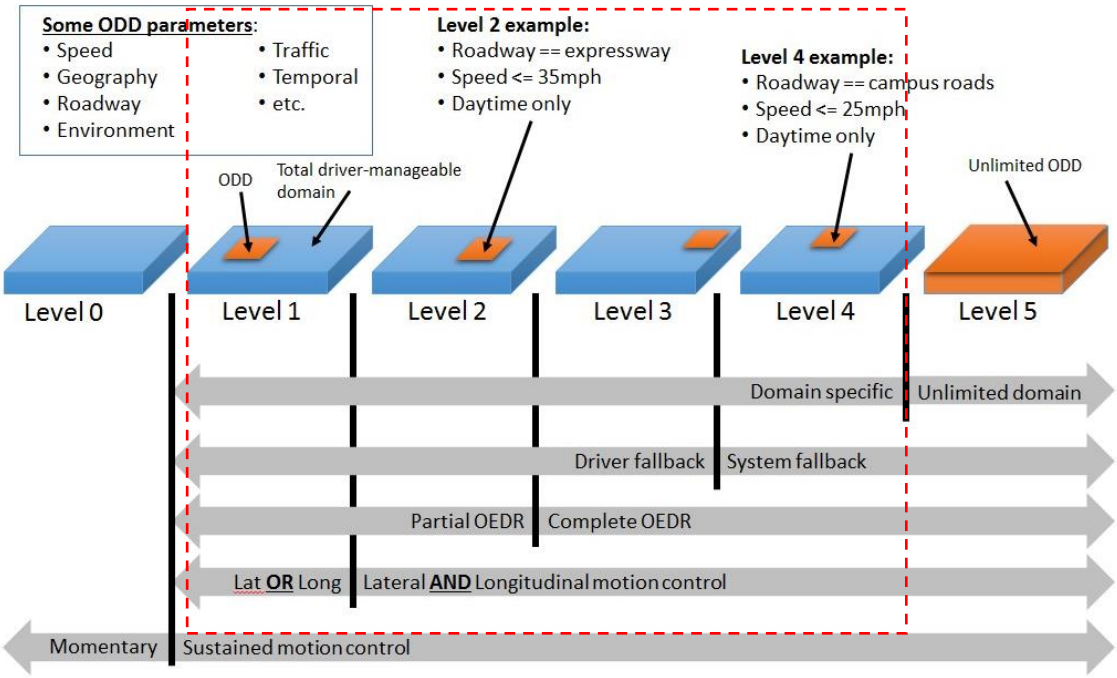


Figure 1-3. ODD relative to driving automation levels (SAE International, 2021a)

automation levels provided by the SAE. Typically, the main attributes of ODD include, but are not limited to, physical infrastructure (e.g., roadway type, lane markings), operational constraints (e.g., speed limit, traffic conditions), connectivity (e.g., vehicle-to-vehicle, vehicle-to-infrastructure communication), and environmental conditions (e.g., weather, illumination). It is worth noting that the ODD for all levels of automation, except for full automation (Level 5), is limited. Moreover, different Original Equipment Manufacturers (OEMs) may prescribe varying ODDs for their AVs, even for the same driving assistant functions (e.g., Adaptive Cruise Control, Lane Keeping Assistance System) within the same level of automation. This variability could lead to uncertainties and drivers' misunderstanding of the capabilities of AVs (Carsten & Martens, 2019; Farah et al., 2020; Noy et al., 2018; Wood et al., 2019). To enhance traffic safety and efficiency, it is essential to minimise instances where AVs exceed their ODD (Gyllenhammar et al., 2020). Therefore, enlarging the ODD of AVs is crucial. By expanding their ODD, AVs can be equipped to handle a wider range of challenging scenarios and adapt to diverse driving conditions, alleviating the occurrences of exceeding ODD. This ultimately enhances the overall capability of AVs to navigate safely and efficiently within mixed-traffic environments.

To expand the ODD of AVs, two key aspects must be addressed: sensing and perception, together with planning and control. Firstly, in terms of sensing and perception, AVs must effectively perceive and interact with the infrastructure and the static environment. This includes factors such as road width, curvature, lane marking types, and degradation conditions. AVs need to accurately sense and interpret their surroundings to navigate safely, especially in challenging driving scenarios that may fall outside their predefined ODD. Thus, enhancing sensing and perception capabilities is pivotal to ensuring AVs can handle diverse driving conditions and environments. Secondly, robust and optimised planning and control are imperative for AVs to interact with other moving road users, particularly HDVs. The complexity of interactions between AVs and HDVs is influenced by various factors, including, among others, different driving styles (aggressive, defensive, pro-social), driving culture, and norms (Negash & Yang, 2023; Orfanou et al., 2022; W. Wang et al., 2022). AVs must be equipped with robust planning and control algorithms to anticipate the behaviour of other road users and navigate safely in mixed-traffic environments. Overall, the development of methods to enhance sensing and perception capabilities and improve planning and control algorithms is essential for expanding the ODD of AVs, which is also the main research target of this thesis. Additionally, AVs must be equipped to handle critical edge cases that may fall beyond their ODD or in instances of system malfunctions within the AD system. In such scenarios, the ability of AVs to swiftly execute emergency response actions or seamlessly transition control back to human drivers is paramount. Consequently, effective anomaly monitoring and detection mechanisms play a crucial role in ensuring the safety and reliability of AVs. This thesis also addresses the challenges associated with anomaly detection.

1.2 Research gaps, questions, and objectives

The main objective of this thesis is to broaden the ODD to augment the capabilities of AVs, thereby enabling the realisation of safe, efficient, and socially compliant automated driving within mixed-traffic environments. As aforementioned, three key perspectives, i.e., sensing and perception, anomaly detection, as well as planning and control, are tackled in this thesis. This sub-section discusses the research gaps identified within each perspective, along with the corresponding research questions and objectives that this thesis aims to tackle.

For sensing and perception, the vision-based approach is chosen due to its widespread usage and practical utility in AD systems (Boukerche & Ma, 2021; Muhammad et al., 2022; Pavel et al., 2022; Zablocki et al., 2022). This thesis focuses on vision-based lane detection for its critical role in current vehicle localisation, ensuring proper positioning within lanes, and for its significance as the foundation of various ADAS systems, such as Lane Keeping Assistance and Lane Departure Warning (Andrade et al., 2019; Bar Hillel et al., 2014; W. Chen et al., 2020; Liang et al., 2020; Xing et al., 2018). Traditional vision-based lane-detection methods rely on hand-crafted low-level features, such as edges, geometric constraints, gradients and texture patterns, and involve several steps such as image pre-processing, feature extraction, line detection and fitting, and post-processing (Bai et al., 2018; Bar Hillel et al., 2014). These methods suffer from many shortcomings. They often require complex and time-consuming hand-crafted features, which may not be suitable or effective enough for AD. Additionally, they usually rely on a single image for lane detection. Recent advancements in computational hardware and deep neural network (DNN) models have enabled the development of deep learning-based lane detection methods. Generally, vision-based lane detection using deep learning approaches is typically categorised into three main perspectives: segmentation-based (Feng et al., 2022; Lee & Liu, 2023; Ren et al., 2022; Hai Wang et al., 2022; Zhang et al., 2021), anchor-based (Huang et al., 2023; Jin et al., 2022; Qin et al., 2022; Tabelini et al., 2021), and parameter-based (R. Liu et al., 2021; Torres et al., 2020), among which the segmentation-based method stands out as the most prevalent and widely utilised approach. These approaches automatically extract useful features and enable end-to-end lane detection, outperforming traditional methods (Hou et al., 2019; Neven et al., 2018; Pan et al., 2018; Tang et al., 2021). However, current deep learning methods used for vision-based lane detection fail to fully leverage the essential characteristics of lanes or account for significant spatial-temporal correlations and dependencies among critical regions in continuous driving image frames. Consequently, the detection results still remain unsatisfactory, particularly under extremely challenging driving conditions. Thus, Research Question 1 (RQ1) is formulated as follows:

Sensing and perception module

RQ1: *How can spatial-temporal features and correlations be effectively utilised to enhance vision-based sensing and perception capabilities (e.g., lane detection), and to what extent can these capabilities be improved?*

Sub research questions:

RQ 1-1: How to develop effective sequential deep neural network architecture or mechanism to effectively capture spatial-temporal correlations?

RQ 1-2: How to speed up the training of sequential deep neural network models? What strategies can be employed?

RQ 1-3: How to make efficient use of the available data, especially the unlabelled ones?

Accordingly, the corresponding research objectives and tasks are delineated as follows:

- To formulate a hybrid sequential DNN architecture towards enhanced spatial-temporal feature extraction and integration with domain knowledge, as well as the identification of spatial-temporal correlations and dependencies within continuous image frames; [*Chapter 2*]
- To develop DNN feature extraction mechanisms (e.g., attention) aimed at capturing spatial-temporal correlations within critical regions of continuous image frames; [*Chapter 3*]
- To devise efficient pipeline and self-supervised pre-training methods for training sequential DNN models and leveraging unlabelled image data; [*Chapter 4*]
- To validate the DNN models and the training methodology using different large-scale datasets; [*Chapters 2, 3,* and *4*]

Regarding anomaly detection, considering data availability, this thesis focuses on two use case studies: (1) detection of abnormal lane rendering images within navigation map apps, and (2) identification of abnormal driving behaviour in naturalistic driving scenarios. Data-driven machine learning (ML)-based anomaly detection methods have been widely employed across various domains, showing great promise (Alqahtani & Kumar, 2024; G. Li & Jung, 2023; Samariya & Thakkar, 2023). These methods can be categorised based on the type of data required to train the ML model into three main approaches: 1) Supervised ML, 2) Unsupervised ML, and 3) Semi-supervised ML. Usually, the data-driven ML methods are applied in a supervised manner, where each instance in the dataset is labelled as a normal sample or an anomaly. These methods train ML models on labelled datasets to fit ML models for automatically detecting anomalies in new input data. Regarding the two selected use cases, most of the available studies relied on (fully-) supervised ML models for anomaly detection, with limited exploration of unsupervised or semi-supervised methods. However, in real-world scenarios, ground truth labels are occasionally absent or inaccurate. Moreover, labelling extensive amounts of data can be tedious and even hazardous in certain critical situations. Additionally, there tends to be a significant data imbalance with an abundance of normal data compared to anomaly data, and some open-source datasets are only partially labelled. Considering these challenges, Research Question 2 (RQ2) is formulated as follows:

Anomaly detection module

RQ2: *How to develop effective semi-supervised/unsupervised machine learning methods for anomaly detection leveraging unlabelled data?*

Sub research questions:
RQ 2-1: What are the key features for anomaly detection, and how can they be identified?
RQ 2-2: How to develop pipeline and method to make efficient use of unlabelled data for enhancing anomaly detection?

Accordingly, the corresponding research objectives and tasks are outlined as follows:

- To customise and implement semi-supervised/self-supervised ML models for anomaly detection of the selected use cases; [*Chapters 5* and *6*]

- To comprehensively compare and evaluate the performance of supervised and semi-supervised ML algorithms regarding anomaly detection; [*Chapters 5* and *6*]
- To carry out feature engineering and identify the key features regarding anomaly detection (for abnormal driving behaviour); [*Chapter 6*]
- To develop a self-supervised pretraining method and a holistic pipeline for making efficient use of unlabelled data to enhance anomaly detection performance (for detection of abnormal lane rendering images); [*Chapter 5*]

Regarding AVs' planning and control, previous studies have traditionally focused on integrating safety, efficiency, comfort, and energy consumption into the development of automated driving algorithms (Du et al., 2022; ElSamadisy et al., 2024; Vasile et al., 2023; M. Zhu et al., 2020). However, ensuring that AVs are socially compliant, understood, and accepted by human drivers is equally crucial for enhancing safety and efficiency, particularly in mixed-traffic conditions. Consequently, the design of socially compliant driving strategies and behaviours is gaining prominence. While there have been some preliminary endeavours in this area (Hang et al., 2021; Kolekar et al., 2020; Schwarting et al., 2019; W. Wang et al., 2022), research examining this emerging topic still remains limited, with an integrated conceptual framework yet to be established. Furthermore, deep reinforcement learning (DRL), which combines the feature-capturing capabilities of deep learning with the decision-making aptitude of reinforcement learning, has been extensively utilised and acknowledged in the development of automated driving (Kiran et al., 2022; Z. Zhu & Zhao, 2022). It has been applied to various driving tasks and diverse driving scenarios, including car following (Yang et al., 2023; M. Zhu et al., 2018, 2020), lane changing (Y. Chen et al., 2019; T. Shi et al., 2019; G. Wang et al., 2022), and highway on-ramp merging (B. Liu et al., 2021; Huanjie Wang et al., 2021; S. Wu et al., 2022). However, significant gaps persist in addressing complex driving scenarios such as navigating roundabouts, which involve large curvature, interaction with multiple participants, and the necessity to manage both lateral and longitudinal control. These gaps are particularly evident in the utilisation and comparison of different DRL algorithms considering integrated reward mechanisms for achieving safe, efficient, comfortable, and energy-saving automated driving through roundabouts within mixed-traffic environments. Additionally, comprehensive evaluations of DRL algorithms across various driving manoeuvres and analyses of their adaptability, such as applying models trained in one scenario to another, are still insufficiently explored. Considering these aforementioned research gaps, Research Question 3 (RQ3) is formulated as follows:

Planning and control module

RQ3: *How to develop and optimise automated vehicles' driving strategies and styles to ensure safety, efficiency, and, particularly, social compliance in mixed-traffic environments?*

Sub research questions:
RQ 3-1: How can social norms and driving-related benefits for human-driven vehicles be effectively integrated into the development of automated driving strategies?
RQ 3-2: How do different deep reinforcement learning algorithms perform across different driving manoeuvres?
RQ 3-3: How can model performance be comprehensively evaluated and compared, particularly in terms of their adaptability to handle scenario shifts?

In accordance with RQ3, the corresponding research objectives and tasks are specified as follows:

- To design a conceptual framework that considers socially compliant decision-making in AD; [*Chapter 7*]
- To devise a learning-based approach, with a particular emphasis on DRL, for the complex scenario of roundabout driving in mixed-traffic conditions; [*Chapter 8*]
- To develop a model-based approach for AD that integrates the driving-related benefits of HDVs; [*Chapter 9*]
- To conduct a comprehensive comparison and assessment of different DRL algorithms across various driving manoeuvres, with special attention to their ability to adapt to scenario shifts; [*Chapter 8*]

By addressing the three research questions outlined above, this thesis contributes to the overarching objective of broadening the ODD to augment the capabilities of AVs, which will thereby facilitate the development, implementation, and deployment of safe, efficient, and socially compliant automated vehicles in mixed-traffic environments.

1.3 Research approach

This thesis employs a diverse range of methodologies and research methods, including literature review, conceptual and pipeline design, online survey, machine learning (particularly deep learning and deep reinforcement learning), model-based approach, as well as simulation-based training and validation. The literature review serves as a foundational component throughout each sub-research question and chapter, especially for [Chapter 7], which focuses on summarising the state-of-the-art on socially compliant automated driving. Conceptual and pipeline designs are utilised for tasks related to sensing and perception, anomaly detection, as well as planning and control, typically for the selected studies of lane detection, abnormal lane rendering image detection, and socially compliant automated driving. For instance, a three-phase pipeline is devised for lane detection [Chapter 4], while a four-phase pipeline is designed for abnormal lane rendering image detection [Chapter 5]. Additionally, for planning and control, a conceptual framework emphasising social compliance is developed [Chapter 7]. An online survey is employed particularly to gather insights and feedback from experts regarding the proposed conceptual framework for socially compliant automated driving [Chapter 7]. As for machine learning techniques, deep learning is adopted for sensing and perception (lane detection) as detailed in [Chapters 2, 3, and 4] together with anomaly detection (two use case studies) as detailed in [Chapters 5 and 6], while deep reinforcement learning is employed for AVs' planning and control in [Chapter 8] accompanied by simulation-based training and validation. Furthermore, a model-based approach is also adopted for AVs' planning and control in [Chapter 9], involving simulation-based training and validation as well.

To elaborate further, regarding enhancing the capabilities of AVs to enable them with a wider ODD, firstly, from the perspective of sensing and perception, vision-based deep learning is adopted for the chosen lane detection use case. Typically, two spatial-temporal DNN models are developed, complemented by a self-supervised pre-training method. The first DNN model, illustrated in [Chapter 2], integrates the spatial convolutional neural network (SCNN) for

single-image feature extraction with spatial-temporal Recurrent Neural Network (RNN) modules to capture correlations and dependencies among continuous images. The second model, detailed in [Chapter 3], focuses on designing customised spatial-temporal attention mechanisms to further enhance the utilisation of spatial-temporal correlations among different image regions in continuous frames for vision-based lane detection. The designed attention blocks are connected with the linear Long Short Term Memory (LSTM) neural networks, ensuring the model's lightweight nature and lower computational complexity. Additionally, as outlined in [Chapter 4], a self-supervised pre-training method using masked sequential autoencoders (MSAE) is proposed to enhance detection accuracy and expedite model training. A customised Focal Loss based PolyLoss is also introduced to further improve detection accuracy [Chapter 4]. The efficacy of the developed models and proposed self-supervised pre-training method is validated using various large-scale datasets (e.g., TuSimple, tvtLANE, and LLAMAS dataset).

Secondly, in the realm of anomaly detection, two distinct use cases are chosen, and both semi-supervised and fully-supervised machine learning approaches are explored. Given the formidable challenges of obtaining ground truth labels and the labour-intensive nature of data labelling, a strategic emphasis is placed on semi-supervised ML models leveraging the Hierarchical Extreme Learning Machine (HELM). Specifically, HELM is meticulously customised to detect abnormal driving behaviour in naturalistic driving scenarios, utilising partially labelled data while endeavouring to incorporate Surrogate Measures of Safety (SMoS) as pivotal input features to enhance detection performance [Chapter 6]. Furthermore, a four-phase pipeline, consisting of data pre-processing, self-supervised pre-training with Masked Image Modelling (MiM), customised fine-tuning using cross-entropy based loss with label smoothing, and post-processing, is proposed for the detection of abnormal lane rendering images within navigation map apps [Chapter 5]. Transforming lane rendering image anomaly detection into a classification problem, state-of-the-art fully supervised deep learning models, especially Transformer-based ones, are adopted with self-supervised pre-training using two MiM methods. The efficacies of the developed semi-supervised ML and fully supervised ML models, together with the proposed pipelines for the selected case studies, are all validated using large-scale real-world data.

Thirdly, for AVs' planning and control in the mixed-traffic context, based on the literature review and expert interview, a conceptual framework is devised, with a primary focus on socially compliant automated driving [Chapter 7]. Then, expert insights and feedback on the proposed framework are collected through an online survey, and the results are visualised and analysed. Both model-based and learning-based approaches are employed to develop AVs' planning and control models for the selected manoeuvre of driving through roundabouts in mixed-traffic conditions. In the learning-based approach, under the DRL framework, the corresponding agent, state, environment, and action space, together with an integrated multi-factor reward function considering safety, efficiency, comfort, and energy consumption, are designed [Chapter 8]. Multiple DRL algorithms, including Deep Deterministic Policy Gradient (DDPG), Proximal Policy Optimisation (PPO), and Trust Region Policy Optimisation (TRPO), are employed and implemented to instruct AVs' driving through roundabouts. Regarding the model-based approach, three interdisciplinary concepts, i.e., human perceived Driving Risk Field (DRF), Social Value Orientation (SVO), and Model Predictive Contouring Control (MPCC), are integrated [Chapter 9]. The DRF is utilised to model the perceived risk of surrounding human drivers, while SVO, from the social sciences field, is adopted to replicate

how AVs balance their own benefits against those of other surrounding HDVs. The model-based DRF-SVO is packaged into the MPCC framework to implement the integration of both planning and control simultaneously. Both the model-based and learning-based approaches undergo extensive simulation-based training and verification to ensure their effectiveness and safety [Chapters 8 and 9].

1.4 Contributions

In this sub-section, the main scientific and practical contributions of this thesis are highlighted.

1.4.1 Scientific contributions

(1) *A self-supervised pretraining method, two hybrid spatial-temporal DNN models, and a three-phased pipeline for enhancing vision-based sensing and perception*: This thesis introduces two novel hybrid spatial-temporal encoder-decoder based sequential DNN models designed for powerful feature extraction. These models seamlessly integrate single-image feature extraction with the detection of correlations and dependencies among continuous images. The developed models exhibit state-of-the-art performance in both normal and challenging driving scenarios while maintaining lower computational complexity and smaller model sizes. Additionally, a self-supervised pre-training method utilising MSAE is proposed to further enhance detection accuracy and speed up the model training process, ensuring robust performance in various driving conditions. Moreover, the introduction of customised Focal Loss based PolyLoss further improves the accuracy.

(2) *Development and evaluation of supervised and semi-supervised ML methods for anomaly detection*: This thesis presents the development of one semi-supervised ML method, namely the Hierarchical Extreme Learning Machine, for the detection of abnormal driving behaviour. The method can effectively detect anomalies using only partially labelled data, addressing challenges related to obtaining ground truth labels and labour-intensive data labelling. It outperformed its semi-supervised counterparts. Moreover, for abnormal driving behaviour detection, the incorporation of SMoS (to be specific, event-based safety indicators) as pivotal input features significantly enhances detection accuracy. Additionally, fully supervised Transformer-based models are developed for detecting abnormal lane rendering images within navigation map applications. These models incorporate self-supervised pre-training using MiM and customised fine-tuning with a cross-entropy based loss function enhanced by label smoothing. The resulting models prove effective in achieving high accuracy, recall ratio, and F1 scores. To the best of the author's knowledge, this thesis represents the first exploration regarding semi-supervised and self-supervised ML methods applied for the selected anomaly detection use cases.

(3) *A conceptual framework for developing and implementing socially compliant automated vehicles*: This thesis marks a pioneering effort in developing an integrated conceptual framework for socially compliant automated driving. Within the proposed conceptual framework, various social components such as culture, norms, and cues, along with different driving styles (e.g., aggressive, cautious, pro-social), are systematically incorporated. Notably, bidirectional behavioural adaptation is introduced as a novel concept, emphasising the dynamic and bidirectional interactions and adaptations between AVs and human drivers. Furthermore, the framework emphasises the importance of balancing the benefits of ego AV in

terms of safety, comfort, and efficiency with the needs and expectations of other road users, highlighting the necessity for a nuanced trade-off strategy on a case-by-case basis. Additionally, the proposal of a spatial-temporal memory module facilitates the long-term and short-term upgradation of knowledge and rules, enabling the implementation and refinement of driving strategies that consider bidirectional behavioural adaptation. The validity and effectiveness of the proposed conceptual framework are assessed through an online questionnaire-based survey, garnering expert insights and feedback to refine and validate its components.

(4) *Model-based social-aware planning and control for automated driving through roundabouts*: This thesis introduces an integrated social-aware planning and control algorithm, i.e., DRF-SVO-MPCC, for AVs' driving through roundabouts, leveraging three interdisciplinary terms (i.e., Driving Risk Field, Social Value Orientation, and Model Predictive Contouring Control). By integrating these elements and referring to the desired velocity, the DRF-SVO-MPCC model facilitates two driving styles, prosocial and egoistic. Through extensive simulation testing utilising an open-sourced platform, the DRF-SVO-MPCC algorithm demonstrates superior performance across various scenarios of roundabout navigation, including single-lane, two-lane, and scenarios with and without surrounding HDVs.

(5) *Implementation and comprehensive evaluation of DRL-based planning and control for AVs*: This thesis implements various DRL algorithms (e.g., DDPG, PPO, and TRPO) to guide AVs' driving through roundabouts in mixed-traffic conditions, considering safety, efficiency, comfort, and energy consumption. This thesis also conducts a comprehensive evaluation and comparison of DRL algorithms, including Deep Q-Network (DQN) and TRPO, across different driving manoeuvres such as highway driving and driving through unsignalised intersections. Additionally, the thesis designs a customised training environment that encompasses various driving manoeuvres and multiple road scenarios to train a unified driving model capable of handling diverse situations. Notably, the thesis also investigates the adaptability of DRL algorithms across different scenarios, assessing their capability to handle scenario shifts. To the best of the author's knowledge, this thesis represents a pioneering effort in conducting such a comprehensive evaluation, particularly with scenario shifting considered, a facet that has been seldom covered by previous studies in this field.

1.4.2 Practical contributions

In addition to the aforementioned scientific contributions, the outcomes of this thesis also yield implications for society and various stakeholders in terms of policy-making, technological design and advancement, as well as the development, implementation, and understanding of AVs in mixed-traffic environments.

For ***original equipment manufacturers***, particularly ***car manufacturers***, the two novel hybrid spatial-temporal DNN models developed in this thesis, coupled with the designed three-phased pipeline featuring the proposed self-supervised pretraining method, offer valuable insights for enhancing vision-based sensing and perception. These insights encompass improvements in accuracy as well as reductions in model complexity. As computational hardware continues to advance, the developed DNN models and associated training methods are poised for deployment within vehicles, underscoring their practical applicability in automotive contexts.

Similarly, the developed semi-supervised and fully-supervised ML models for anomaly detection can aid in various aspects of automotive safety and efficiency. By enabling the detection of abnormal behaviours and edge events, these models support predictive safety maintenance strategies and facilitate early conflict detection and avoidance. Moreover, they can enable proactive interventions to mitigate potential risks stemming from AV system anomalies or abnormal human driving behaviours, thereby enhancing traffic flow safety and efficiency.

Furthermore, the proposed conceptual framework for socially compliant automated driving can serve as a guiding framework for the development and implementation of socially compliant AVs. By systematically integrating various social components and driving styles, this framework provides valuable insights and assistance in designing AVs that seamlessly interact with other road users in a socially acceptable manner. This framework will be particularly beneficial for car manufacturers when developing their future vehicles, as it offers a structured approach to incorporating social considerations into AV design. The framework also emphasises the importance of striking a balance between the benefits of AVs and the needs of other road users, thereby it can foster greater acceptance and integration of AV technology into existing transportation systems. This will not only enhance safety and efficiency but also promote harmonious coexistence between AVs and HDVs on the roads.

Additionally, the developed model-based and learning-based planning and control algorithms represent significant practice in the realm of social-aware automated driving. These algorithms enable AVs to navigate complex traffic scenarios while considering safety, efficiency, energy consumption, and the benefits of surrounding HDVs. These algorithms provide valuable insights for car manufacturers when developing their AVs' planning and control modules.

For ***road operators***, the developed lane detection methods outlined in this thesis can offer significant potential benefits concerning lane marking inspection and maintenance. These methods streamline the process by automating or semi-automating the detection of lane markings, thereby reducing the necessity for manual inspection and intervention. As a result, road maintenance tasks can be performed more efficiently, leading to cost savings and improved productivity.

Additionally, the use of these lane detection methods contributes to improved road safety. Clear and well-maintained lane markings enhance visibility and guidance for drivers and AVs, reducing the risk of accidents. By promptly identifying and rectifying any issues with lane markings, road maintenance operators can play a vital role in ensuring safe and efficient road networks.

Furthermore, the implementation of the proposed advanced lane detection technologies can support data-driven decision-making in road maintenance planning and prioritisation. The results collected through automated lane detection can inform maintenance schedules, infrastructure investments, and resource allocation strategies, leading to optimised maintenance practices and improved overall road quality.

For ***authorities and insurance companies***, this thesis presents possible methods for monitoring and detecting abnormal driving behaviours, which can prove invaluable for driver training initiatives and insurance pricing strategies. By leveraging the developed anomaly detection models and techniques, authorities can establish effective mechanisms for monitoring driver behaviour on the roads, identifying and addressing unsafe driving practices, thereby

contributing to enhanced road safety and accident prevention efforts. Furthermore, insurance companies can utilise the insights provided by the developed abnormal driving detection methods to assess risk profiles more accurately. By incorporating data-driven assessments of driver behaviour into their pricing models, insurers can offer more personalised and fair insurance premiums. This approach not only promotes safer driving habits among policyholders but also incentivises responsible behaviour behind the wheel.

For ***policymakers***, this thesis offers valuable insights and potential strategies for the implementation and deployment of socially compliant AVs. The conceptual framework for socially compliant automated driving outlined in this thesis provides a structured approach for policymakers to understand and address the complex social dynamics involved in AV development and deployment. By considering factors such as culture, norms, driving styles, and bidirectional behavioural adaptation, policymakers can formulate comprehensive strategies that account for diverse societal contexts and expectations and promote the adoption and integration of AV technology into existing transportation systems in a socially responsible manner.

The emphasis on balancing the benefits of AVs with the benefits of other road users underscores the importance of stakeholder engagement and collaboration in policymaking processes. By actively involving various stakeholders, including government agencies, industry representatives, transportation experts, and community groups, policymakers can develop inclusive and consensus-driven policies that promote the equitable integration of AVs into transportation systems.

Furthermore, policymakers can leverage the insights provided by this thesis to develop regulatory frameworks, standards, and guidelines that govern the design, operation, and deployment of AVs. By establishing clear rules, standards, and requirements for AV manufacturers, operators, and users, policymakers can ensure that AVs adhere to socially acceptable norms and behaviours, thereby enhancing public acceptance and trust in AV technology.

For ***human drivers and the general public***, this thesis provides valuable insights into understanding AV technology and its implications for transportation systems, empowering individuals to better comprehend the complexities and potential benefits of AV technology. Through the exploration of topics such as deep learning-based sensing, perception, and anomaly detection, together with model-based and reinforcement learning-based planning and control, this thesis demystifies the inner workings of AV systems and elucidates their role in shaping the future of transportation. By offering clear explanations and real-world examples, this thesis fosters greater awareness and literacy regarding AV technology among human drivers and the general public.

Moreover, by highlighting the importance of social acceptance and trust in AV technology, this thesis encourages informed dialogue and engagement among stakeholders and empowers individuals to participate in discussions about the future of mixed traffic and the development of mobility and transportation.

Furthermore, by showcasing the potential benefits of AV technology, such as improved safety, efficiency, and accessibility, this thesis helps to dispel misconceptions and concerns that may exist among human drivers and the general public. By demonstrating the transformative impact

that AVs can have on transportation systems, this thesis encourages openness to innovation and adaptation to emerging technologies.

Overall, these practical contributions are instrumental in driving forward the development and deployment of AVs that are not only technologically advanced but also socially responsible, paving the way for the smoother transition and widespread adoption of AV technology in real-world environments.

1.5 Thesis outline

The thesis outline is illustrated in **Figure 1-4**. This thesis consists of ten chapters in total. The lines with arrows depict the relationship between the chapters. **Chapter 1** serves as an introduction, providing background information, identifying research gaps, stating research questions and objectives, describing research methods, and highlighting the main contributions. The subsequent eight chapters address sub-research questions and objectives proposed in *Section 1.2* and are arranged into three main pillars tackling the modules of sensing and perception, anomaly detection, and AVs' planning and control, respectively. Theoretically and ideally, sensing and perception should aid in anomaly detection, while reliable sensing, perception, and anomaly detection will contribute to planning and control; however, in this thesis, the datasets and targeting manoeuvres used for each pillar are different, so there is no direct linkage. Thus, the three main pillars are connected with dashed arrows. The remaining chapters in this thesis are structured as follows, followed by further elaboration:

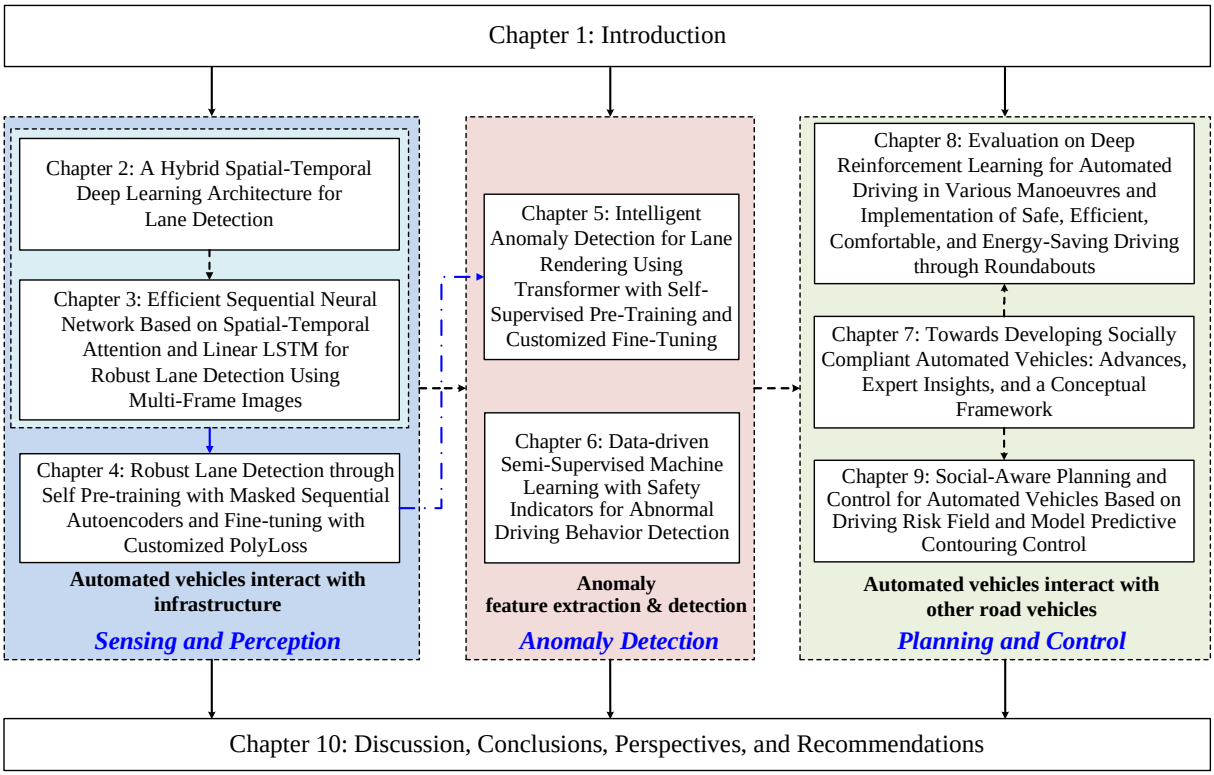


Figure 1-4. The outline of the thesis: thesis structure and relations between chapters

Chapters 2-4 focus on the vision-based sensing and perception task and address RQ1, with Chapter 2 and Chapter 3 designing two hybrid spatial-temporal DNN models (*RQ 1-1*), and Chapter 4 proposing a three-phase pipeline featuring self-supervised pre-training with Masked

Sequential Autoencoders (*RQ 1-2* and *RQ 1-3*). Specifically, Chapter 2 develops a hybrid spatial-temporal deep learning architecture integrating single-image feature extraction using SCNN with capturing correlations and dependencies among continuous images by spatial-temporal RNN modules. Chapter 3 focuses on designing customised spatial-temporal attention mechanisms to make efficient use of spatial-temporal correlations among different image regions across continuous frames. In this chapter, the linear LSTM neural networks are connected with the proposed attention blocks, which ensures lightweight and lower computational complexity. Chapter 4 proposes a self-supervised pre-training method using MSAE to enhance detection accuracy and expedite training of the aforementioned two DNN models (so there is a blue arrow that connects the group of Chapters 2 and 3 with Chapter 4). Additionally, to further improve detection accuracy, a customised Focal Loss based PolyLoss is introduced in Chapter 4.

Chapters 5-6 focus on anomaly detection for two selected use cases related to automated driving and address RQ2. Chapter 5 focuses on intelligent anomaly detection for lane rendering using Transformer models with a pipeline integrating self-supervised pre-training and customised fine-tuning (*RQ 2-2*), and Chapter 6 implements a data-driven semi-supervised machine learning model (i.e., HELM) enhanced with SMoS (i.e., event-based safety indicators) as the important features (*RQ 2-1*) for abnormal driving behaviour detection. Both use case studies leverage unlabelled data to learn the feature patterns and to enhance the detection accuracy (*RQ 2-2*).

The dashed blue arrow connecting Chapter 4 and Chapter 5 indicates their use of similar pipelines with self-supervised pre-training and their shared focus on vision-based tasks.

Chapters 7-9 address RQ3 and emphasise developing AVs' planning and control with a conceptual design, together with model-based and learning-based approaches for simulation-based implementation. To be specific, Chapter 7 devises a conceptual framework with a primary focus on socially compliant automated driving, considering various social components, different driving styles, bidirectional behavioural adaptation, and balancing the benefits of ego AVs with the benefits of other road users (*RQ 3-1*). This conceptual framework provides guidance for the development of AVs' planning and control in Chapters 8 and 9, where specific aspects of the conceptual module design introduced in Chapter 7 are validated to a certain extent. Therefore, there are dashed arrows that connect Chapter 7 with Chapters 8 and 9. Chapter 8 depicts the learning-based approach by designing the DRL agent, state, environment, and action space, together with an integrated multi-factor reward function considering safety, efficiency, comfort, and energy consumption. Multiple DRL algorithms are employed and implemented to instruct AVs' driving through various scenarios. Chapter 8 also assesses the adaptability of DRL algorithms across different scenarios (*RQ 3-2*), considering their capability to handle scenario shifts (*RQ 3-3*). While Chapter 9 designs a model-based approach, integrating three interdisciplinary concepts, i.e., DRF, SVO, and MPCC, to form the DRF-SVO-MPCC model for social-aware planning and control of AVs driving through roundabouts. The developed DRF-SVO-MPCC model facilitates two driving styles, prosocial and egoistic, with the prosocial style enabling AVs to navigate complex traffic scenarios considering both their ego safety, efficiency, and the benefits of surrounding HDVs (*RQ 3-1*). The models developed in both Chapter 8 and Chapter 9 are tested and verified through simulation-based experiments.

Finally, **Chapter 10** discusses the key findings and the limitations of the thesis. Prospective recommendations for practice and future research are also provided.

References

Alqahtani, H., & Kumar, G. (2024). Machine learning for enhancing transportation security: A comprehensive analysis of electric and flying vehicle systems. Engineering Applications of Artificial Intelligence, 129(November 2023), 107667. https://doi.org/10.1016/j.engappai.2023.107667

Andrade, D. C., Bueno, F., Franco, F. R., Silva, R. A., Neme, J. H. Z., Margraf, E., Omoto, W. T., Farinelli, F. A., Tusset, A. M., Okida, S., Santos, M. M. D., Ventura, A., Carvalho, S., & Amaral, R. D. S. (2019). A novel strategy for road lane detection and tracking based on a vehicle's forward monocular camera. IEEE Transactions on Intelligent Transportation Systems, 20(4), 1497–1507. https://doi.org/10.1109/TITS.2018.2856361

Bai, M., Mattyus, G., Homayounfar, N., Wang, S., Lakshmikanth, S. K., & Urtasun, R. (2018). Deep multi-sensor lane detection. IEEE International Conference on Intelligent Robots and Systems, 3102–3109. https://doi.org/10.1109/IROS.2018.8594388

Bar Hillel, A., Lerner, R., Levi, D., & Raz, G. (2014). Recent progress in road and lane detection: A survey. Machine Vision and Applications, 25(3), 727–745. https://doi.org/10.1007/s00138-011-0404-2

Boukerche, A., & Ma, X. (2021). Vision-based autonomous vehicle recognition: A new challenge for deep learning-based systems. ACM Computing Surveys. https://doi.org/10.1145/3447866

Carsten, O., & Martens, M. H. (2019). How can humans understand their automated cars? HMI principles, problems and solutions. Cognition, Technology and Work, 21(1), 3–20. https://doi.org/10.1007/s10111-018-0484-0

Chen, W., Wang, W., Wang, K., Li, Z., Li, H., & Liu, S. (2020). Lane departure warning systems and lane line detection methods based on image processing and semantic segmentation–a review. Journal of Traffic and Transportation Engineering (English Edition), xxx. https://doi.org/10.1016/j.jtte.2020.10.002

Chen, Y., Dong, C., Palanisamy, P., Mudalige, P., Muelling, K., & Dolan, J. M. (2019). Attention-based hierarchical deep reinforcement learning for lane change behaviors in autonomous driving. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2019-June, 1326–1334. https://doi.org/10.1109/CVPRW.2019.00172

Du, Y., Chen, J., Zhao, C., Liu, C., Liao, F., & Chan, C. Y. (2022). Comfortable and energy-efficient speed control of autonomous vehicles on rough pavements using deep reinforcement learning. Transportation Research Part C: Emerging Technologies, 134(December 2021), 103489. https://doi.org/10.1016/j.trc.2021.103489

ElSamadisy, O., Shi, T., Smirnov, I., & Abdulhai, B. (2024). Safe, efficient, and comfortable reinforcement-learning-based car-following for AVs with an analytic safety guarantee and dynamic target speed. Transportation Research Record, 2678(1), 643–661. https://doi.org/10.1177/03611981231171899

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. Transportation Research Part A: Policy and Practice, 77, 167–181. https://doi.org/10.1016/j.tra.2015.04.003

Farah, H., Bhusari, S., van Gent, P., Babu, F. A. M., Morsink, P., Happee, R., & van Arem, B. (2020). An empirical analysis to assess the operational design domain of lane keeping system equipped vehicles combining objective and subjective risk measures. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2020.2969928

Feng, Z., Guo, Y., Liang, Q., Bhutta, M. U. M., Wang, H., Liu, M., & Sun, Y. (2022). MAFNet: Segmentation of road potholes with multimodal attention fusion network for autonomous vehicles. IEEE Transactions on Instrumentation and Measurement, 71. https://doi.org/10.1109/TIM.2022.3200100

Fraedrich, E., Beiker, S., & Lenz, B. (2015). Transition pathways to fully automated driving and its implications for the sociotechnical system of automobility. European Journal of Futures Research, 3(1). https://doi.org/10.1007/s40309-015-0067-8

Greenblatt, J. B., & Shaheen, S. (2015). Automated vehicles, on-demand mobility, and environmental impacts. Current Sustainable/Renewable Energy Reports, 2(3), 74–81. https://doi.org/10.1007/s40518-015-0038-5

Gyllenhammar, M., Johansson, R., Warg, F., Chen, D., Heyn, H.-M., Sanfridson, M., Söderberg, J., Thorsén, A., Ursing, S., Ab, Z., & Com, M. G. (2020). Towards an operational design domain that supports the safety argumentation of an automated driving system. 10th European Congress on Embedded Real Time Systems, 1–10. https://www.diva-portal.org/smash/get/diva2:1390550/FULLTEXT01.pdf

Hang, P., Lv, C., Xing, Y., Huang, C., & Hu, Z. (2021). Human-like decision making for autonomous driving : A noncooperative game theoretic approach. 22(4), 2076–2087.

Hou, Y., Ma, Z., Liu, C., & Loy, C. C. (2019). Learning lightweight lane detection CNNS by self attention distillation. Proceedings of the IEEE International Conference on Computer Vision, 2019-Octob, 1013–1021. https://doi.org/10.1109/ICCV.2019.00110

Huang, S., Shen, Z., Huang, Z., Ding, Z., Dai, J., Han, J., Wang, N., & Liu, S. (2023). Anchor3DLane: Learning to regress 3D anchors for monocular 3D lane detection. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 17451-17460, https://doi.org/10.1109/CVPR52729.2023.01674

Jamson, H., Merat, N., Carsten, O., & Lai, F. (2011). Fully-automated driving: The road to future vehicles. In Driving Assessment Conference (Vol. 6, No. 2011). University of Iowa. https://doi.org/10.17077/drivingassessment.1370

Jia, S., Hui, F., Li, S., Zhao, X., & Khattak, A. J. (2020). Long short-term memory and convolutional neural network for abnormal driving behaviour recognition. IET Intelligent Transport Systems, 14(5), 306–312. https://doi.org/10.1049/iet-its.2019.0200

Jin, D., Park, W., Jeong, S.-G., Kwon, H., & Kim, C.-S. (2022). Eigenlanes: Data-driven lane descriptors for structurally diverse lanes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17163-17171).

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Sallab, A. A. A., Yogamani, S., & Perez, P. (2022). Deep reinforcement learning for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems, 23(6), 4909–4926. https://doi.org/10.1109/TITS.2021.3054625

Kolekar, S., de Winter, J., & Abbink, D. (2020). Human-like driving behaviour emerges from a risk-based driver model. Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-18353-4

Lee, D. H., & Liu, J. L. (2023). End-to-end deep learning of lane detection and path prediction for real-time autonomous driving. Signal, Image and Video Processing. https://doi.org/10.1007/s11760-022-02222-2

Li, G., & Jung, J. J. (2023). Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges. Information Fusion, 91, 93–102. https://doi.org/10.1016/j.inffus.2022.10.008

Liang, D., Guo, Y. C., Zhang, S. K., Mu, T. J., & Huang, X. (2020). Lane detection: A survey with new results. Journal of Computer Science and Technology, 35(3), 493–505. https://doi.org/10.1007/s11390-020-0476-4

Liu, B., Tang, Y., Ji, Y., Shen, Y., & Du, Y. (2021). A deep reinforcement learning approach for ramp metering based on traffic video data. Journal of Advanced Transportation, 2021. https://doi.org/10.1155/2021/6669028

Liu, R., Yuan, Z., Liu, T., & Xiong, Z. (2021). End-to-end lane shape prediction with transformers. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 3694–3702.

Muhammad, K., Hussain, T., Ullah, H., Ser, J. Del, Rezaei, M., Kumar, N., Hijji, M., Bellavista, P., & De Albuquerque, V. H. C. (2022). Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2022.3207665

Negash, N. M., & Yang, J. (2023). Driver behavior modeling toward autonomous vehicles: Comprehensive review. In IEEE Access. https://doi.org/10.1109/ACCESS.2023.3249144

Neven, D., De Brabandere, B., Georgoulis, S., Proesmans, M., & Van Gool, L. (2018). Towards end-to-end lane detection: An instance segmentation approach. IEEE Intelligent Vehicles Symposium, Proceedings, 2018-June, 286–291. https://doi.org/10.1109/IVS.2018.8500547

Noy, I. Y., Shinar, D., & Horrey, W. J. (2018). Automated driving: Safety blind spots. Safety Science, 102(October 2017), 68–78. https://doi.org/10.1016/j.ssci.2017.07.018

Orfanou, F. P., Vlahogianni, E. I., Yannis, G., & Mitsakis, E. (2022). Humanizing autonomous vehicle driving: Understanding, modeling and impact assessment. Transportation Research Part F: Traffic Psychology and Behaviour. https://doi.org/10.1016/j.trf.2022.04.008

Pan, X., Shi, J., Luo, P., Wang, X., & Tang, X. (2018). Spatial as deep: Spatial CNN for traffic scene understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

Pavel, M. I., Tan, S. Y., & Abdullah, A. (2022). Vision-based autonomous vehicle systems based on deep learning: A systematic literature review. In Applied Sciences (Switzerland). https://doi.org/10.3390/app12146831

Qin, Z., Zhang, P., & Li, X. (2022). Ultra fast deep lane detection with hybrid anchor driven ordinal classification. IEEE Transactions on Pattern Analysis and Machine Intelligence. 46(5), 2555-2568. https://doi.org/10.1109/TPAMI.2022.3182097

Ren, X., Li, M., Li, Z., Wu, W., Bai, L., & Zhang, W. (2022). Curiosity-driven attention for anomaly road obstacles segmentation in autonomous driving. IEEE Transactions on Intelligent Vehicles, 1–11. https://doi.org/10.1109/TIV.2022.3204714

SAE International, (2021a). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles J3016_202104, https://www.sae.org/standards/content/j3016_202104/

SAE International, (2021b). SAE levels of driving automation refined for clarity and international audience [WWW Document]. SAE Int. URL https://www.sae.org/blog/saej3016-update (Accessed 2024-03-18)

Samariya, D., & Thakkar, A. (2023). A comprehensive survey of anomaly detection algorithms. Annals of Data Science, 10(3), 829–850. https://doi.org/10.1007/s40745-021-00362-9

Sar, I., Routray, A., & Mahanty, B. (2023). A review on existing technologies for the identification and measurement of abnormal driving. International Journal of Intelligent Transportation Systems Research, 21(1), 159–177. https://doi.org/10.1007/s13177-023-00343-7

Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., & Rus, D. (2019). Social behavior for autonomous vehicles. Proceedings of the National Academy of Sciences of the United States of America, 116(50), 2492–24978. https://doi.org/10.1073/pnas.1820676116

Shi, T., Wang, P., Cheng, X., Chan, C. Y., & Huang, D. (2019). Driving decision and control for automated lane change behavior based on deep reinforcement learning. 2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019, 2895–2900. https://doi.org/10.1109/ITSC.2019.8917392

Tabelini, L., Berriel, R., Paixão, T. M., Badue, C., de Souza, A. F., & Oliveira-Santos, T. (2021). Keep your eyes on the lane: Real-time attention-guided lane detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 294–302. https://doi.org/10.1109/CVPR46437.2021.00036

Talebpour, A., & Mahmassani, H. S. (2016). Influence of connected and autonomous vehicles on traffic flow stability and throughput. Transportation Research Part C: Emerging Technologies. https://doi.org/10.1016/j.trc.2016.07.007

Tang, J., Li, S., & Liu, P. (2021). A review of lane detection methods based on deep learning. Pattern Recognition, 111, 107623. https://doi.org/10.1016/j.patcog.2020.107623

Toroyan, T., Peden, M. M., & Iaych, K. (2013). WHO launches second global status report on road safety. Injury Prevention. https://doi.org/10.1136/injuryprev-2013-040775

Torres, L. T., Berriel, R. F., Paixão, T. M., Badue, C., De Souza, A. F., & Oliveira-Santos, T. (2020). PolyLaneNet: Lane estimation via deep polynomial regression. ICPR, 6150–6156.

Vasile, L., Dinkha, N., Seitz, B., Dasch, C., & Schramm, D. (2023). Comfort and safety in conditional automated driving in dependence on personal driving behavior. IEEE Open Journal of Intelligent Transportation Systems, 4(June), 772–784. https://doi.org/10.1109/OJITS.2023.3323431

Wang, G., Hu, J., Li, Z., & Li, L. (2022). Harmonious lane changing via deep reinforcement learning. IEEE Transactions on Intelligent Transportation Systems, 23(5), 4642–4650. https://doi.org/10.1109/TITS.2020.3047129

Wang, Hai, Chen, Y., Cai, Y., Chen, L., Li, Y., Sotelo, M. A., & Li, Z. (2022). SFNet-N: An improved sfnet algorithm for semantic segmentation of low-light autonomous driving road

scenes. IEEE Transactions on Intelligent Transportation Systems, 23(11), 21405–21417. https://doi.org/10.1109/TITS.2022.3177615

Wang, Huanjie, Yuan, S., Guo, M., Li, X., & Lan, W. (2021). A deep reinforcement learning-based approach for autonomous driving in highway on-ramp merge. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 235(10–11), 2726–2739. https://doi.org/10.1177/0954407021999480

Wang, W., Wang, L., Zhang, C., Liu, C., & Sun, L. (2022). Social interactions for autonomous driving: A review and perspectives. Foundations and Trends® in Robotics. https://doi.org/10.1561/2300000078

Wu, S., Tian, D., Zhou, J., Duan, X., Sheng, Z., & Zhao, D. (2022). Autonomous on-ramp merge strategy using deep reinforcement learning in uncertain highway environment. Proceedings of 2022 IEEE International Conference on Unmanned Systems, ICUS 2022, 658–663. https://doi.org/10.1109/ICUS55513.2022.9986560

Xing, Y., Lv, C., Chen, L., Wang, H., Wang, H., Cao, D., Velenis, E., & Wang, F. Y. (2018). Advances in vision-based lane detection: Algorithms, integration, assessment, and perspectives on ACP-based parallel vision. IEEE/CAA Journal of Automatica Sinica, 5(3), 645–661. https://doi.org/10.1109/JAS.2018.7511063

World Health Organization. (2023). Global status report on road safety 2023. https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023

Yang, X., Zou, Y., Zhang, H., Qu, X., & Chen, L. (2023). Improved deep reinforcement learning for car-following decision-making. Physica A: Statistical Mechanics and Its Applications, 624(4800), 128912. https://doi.org/10.1016/j.physa.2023.128912

Yaqoob, I., Khan, L. U., Kazmi, S. M. A., Imran, M., Guizani, N., & Hong, C. S. (2020). Autonomous driving cars in smart cities: Recent advances, requirements, and challenges. IEEE Network. https://doi.org/10.1109/MNET.2019.1900120

Zablocki, É., Ben-Younes, H., Pérez, P., & Cord, M. (2022). Explainability of deep vision-based autonomous driving systems: Review and challenges. International Journal of Computer Vision. https://doi.org/10.1007/s11263-022-01657-x

Zhang, Y., Gao, B., Guo, L., Guo, H., & Chen, H. (2021). Adaptive decision-making for automated vehicles under roundabout scenarios using optimization embedded reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems, 32(12), 5526–5538. https://doi.org/10.1109/TNNLS.2020.3042981

Zhu, M., Wang, X., & Wang, Y. (2018). Human-like autonomous car-following model with deep reinforcement learning. Transportation Research Part C: Emerging Technologies, 97(October), 348–368. https://doi.org/10.1016/j.trc.2018.10.024

Zhu, M., Wang, Y., Pu, Z., Hu, J., Wang, X., & Ke, R. (2020). Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving. Transportation Research Part C: Emerging Technologies, 117, 102662. https://doi.org/10.1016/j.trc.2020.102662

Zhu, Z., & Zhao, H. (2022). A survey of deep RL and IL for autonomous driving policy learning. IEEE Transactions on Intelligent Transportation Systems, 23(9), 14043–14065. https://doi.org/10.1109/TITS.2021.3134702

2 A hybrid spatial-temporal deep learning architecture for lane detection

Abstract

Accurate and reliable lane detection is vital for the safe performance of Lane Keeping Assistance and Lane Departure Warning systems. However, under certain challenging circumstances, it is difficult to get satisfactory performance in accurately detecting the lanes from one single image as is mostly done in current literature. Since lane markings are continuous lines, the lanes that are difficult to be accurately detected in the current single image can potentially be better deduced if information from previous frames is incorporated. This study proposes a novel hybrid spatial-temporal sequence-to-one deep learning architecture. This architecture makes full use of the spatial-temporal information in multiple continuous image frames to detect the lane markings in the very last frame. Specifically, the hybrid model integrates the following aspects: (a) the single image feature extraction module equipped with the spatial convolutional neural network (SCNN); (b) the spatial-temporal feature integration module constructed by spatial-temporal recurrent neural network (ST-RNN); (c) the encoder-decoder structure, which makes this image segmentation problem work in an end-to-end supervised learning format. Extensive experiments reveal that the proposed model architecture can effectively handle challenging driving scenes and outperforms available state-of-the-art methods.

This chapter is based on the journal publication:

Dong, Y., Patil, S., Van Arem, B., & Farah, H. (2023). A Hybrid Spatial-Temporal Deep Learning Architecture for Lane Detection. Computer-Aided Civil and Infrastructure Engineering, 38(1), 67–86. https://doi.org/10.1111/mice.12829

2.1 Introduction

The interest in developing automated driving functionalities, and in the end, fully automated vehicles, has been increasing vastly over the last decade. The safety of these automated functionalities is a crucial element and a priority for academic researchers, manufacturers, policymakers, and their potential future users. Automated driving requires a full understanding of the environment around the automated vehicle through its sensors. Vision-based methods have lately been boosted by advancements in computer vision and machine learning. Regarding environmental perception, camera-based lane detection is important as it allows the vehicle to position itself within the lane. This is also the foundation of most Lane Keeping Assistance and Lane Departure Warning systems (Andrade et al., 2019; Bar Hillel et al., 2014; Chen et al., 2020; Liang et al., 2020; Xing et al., 2018).

Traditional vision-based lane-detection methods rely on hand-crafted low-level features (e.g., colour, gradient, and ridge features) and usually work in a four-step procedure, that is, image pre-processing, feature extraction, line detection and fitting, and post-processing (Bar Hillel et al., 2014; Haris & Glowacz, 2021). Traditional computer vision techniques, for example, inverse perspective mapping (Aly, 2008; B. F. Wang et al., 2014), Hough transform (Berriel et al., 2017; Jiao et al., 2019; Zheng et al., 2018), Gaussian filters (Aly, 2008; Sivaraman & Trivedi, 2013; Y. Wang et al., 2012), and random sample consensus (Aly, 2008; Choi et al., 2018; Du et al., 2018; Guo et al., 2015; Lu et al., 2019), are usually adopted in the four-step procedure. The problems of traditional methods are: (a) hand-crafted features are cumbersome to manage and not always useful, suitable, or powerful; and (b) the detection results are always based on one single image. Thus, the detection accuracies are relatively not high.

During the last decade, with the advancements in deep learning algorithms and computational power, many deep neural network-based methods have been developed for lane detection with good performance. There are generally two dominant approaches (Tabeli et al., 2021b), that is, (1) segmentation-based pipeline (Kim & Park, 2017; Ko et al., 2020; T. Liu et al., 2020; Pan et al., 2018; Zhang et al., 2021; Zou et al., 2020), in which predictions are made on the per-pixel basis, classifying each pixel as either lane or not; (2) the pipeline using row-based prediction (Hou et al., 2020; Qin et al., 2020; Yoo et al., 2020), in which the image is split into a (horizontal) grid, and the model predicts the most probable location to contain a part of a lane marking in each row. Recently, Lizhe Liu et al. (2021) summarised two additional categories of deep learning-based lane-detection methods, that is, the anchor-based approach (Z. Chen et al., 2019; Li et al., 2020; Tabeli et al., 2021b; Xu et al., 2020), which focuses on optimising the line shape by regressing the relative coordinates with the help of predefined anchors, and the parametric prediction-based method, which directly outputs parametric lines expressed by curve equation (R. Liu et al., 2020; Tabeli et al., 2021a). Apart from these dominant approaches, some other less common methods were proposed recently. For instance, Lin et al. (2020) fused the adaptive anchor scheme (designed by formulating a bilinear interpolation algorithm) aided informative feature extraction and object detection into a single deep convolutional neural network (CNN) for lane detection from a top-view perspective. Philion (2019) developed a novel learning-based approach with a fully convolutional model to decode the lane structures directly rather than delegating structure inference to post-processing, plus an effective approach to adapt the model to new contexts by unsupervised transfer learning.

Similar to traditional vision-based lane-detection methods, most available deep learning models utilise only the current image frame to perform the detection. Until very recently, a few studies have explored the combination of CNN and recurrent neural network (RNN) to detect lane markings or simulate autonomous driving using continuous driving scenes (Chen et al., 2020; Zhang et al., 2021; Zou et al., 2020). However, the available methods do not take full advantage of the essential properties of the lane being long continuous solid or dashed line structures. Also, they do not yet make the utmost of the spatial-temporal (ST) information together with correlation and dependencies in the continuous driving frames. Thus, for certain extremely challenging driving scenes, their detection results are still unsatisfactory.

In this study, lane detection is treated as a segmentation task, in which a novel hybrid ST sequence-to-one deep learning architecture is developed for lane detection through a continuous sequence of images in an end-to-end approach. To cope with challenging driving situations, the hybrid model takes multiple continuous frames of an image sequence as inputs, and integrates the single image feature extraction module, the ST feature integration module, together with the encoder-decoder structure to make full use of the ST information in the image sequence. The single image feature extraction module utilises modified common backbone networks with embedded spatial CNN (SCNN; Pan et al., 2018) layers to extract the features in every single image throughout the continuous driving scene. SCNN is powerful in extracting spatial features and relationships in one single image, especially for long continuous shape structures. Next, the extracted features are fed into ST-RNN layers to capture the ST dependencies and correlations among the continuous frames. An encoder-decoder structure is adopted with the encoder consisting of SCNN and several fully convolutional layers to downsample the input image and abstract the features, while the decoder, constructed by CNNs, upsample the abstracted outputs of previous layers to the same size as the input image. With the labelled ground truth of the very last image in the continuous frames, the model training works in an end-to-end way as a supervised learning approach. To train and validate the proposed model on two large-scale open-sourced datasets, that is, tvtLANE (Zou et al., 2020) and TuSimple, a corresponding training strategy has been also developed. To summarise, the main contributions of this study lie in:

1. A hybrid ST sequence-to-one deep neural network architecture integrating the advantages of the encoder-decoder structure, SCNN-embedded single image feature extraction module, and ST-RNN module, is proposed.
2. The proposed model architecture is the first attempt that tries to strengthen both spatial relation feature extraction in every single image frame and ST correlation together with dependencies among continuous image frames for lane detection.
3. The implementation utilised two widely used neural network backbones, that is, UNet (Ronneberger et al., 2015) and SegNet (Badrinarayanan et al., 2017) and included extensive evaluation experiments on commonly used datasets, demonstrating the effectiveness and strength of the proposed model architecture.
4. The proposed model can tackle lane detection in challenging scenes such as curves, dirty roads, serious vehicle occlusions, and so forth, and outperforms all the available state-of-the-art baseline models in most cases with a large margin.
5. Under the proposed architecture, the light version model variant can achieve beyond state-of-the-art performance while using fewer parameters.

2.2 Proposed method

Although many sophisticated methods have been proposed for lane detection, most of the available methods use only one single image, resulting in unsatisfactory performance under some extremely challenging scenarios, for example, dazzle lighting and serious occlusion. This study proposes a novel hybrid ST sequence-to-one deep neural network architecture for lane detection. The architecture was inspired by: (a) the successful precedents of hybrid deep neural network architectures that fuse CNN and RNN to make use of information in continuous multiple frames (Zhang et al., 2021; Zou et al., 2020); (b) the domain prior knowledge that traffic lanes are long continuous shape line structure with strong spatial relationship. The architecture integrates two modules of two distinctive neural networks with complementary merits, that is, SCNN and convolutional long short-term memory (ConvLSTM) neural network, under an end-to-end encoder-decoder structure, to tackle lane detection in challenging driving scenes.

2.2.1 Overview of the proposed model architecture

The proposed deep neural network architecture adopts a sequence-to-one end-to-end encoder-decoder structure as shown in **Figure 2-1**.

Here “sequence-to-one” means that the model gets a sequence of multi-images as input and outputs the detection result of the last image (please note that essentially the model is still utilising sequence-to-sequence neural networks); “end-to-end” means that the learning algorithm goes directly from the input to the desired output, which refers to the lane-detection result in this study, bypassing the intermediate states (Levinson et al., 2011; Neven et al., 2017); the encoder-decoder structure is a modular structure that consists of an encoder network and a decoder network and is often employed in sequence-to-sequence tasks, such as language translation (e.g., Sutskever et al., 2014), and speech recognition (e.g., Wu et al., 2017). Here, the proposed model adopts an encoder CNN with SCNN layers and a decoder CNN using fully convolutional layers. The encoder takes a sequence of continuous image frames, that is, time-series images, as input and abstracts the feature map(s) in smaller sizes. To make use of the prior knowledge that traffic lanes are solid- or dashed-line structures with a continuous shape, one special kind of CNN, that is, SCNN, is adopted after the first CNN hidden layer. With the help of SCNN, spatial features and relationships in every single image will be better extracted. Following this, the extracted feature maps of the continuous frames, constructed in a time-series manner, will be fed to ST-RNN blocks for sequential feature extraction and spatial-temporal information integration. Finally, the decoder network upsamples the abstracted feature maps obtained from the ST-RNN and decodes the content to the original input image size with the detection results. The proposed model architecture is implemented with two backbones, UNet (Ronneberger et al., 2015) and SegNet (Badrinarayanan et al., 2017). Note, in the UNet-based architecture, similar to (Ronneberger et al., 2015), the proposed model employs the skip connection between the encoder and decoder phase by concatenating operation to reuse features and retain information from previous encoder layers for more accurate predictions; while in the SegNet-based networks, at the decoder stage, similar to (Badrinarayanan et al., 2017), the proposed model reuses the pooling indices to capture, store, and make use of the vital boundary information in the encoder feature maps. The detailed network implementation is elaborated in the remaining parts of *Section 2.2*.

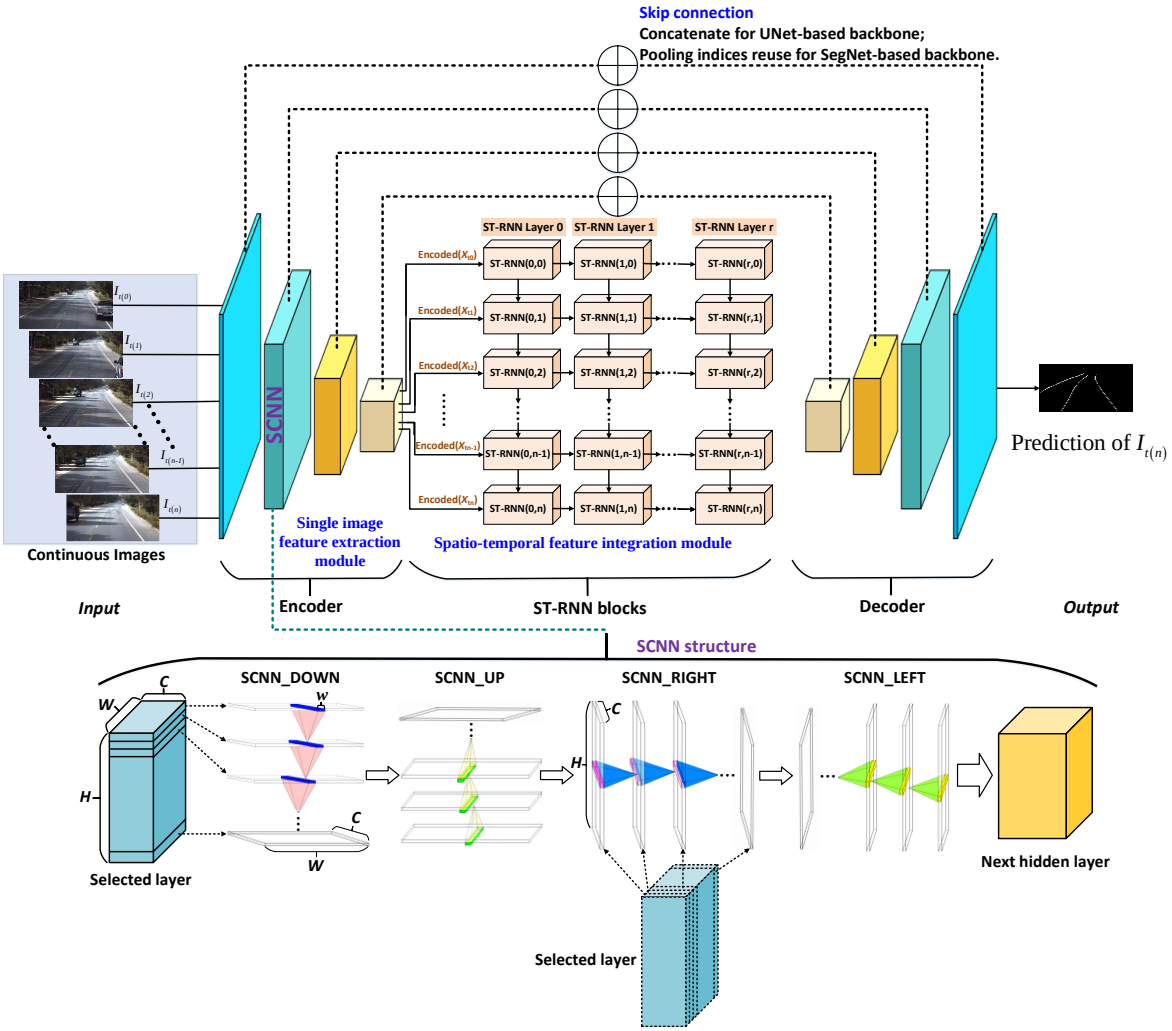


Figure 2-1. The architecture of the proposed model

2.2.2 Network design

1) End-to-end encoder-decoder: Regarding lane detection as an image segmentation problem, the encoder-decoder structure-based neural network can be implemented and trained in an end-to-end way. Inspired by the excellent performance of CNN-based encoder-decoder for image semantic-segmentation tasks in various domains (Badrinarayanan et al., 2017; S. Wang et al., 2020; Yasrab et al., 2017), this study also adopts the "symmetrical" encoder-decoder as the main backbone structure. Convolution and pooling operations are employed to extract and abstract the features in every image in the encoder stage; while in the decoder subset, the inverted convolution and upsampling operation are adopted to grasp the extracted high-order features and construct the outputs layer by layer with regard to the targets. By setting the output target size the same as the input image size, the whole network can work in an end-to-end approach. In the implementation, two widely used backbones, UNet and Seg-Net, are adopted. To better extract and make use of the spatial relations in every image frame, the SCNN layer is introduced in the encoder part of the single image feature extraction module. Furthermore, to excavate and make use of the ST correlations and dependencies among the input continuous image frames, ST-RNN blocks are embedded in the middle of the encoder-decoder networks.

2) SCNN: The SCNN was first proposed by Pan et al. (2018). The "spatial" here means that the specially designed CNN can propagate spatial information via slice-by-slice message passing. The detailed structure of SCNN is demonstrated in the bottom part of **Figure 2-1**.

SCNN can propagate the spatial information in one image through four directions as shown with the suffix "DOWN," "UP," "RIGHT," and "LEFT" in **Figure 2-1**, which denotes downward, upward, rightward, and leftward, respectively. Take the "SCNN_DOWN" module as an example, considering that SCNN is adopted on a three-dimensional tensor of size $C \times W \times H$, wherein the lane-detection task, C, W, and H denote the number of channels, image (or its feature map) width, and height, respectively. For SCNN_D, the input tensor would be split into H slices, and the first slice will then be sent into a convolution operation layer with C kernels of size $C \times w$, in which w is the kernel width. Different from the traditional CNN in which the output of one convolution layer is introduced into the next layer directly, in SCNN_D, the output is added to the next adjacent slice to produce a new slice and iteratively to the next convolution layer, continuing until the last slice in the selected direction is updated. The convolution kernel weights are shared throughout all slices, and the same mechanism works for other directions of SCNNs.

With the above properties, SCNN has demonstrated its strengths in extracting spatial relationships in the image, which makes it suitable for detecting long continuous shape structures, for example, traffic lanes, poles, and walls (Pan et al., 2018). However, using only one image to do the detection, SCNN still could not produce satisfying performances under extremely challenging conditions. And that is why a sequence-to-one architecture with continuous image frames as inputs and ST-RNN blocks to capture the ST correlations in the continuous frames is proposed in this study.

3) ST-RNN module: In this proposed framework, the multiple continuous frames of images are modelled as "image-time-series" inputs. To capture the ST dependencies and correlations among the image-time-series, the ST-RNN module is embedded in the middle of the encoder-decoder structure, which takes over the output extracted features of the encoder as its input and outputs the integrated ST information to the decoder.

Various versions of RNNs have been proposed, for example, LSTM together with its multivariate version, that is, fully connected LSTM (FC-LSTM), and gated recurrent unit (GRU), to tackle time-series data in different application domains. In this study, two state-of-the-art RNN networks, that is, ConvLSTM (Shi et al., 2015) and convolutional GRU (ConvGRU) (Ballas et al., 2016), are employed. These models, considering their abilities in ST feature extraction, generally outperform other traditional RNN models.

A general critical problem for the vanilla RNN model is the gradients vanishing (Hochreiter and Schmidhuber, 1997; Pascanu et al., 2013; Ribeiro, 2020). For this, LSTM introduces memory cells and gates to control the information flow to trap the gradient preventing it from vanishing during the back-propagation. In LSTM, the information of the new time-series inputs will be accumulated to the memory cell $\mathcal{C}_t$ if the input gate i_t is on. In contrast, if the information is not "important", the past cell status $\mathcal{C}_{t-1}$ could be "forgotten" by activating the forget gate f_t. Also, there is the output gate o_t which decides whether the latest cell output $\mathcal{C}_t$ will be propagated to the final state $\mathcal{H}_t$. The traditional FC-LSTM contains too much redundancy for spatial information, which makes it time-consuming and computational-expensive. To address this, the ConvLSTM (Shi et al., 2015) is selected to build the ST-RNN

block of the proposed framework. In ConvLSTM, the convolutional structures and operations are introduced in both the input-to-state and state-to-state transitions to do spatial information encoding, which also alleviates the problem of time- and computation-consuming.

The key formulation of the ConvLSTM is shown by equations (2-1)-(2-5), where $\odot$ denotes the Hadamard product, $*$ denotes the convolution operation, $\sigma(\cdot)$ represents the sigmoid function, and $\tanh(\cdot)$ represents the hyperbolic tangent function; X_t, $\mathcal{C}_t$, and $\mathcal{H}_t$ are the input (i.e., the extracted features from the encoder in the proposed framework), memory cell status, and output at time t; i_t, f_t, and o_t are the function values of the input gate, forget gate, and output gate, respectively; W denotes the weight matrices, whose subscripts indicate the two corresponding variables are connected by this matrix. For instance, W_{xc} is the weight matrix between the input extracted features X_t and the memory cell $\mathcal{C}_t$; "b"s are biases of the gates, e.g., b_i is the input gate's bias.

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * \mathcal{H}_{t-1} + W_{ci} \odot \mathcal{C}_{t-1} + b_i) \tag{2-1}$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * \mathcal{H}_{t-1} + W_{cf} \odot \mathcal{C}_{t-1} + b_f) \tag{2-2}$$

$$\mathcal{C}_t = f_t \odot \mathcal{C}_{t-1} + i_t \odot \tanh(W_{xc} * X_t + W_{hc} * \mathcal{H}_{t-1} + b_c) \tag{2-3}$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * \mathcal{H}_{t-1} + W_{co} \odot \mathcal{C}_t + b_o) \tag{2-4}$$

$$\mathcal{H}_t = o_t \odot \tanh(C_t) \tag{2-5}$$

The ConvGRU (Ballas et al., 2016) further lightens the computational complexity by reducing a gate structure but could perform similarly or slightly better compared with the traditional RNNs or even ConvLSTM. The procedure of computing different gates and hidden states/outputs of ConvGRU is demonstrated with equations (2-6)-(2-9), in which the symbols have the same meaning as described before, while additional z_t and r_t mean the update gate and the reset gate, respectively, plus $\widetilde{\mathcal{H}}$ represents the current candidate hidden representation.

$$z_t = \sigma(W_{zx} * X_t + W_{zh} * \mathcal{H}_{t-1} + b_z) \tag{2-6}$$

$$r_t = \sigma(W_{rx} * X_t + W_{rh} * \mathcal{H}_{t-1} + b_r) \tag{2-7}$$

$$\widetilde{\mathcal{H}}_t = \tanh(W_{ox} * X_t + W_{oh} * (r_t \odot \mathcal{H}_{t-1}) + b_o) \tag{2-8}$$

$$\mathcal{H}_t = z_t\widetilde{\mathcal{H}} + (1\text{-}z_t)\mathcal{H}_{t-1} \tag{2-9}$$

In ConvGRU, there are only two gate structures, i.e., the update gate z_t and the reset gate r_t. It is the update gate z_t that decides how to update the hidden representation when generating the ultimate result of $\mathcal{H}_t$ at the current layer, as shown in equation (2-9). While the reset gate r_t is served to control to what extent the feature information captured in the previous hidden state is supposed to be forgotten through an element-wise multiplication operation when calculating the current candidate hidden representation. From the equations, it is concluded that the information of $\mathcal{H}_t$ mainly comes from $\widetilde{\mathcal{H}}_t$, while $\mathcal{H}_{t-1}$ as the previous hidden-state representation also contributes to the process of computing the final representation of $\mathcal{H}_t$; thus, the temporal dependencies are captured.

In practice, both ConvLSTM and ConvGRU with different numbers of hidden layers were employed to serve as the ST-RNN module in the proposed architecture, and the corresponding performances were evaluated, respectively. To be specific, in the proposed network, the input

and the output sizes of the ST-RNN block are equivalent to the feature map size extracted through the encoder, which are 8 × 16 and 4 × 8 for the UNet-based and SegNet-based backbone, respectively. The convolutional kernel size in ConvLSTM and ConvGRU is 3 × 3, and the dimension of each hidden layer is 512. The detailed implementations are described in the following section.

2.2.3 Detailed implementation

1) Network design details: The proposed spatial-temporal sequence-to-one neural network was developed for the lane detection task with K (in this study $K = 5$ if not specified) continuous image frames as inputs. The image frames were first fed into the encoder for feature extraction and abstraction. Different from the normal CNN-based encoder, the SCNN layer was utilised to effectively extract the spatial relationships within every image. Different locations of the SCNN layer were tested, i.e., embedding the SCNN layer after the first hidden convolutional layer or at the very beginning. The outputs of the encoder network were modelled in a time-series manner and fed into the ST-RNN blocks (i.e., ConvLSTM or ConvGRU layers) to further extract more useful and accurate features, especially the spatial-temporal dependencies and correlations among different image frames. In short, the encoder network is primarily responsible for spatial feature extraction and abstraction transforming input images into specified feature maps, while the ST-RNN blocks accept the extracted features from the continuous image frames in a time-series manner to capture the spatial-temporal dependencies.

The outputs of the ST-RNN blocks were then transferred into the decoder network that adopts deconvolution and upsampling operations to highlight and make full use of the features and rebuild the target to the original size of the input image. Note that there is the skip concatenate connection (for UNet-based architecture) or pooling indices reusing (for SegNet-based architecture) between the encoder and decoder to reuse the retained features from previous encoder layers for more accurate predictions at the decoder phase. After the decoder phase, the lane detection result is obtained as an image of the equivalent size to the input image frame. With the labelled ground truth and the help of the encoder-decoder structure, the proposed model can be trained and implemented in an end-to-end way. The detailed input, output sizes, and parameters of the layers in the entire neural network are listed in Appendix **Table 2-A1** and **Table 2-A2**.

For both SegNet-based and UNet-based implementations, two types of RNN layers, i.e., ConvLSTM and ConvGRU, were tested to serve as the ST-RNN block. Besides, the ST-RNN blocks were tested with 1 hidden layer and 2 hidden layers, respectively. So there are four variants in the proposed SegNet-based models, i.e., SCNN_SegNet_ConvGRU1, SCNN_SegNet_ConvGRU2, SCNN_SegNet_ConvLSTM1, and SCNN_SegNet_ConvLSTM2. SCNN_SegNet_ConvGRU1 means the model is using SegNet as the backbone with SCNN layer embedded encoder, and 1 hidden layer of ConvGRU as the ST-RNN block. This naming rule applies to the other 3 variants. Also, there are four variants of the proposed UNet-based models, with a similar naming rule.

In the proposed models with UNet as the backbone, the number of kernels used in the last convolutional block of the encoder part differs from the original UNet's settings. Here, the number of output kernels (channels) of the last convolutional block in the proposed encoder does not double its input kernels, which applies to all the previous convolutional blocks. This

is done, similar to (Zou et al., 2020), to better connect the output of the encoder with the ST-RNN block (ConvLSTM or ConvGRU layers). To do so, the parameters of the full-connection layer are designed to be quadrupled while the side lengths of the feature maps are reduced to half, at the same time, the number of kernels remains unchanged. This strategy also somewhat contributes to reducing the parameter size of the whole network.

A modified light version of UNet (UNetLight) was also tested to serve as the network backbone to reduce the total parameter size, increase the model's ability to operate in real time, and further verify the proposed network architecture's effectiveness. The UNetLight has a similar network design to the demonstration in **Table 2-A2**. The only difference is that all the numbers of kernels in the ConvBlocks are reduced to half, except for the *Input* in *In_ConvBlock* (with the input channel of 3 unchanged) and the *Output* in *Out_ConvBlock* (with the output channel of 2 unchanged). To save space, the parameter settings of the UNetLight-based implementation will not be illustrated.

2) Loss function: Since lane detection is modelled as a segmentation task and a pixel-wise binary classification problem, cross-entropy is a suitable candidate to serve as the loss function. However, because the pixels classified to be lanes are always quite less than those classified to be the background (meaning that it is an imbalanced binary classification and discriminative segmentation task), in the implementation, the loss was built upon the weighted cross-entropy. The adopted loss function, as the standard weighted binary cross-entropy function, is given in equation (2-10),

$$Loss = -\frac{1}{S}\sum_{i=1}^{S}\left[\omega * y_i * log\left(h_\theta(x_i)\right) + (1\text{-}y_i) * log\left(1 - h_\theta(x_i)\right)\right] \quad (2\text{-}10)$$

where S is the number of training examples, ω stands for the weight which is set according to the ratio between the total lane pixel quantities and non-lane pixel quantities throughout the whole training set, y_i is the true target label for the training example i, x_i is the input for the training example i, and h_θ stands for the model with neural network weights θ.

3) Training details: The proposed neural networks with different variants, together with the baseline models were trained on the Dutch high-performance supercomputer clusters, Cartesius and Lisa, using 4 Titan RTX GPUs with the data parallel mechanism in PyTorch. The input image size was set as 128 × 256 to reduce the computational payload. The batch size was set to be as large as possible (e.g., 64 for UNet-based network architecture, 100 for SegNet-based ones, and 136 for UNetLight-based ones), and the learning rate was initially set to 0.03. The RAdam optimiser (Liyuan Liu et al., 2019) was first used in this work for training the model at the beginning. At the later stage, when the training accuracy was beyond 95%, the optimiser was switched to the Stochastic Gradient Descent (SGD) (Bottou, 2010) optimiser with decay. With the labelled ground truth, the models were trained through iteratively updating the parameters in the weight matrices and the losses on the basis of the deviation between outputs of the proposed neural network and the ground truth using the backpropagation mechanism. To speed up the training process, the pre-trained weights of SegNet and UNet on ImageNet (Deng et al., 2009) were adopted.

2.3 Experiments and results

Extensive experiments were carried out to inspect and verify the accuracy, effectiveness, and robustness of the proposed lane-detection model using two large-scale open-sourced datasets. The proposed models were evaluated on different driving scenes and were compared with several state-of-the-art baseline lane-detection methods, which also employ deep learning, for example, UNet (Ronneberger et al., 2015), Seg-Net (Badrinarayanan et al., 2017), SCNN (Pan et al., 2018), LaneNet (Neven et al., 2018), UNet_ConvLSTM (Zou et al., 2020), and SegNet_ConvLSTM (Zou et al., 2020).

2.3.1 Datasets

1) tvtLANE training set: To verify the proposed model performance, the tvtLANE dataset (Zou et al., 2020) based upon the TuSimple lane marking challenge dataset, was first utilised for training, validating, and testing. The original dataset of the TuSimple lane marking challenge includes 3,626 clips of training and 2,782 clips of testing, which are collected under various weather conditions and during different periods. In each clip, there are 20 continuous frames saved in the same folder. In each clip, only the lane marking lines of the very last frame, i.e., the 20^{th} frame, are labelled with the ground truth officially. Zou et al. (2020) additionally labelled every 13^{th} image in each clip and added their own collected lane dataset, which includes 1,148 sequences of rural driving scenes collected in China. This immensely expanded the variety of the road and driving conditions since the original TuSimple dataset only covers highway driving conditions. *K* continuous frames of each clip are used as the inputs, with the ground truth of the labelled 13^{th} or 20^{th} frame to train the models.

To further augment the training dataset, crop, flip, and rotation operations were employed. Thus, a total number of $(3{,}626 + 1{,}148) \times 4 = 19{,}096$ continuous sequences were produced, in which 38,192 images are labelled with ground truth. To adapt to different driving speeds, the input image sequences were sampled at 3 strides with a frame interval of 1, 2, or 3, respectively. Then, 3 sampling methods were employed to construct the training samples regarding the labelled 13^{th} and 20^{th} frames in each sequence, as demonstrated in **Table 2-1**.

2) tvtLANE testing set: Two different datasets were used for testing, i.e., testset #1 (normal) and testset #2 (challenging), which are also formatted with 5 continuous images as the input to detect the lane markings in the very last frame with the labelled ground truth. To be specific, testset #1 is built upon the original TuSimple test set for normal driving scene testing; while testset #2 is constructed with 12 challenging driving situations, especially used for robustness evaluation. The detailed descriptions of the trainset and testset in tvtLANE are illustrated in **Table 2-1**, with examples shown in **Figure 2-2**.

2.3.2 Qualitative evaluation

Qualitative evaluation with the visualisation of the lane detection results is the most intuitive approach to compare and evaluate the properties of different models, and it helps to find insights regarding their pros and cons.

1) tvtLANE Testset #1: normal situations

Samples of the lane detection results on tvtLANE testset #1 of the proposed models and other

Table 2-1. Trainset and testset in tvtLANE

Trainset				
Subset				**Labelled Images Num**
Original TuSimple Dataset (Highway)				7,252
Zou et al. (2020) added (Rural Road)				2,296
Sample Methods				
Labelled Ground Truth			**Sample Stride**	**Train Sample Frames**
13^{th}			3	1^{st}, 4^{th}, 7^{th}, 10^{th}, 13^{th}
			2	5^{th}, 7^{th}, 9^{th}, 11^{th}, 13^{th}
			1	9^{th}, 10^{th}, 11^{th}, 12^{th}, 13^{th}
20^{th}			3	8^{th}, 11^{th}, 14^{th}, 17^{th}, 20^{th}
			2	12^{th}, 14^{th}, 16^{th}, 18^{th},20^{th}
			1	16^{th}, 17^{th}, 18^{th}, 19^{th},20^{th}
Testset				
Subset	**Labelled Images Num**	**Labelled Ground Truth**	**Sample Stride**	**Test Sample Frames**
Testset #1 Normal	540	13^{th}	1	9^{th}, 10^{th}, 11^{th}, 12^{th}, 13^{th}
		20^{th}	1	16^{th}, 17^{th}, 18^{th},19^{th},20^{th}
Testset #2 Challenging	728	All	1	1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}
				2^{nd}, 3^{rd}, 4^{th}, 5^{th}, 6^{th}
				3^{rd}, 4^{th}, 5^{th}, 6^{th}, 7^{th}
				…

Figure 2-2. Sample data in trainset and testset

(a) original TuSimple dataset (Highway), (b) Zou et al., (2020) added Rural Road situations, (c) Testset #1 Normal situations, and (d) Testset #2 Challenging situations. In each row, the first five images are the input image sequence the last image is the labelled ground truth

state-of-the-art models are demonstrated in **Figure 2-3 (1)**. All these results are without post-processing.

In general, good lane detection should include the following 5 properties:

- The number of lines needs to be predicted correctly. A wrong detection or a misprediction might cause the automated vehicles to consider unsafe or unreachable areas as drivable

areas resulting in potential accidents. As illustrated in the 1st and 2nd columns in **Figure 2-3 (1)**, the proposed models can identify the correct number of lane lines, while the baseline models, especially the ones using a single image, somewhat cannot detect the correct number of lines compared with the ground truth.

- The positions of each lane marking line should be predicted precisely in accordance with the ground truth. As illustrated in **Figure 2-3 (1)**, the proposed models in row (j) with the model named by SCNN_SegNet_ConvLSTM2 and row (n) with the model named by SCNN_UNet_ConvLSTM2, could deliver better lane location predictions with thinner lines, compared with the baseline models. Superior to scattering points around, thinner predicted lane lines indicate a more precise model prediction of the lane position.

- The predicted lane lines should not merge or be broken. As illustrated in the 1st, 2nd, 6th, 7th, and 8th columns of **Figure 2-3 (1)**, some baseline models' output lane lines either merge at the far end or break the continuity with dashed lines. The proposed models perform slightly better, although in a few cases, the lines are also discontinuous.

- The lanes should be predicted correctly even at the boundary of the image. As can be found in **Figure 2-3 (1)**, some baseline models, e.g., row (c), (d), and (e), run across difficulties at the top boundary of the image with merge lanes on the top. This also accords with the aforementioned property.

- The lane detection models should deliver accurate predictions under different driving scenes, even under some challenging situations. For example, in the 2nd, 3rd, 5th, and 7th columns of **Figure 2-3 (1)**, vehicles are occluding the lanes. A good lane detection model should be able to handle these. The proposed models perform well under these slightly challenging cases; more challenging situations are further discussed later.

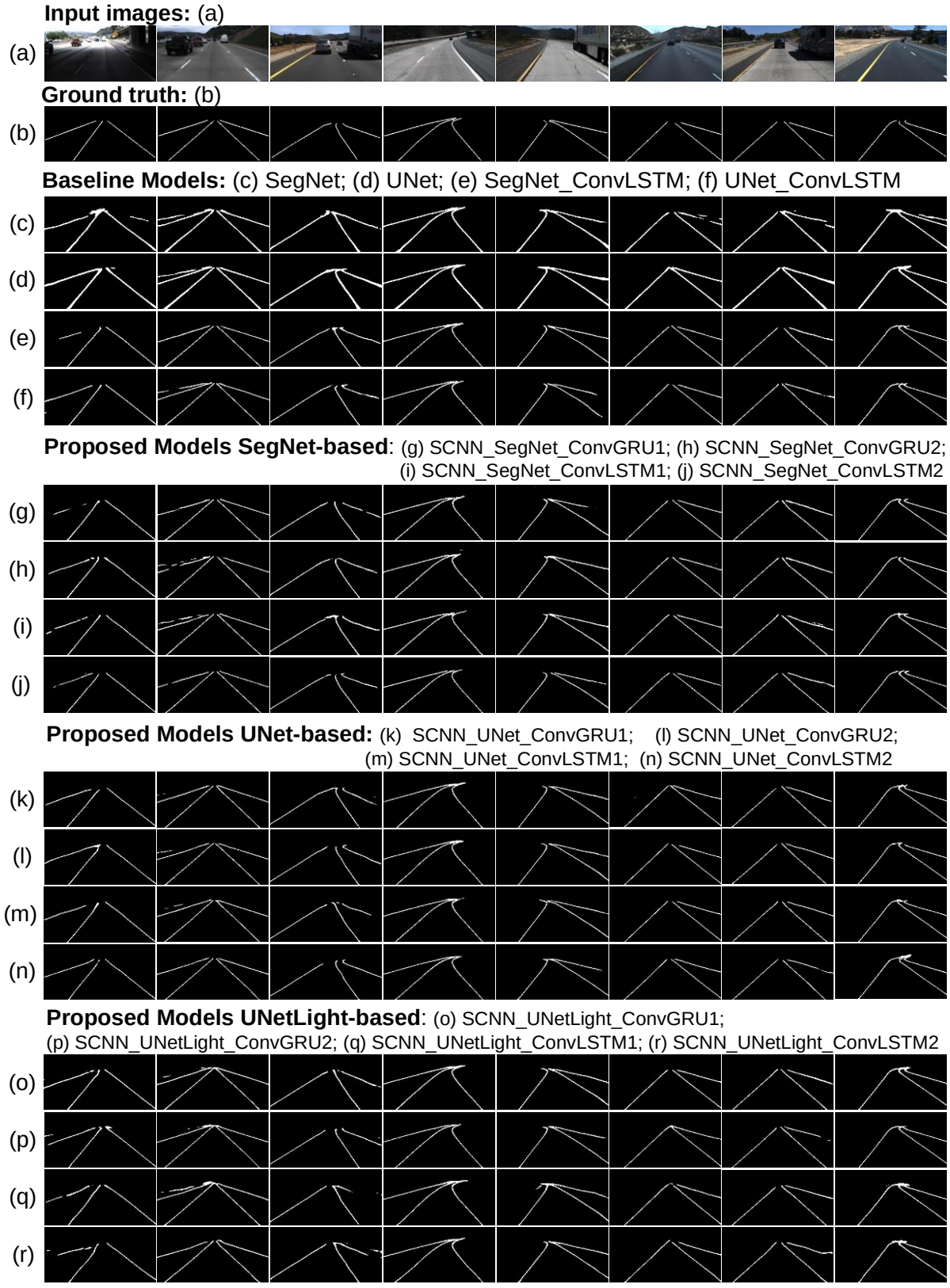
2) tvtLANE testset *#2: 12 challenging driving cases*

Figure 2-3 (2) shows the comparison of the proposed models with the baseline models under some extremely challenging driving scenes in the tvtLANE testset #2. All the results are not post-processed. These challenging scenes cover wide situations including serious vehicle occlusion, bad lighting conditions (e.g., shadow, dim), tunnel situations, and dirt road conditions. In some extremely challenging cases, the lanes are totally occluded by vehicles, other objects, and/or shadows, which could be very difficult even for humans to do the detection.

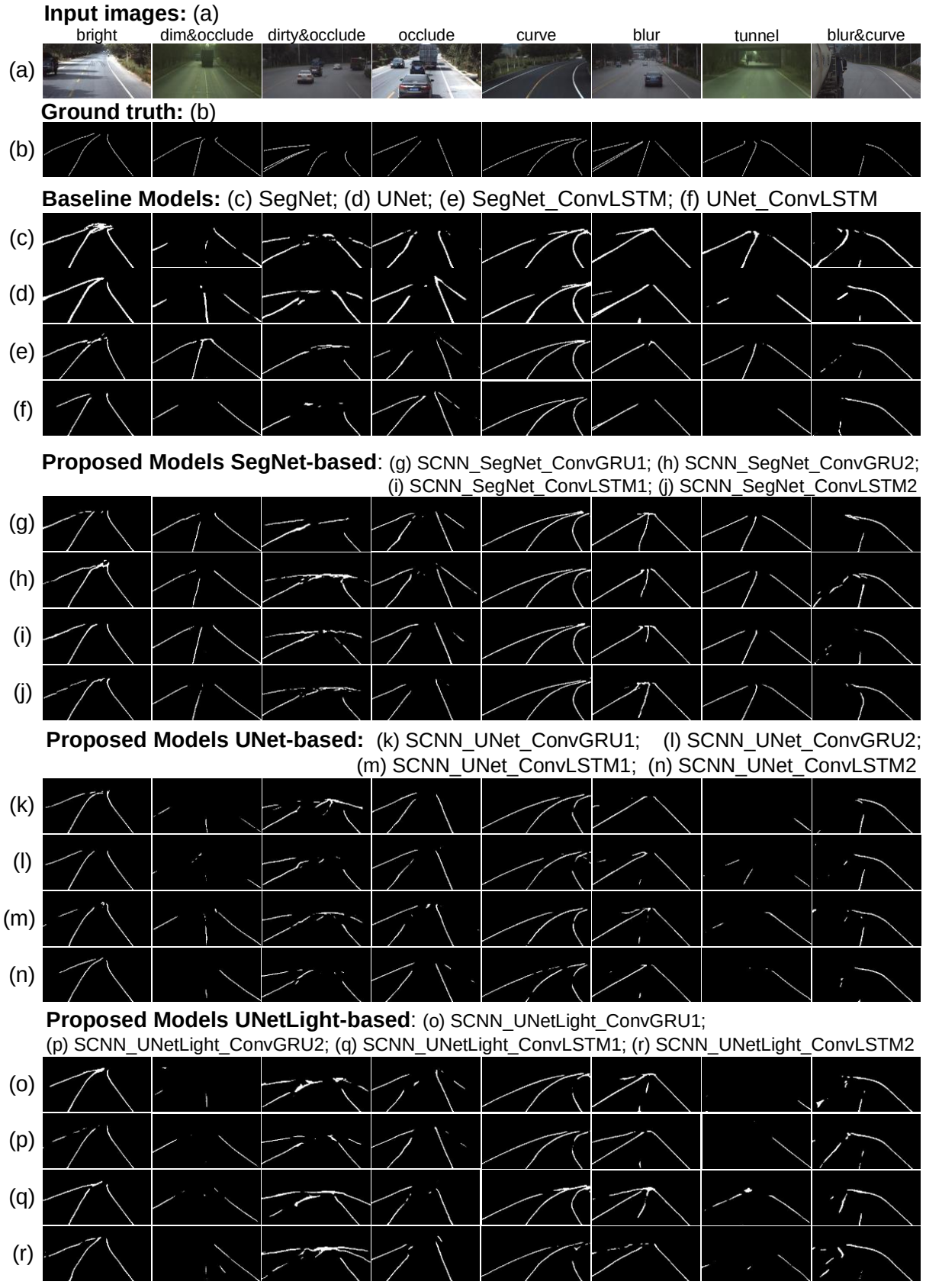
As can be observed in **Figure 2-3 (2)**, although all the baseline models fail in these challenging cases, the proposed models, especially the one named SCNN_SegNet_ConvLSTM2 illustrated in row (k), could still deliver good predictions in almost every situation listed in **Figure 2-3 (2)**. The only flaw is that in the 3rd column, where vehicle occlusion and blurred road conditions happen simultaneously, the proposed models also find it hard to predict precisely. With the results in the 4th, 7th, and 8th columns, the robustness of SCNN_SegNet_ConvLSTM2's property in detecting the correct number of lane lines is further verified, especially, one can observe in the 4th column, where almost all the other models are defeated, SCNN_SegNet_ConvLSTM2 can still predict the correct number of lanes.

Furthermore, it should be noticed that correct lane location predictions in these challenging situations are of vital importance for safe driving. For example, regarding the situation in the last column where a heavy vehicle totally shadows the field of vision on the left side, it will be

very dangerous if the automated vehicle is driving according to the lane detection results demonstrated in the 3rd to 5th rows.



(1) Visualisation of the lane-detection results on tvtLANE testset #1 (normal situations)



(2) Visualisation of the lane-detection results on tvtLANE testset #2 (challenging situations)

Figure 2-3. Qualitative evaluation: visualisation of the lane-detection results on (1) tvtLANE testset #1 and (2) tvtLANE testset #2

2.3.3 Quantitative evaluation

1) Evaluation metrics: This subsection examines the proposed models' properties regarding quantitative evaluations. When treated as a pixel-wise classification task, accuracy must be the simplest criterion for the performance evaluation of lane detection (Zou et al., 2017), which represents the overall classification performance in terms of correctly classified pixels, indicated in equation (2-11).

$$\text{Accuracy} = \frac{Truly\ Classified\ Pixels}{Total\ Number\ of\ Pixels} \tag{2-11}$$

However, since it is an imbalanced binary classification problem, where the lane pixels are far less than the background pixels, using only accuracy to evaluate the model is not suitable. Thus, Precision, Recall, and F-measure, illustrated by equations (2-12)-(2-14), are commonly employed.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \tag{2-12}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \tag{2-13}$$

$$\text{F-measure} = (1+\beta^2)\frac{\text{Precision}*\text{Recall}}{\beta^2\text{Precision+Recall}} \tag{2-14}$$

In the above equation, true positive indicates the number of image pixels that are lane marking and are correctly identified; false positive means the number of image pixels that are background but are wrongly classified as lane markings; false negative stands for the number of image pixels which are lane marking but are wrongly classified as the background.

Specifically, this study chooses $\beta = 1$, which corresponds to the F1-measure (harmonic mean) shown in equation (2-15).

$$\text{F1-measure} = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \tag{2-15}$$

The F1-measure, which balances Precision and Recall, is always selected as the main benchmark for model evaluation, e.g., (Lizhe Liu et al., 2021; Pan et al., 2018; Xu et al., 2020; Zhang et al., 2021; Zou et al., 2020).

Furthermore, the model parameter size, i.e., Params (M), together with the multiply-accumulate (MAC) operations, i.e., MACs (G), are provided as indicators of the model complexity. The two indicators are commonly used to estimate models' computational complexities and real-time capabilities.

2) Performance and comparisons on tvtLANE testset #1(normal situations)

As shown in **Table 2-2**, the proposed model of SCNN_UNet_ConvLSTM2, performs the best when evaluating on tvtLANE testset #1, with the highest Accuracy and F1-measure, while the proposed model of SCNN_SegNet_ConvLSTM2 delivers the best Precision.

Incorporating the quantitative evaluation with the qualitative evaluation, it could be easily interpreted that the highest Precision, Accuracy, and F1-measure are mainly derived from (i) the correct lane number, (ii) the accurate lane position, (iii) the sound continuity in the detected lanes, and (iv) the thinness of the predicted lanes with less blurriness, which accords with (ii).

The correct prediction directly reduces the number of False Positives, and a good Precision contributes to better Accuracy and F1-measure. Considering the structure of the proposed model architecture, a further explanation of the high F1-measure, Accuracy, and Precision can be explained as follows:

Firstly, the SCNN layer embedded in the encoder equips the proposed model with better information extracting ability regarding the low-level features and spatial relations in each image.

Secondly, the ST-RNN blocks, i.e., ConvLSTM / ConvGRU layers, can effectively capture the temporal dependencies among the continuous image frames, which could be very helpful for challenging situations where the lanes are shadowed or covered by other objects in the current frame.

Finally, the proposed architecture could make the best of the spatial-temporal information among the processed K continuous frames by regulating the weights of the convolutional kernels within the SCNN and ConvLSTM / ConvGRU layers.

All in all, with the proposed architecture, the proposed model tries to not only strengthen feature extraction regarding spatial relation in one image frame but also the spatial-temporal correlation and dependencies among image frames for lane detection.

Table 2-2. Model performance comparison on tvtLANE testset #1 (normal situations)

Model		Test Acc (%)	Precision	Recall	F1-measure	MACs (G)	Params (M)
Using single image as input	**Baseline Models**						
	UNet	96.54	0.790	**0.985**	0.877	**15.5**	13.4
	SegNet	96.93	0.796	0.962	0.871	50.2	29.4
	SCNN*	96.79	0.654	0.808	0.722	77.7	19.2
	LaneNet*	97.94	0.875	0.927	0.901	44.5	19.7
Using continuous images sequence as inputs	SegNet_ConvLSTM**	97.92	0.874	0.931	0.901	217.0	67.2
	UNet_ConvLSTM**	98.00	0.857	0.958	0.904	69.0	51.1
	Proposed Models (SegNet-Based)						
	SCNN_SegNet_ConvGRU1	98.00	0.878	0.935	0.905	219.2	43.7
	SCNN_SegNet_ConvGRU2	98.05	0.888	0.918	0.903	221.5	57.9
	SCNN_SegNet_ConvLSTM1	98.01	0.881	0.935	0.907	220.0	48.5
	SCNN_SegNet_ConvLSTM2	98.07	**0.893**	0.928	0.910	223.0	67.3
	Proposed Models (UNet-Based)						
	SCNN_UNet_ConvGRU1	98.13	0.878	0.957	0.916	77.9	27.7
	SCNN_UNet_ConvGRU2	**98.19**	0.887	0.950	0.917	87.0	41.9
	SCNN_UNet_ConvLSTM1	98.18	0.886	0.948	0.916	81.0	32.4
	SCNN_UNet_ConvLSTM2	**98.19**	0.889	0.950	**0.918**	93.0	51.3
	Proposed Models (Light Version UNet-Based)						
	SCNN_UNetLight_ConvGRU1	97.83	0.850	0.960	0.902	19.6	**6.9**
	SCNN_UNetLight_ConvGRU2	98.01	0.863	0.950	0.905	21.9	10.5
	SCNN_UNetLight_ConvLSTM1	97.71	0.830	0.950	0.886	20.4	8.1
	SCNN_UNetLight_ConvLSTM2	97.76	0.840	0.953	0.893	23.4	12.8

** Results reported in (Zhang et al., 2021).*

*** There are two hidden layers of ConvLSTM in SegNet_ConvLSTM and UNet_ConvLSTM.*

Looking at the main metric, F1-measure, it is demonstrated that increasing only Precision or only Recall will not improve the F1-measure. Although the baseline models of UNet, SegNet, and SegNet_ConvLSTM get better Recalls, they do not deliver good F1-measure since their Precisions are much lower than the proposed model of SCNN_SegNet_ConvLSTM2 or SCNN_UNet_ConvLSTM2. Regarding the good Recall of UNet and SegNet, it could be speculated from the qualitative evaluation, where one can find that UNet and SegNet tend to produce thicker lane lines. With thicker lines and blurry areas, the two models can somehow reduce the False Negatives, which will contribute to better Recall. This also demonstrates that Recall and Precision antagonise each other, which further proves that F1-measure should be a more reasonable evaluation measure compared with Precision and Recall.

3) Performance and comparisons on tvtLANE testset #2 (challenging situations)

To further evaluate the proposed models' performance and verify the models' robustness, the models were evaluated on a brand-new dataset, i.e., the tvtLANE testset #2. As introduced in *2.3.1 Datasets,* tvtLANE testset #2 includes 728 images in highway, urban, and rural driving scenes. These challenging driving scenes' data were obtained by data recorders at various locations, outside and inside the car's front windshield under different road and weather conditions. Testset #2 is a challenging and comprehensive dataset for model evaluation, from which some cases would be difficult enough for humans to do the correct detection.

Table 2-3 demonstrates the model performance comparison on the 12 types of challenging scenes in tvtLANE testset #2. Following the results and discussions in *2)* Performance and comparisons on tvtLANE testset #1 (normal situations), here **Table 2-3** provides the Precision and F1-measure for the evaluation reference.

As indicated by the bold numbers, the proposed model, SCNN_SegNet_ConvLSTM2, results in the best F1-measure at the overall level and in more situations, while the UNet_ConvLSTM results in the best Precision at the overall level and in more situations. Incorporating the qualitative evaluation in **Figure 2-3 (2)**, it is shown that UNet_ConvLSTM tends to not classify pixels into lane lines for uncertain areas under some challenging situations (e.g., the 2nd and 7th columns in **Figure 2-3 (2)**). This might be the reason for its obtaining better Precision. To further confirm this speculation, **Figure 2-4** compares the lane detection results of SCNN_SegNet_ConvLSTM2 and UNet_ConvLSTM under challenging situations, *8-blur&curve* and *10-shadow-dark*, where UNet_ConvLSTM delivers very good Precisions.

As illustrated in **Figure 2-4**, truly UNet_ConvLSTM tries not to classify pixels into lane lines under uncertain areas as much as possible. This leads to fewer False Negatives which helps for raising a better Precision. However, in real application scenarios, this is not wise and not acceptable. On the contrary, the proposed model SCNN_SegNet_ConvLSTM2 tries to make tough but valuable detections classifying candidate points into lane lines in the challenging uncertain areas with dirt, dark road conditions, and/or vehicle occlusions. This may lead to more False Negatives and a worse Precision but is praiseworthy. These analyses further demonstrate that F1-measure is a better measure compared with Precision. Finally, it can be concluded that the proposed model, SCNN_SegNet_ConvLSTM2, delivers the best performance on the challenging tvtLANE testset #2, which verified the proposed model architecture's robustness.

To sum up, the proposed model architecture demonstrates its effectiveness in both normal and challenging driving scenes, with the UNet-based model, SCNN_UNet_ConvLSTM2, beating

the baseline models with a large margin on normal situations, while the SegNet-based model, SCNN_SegNet_ConvLSTM2, performs the best, handling almost all the challenging driving scenes. The finding that, compared with UNet-based models, SegNet-based neural network models are more robust in coping with challenging driving environments accords with the results in (Zou et al., 2020).

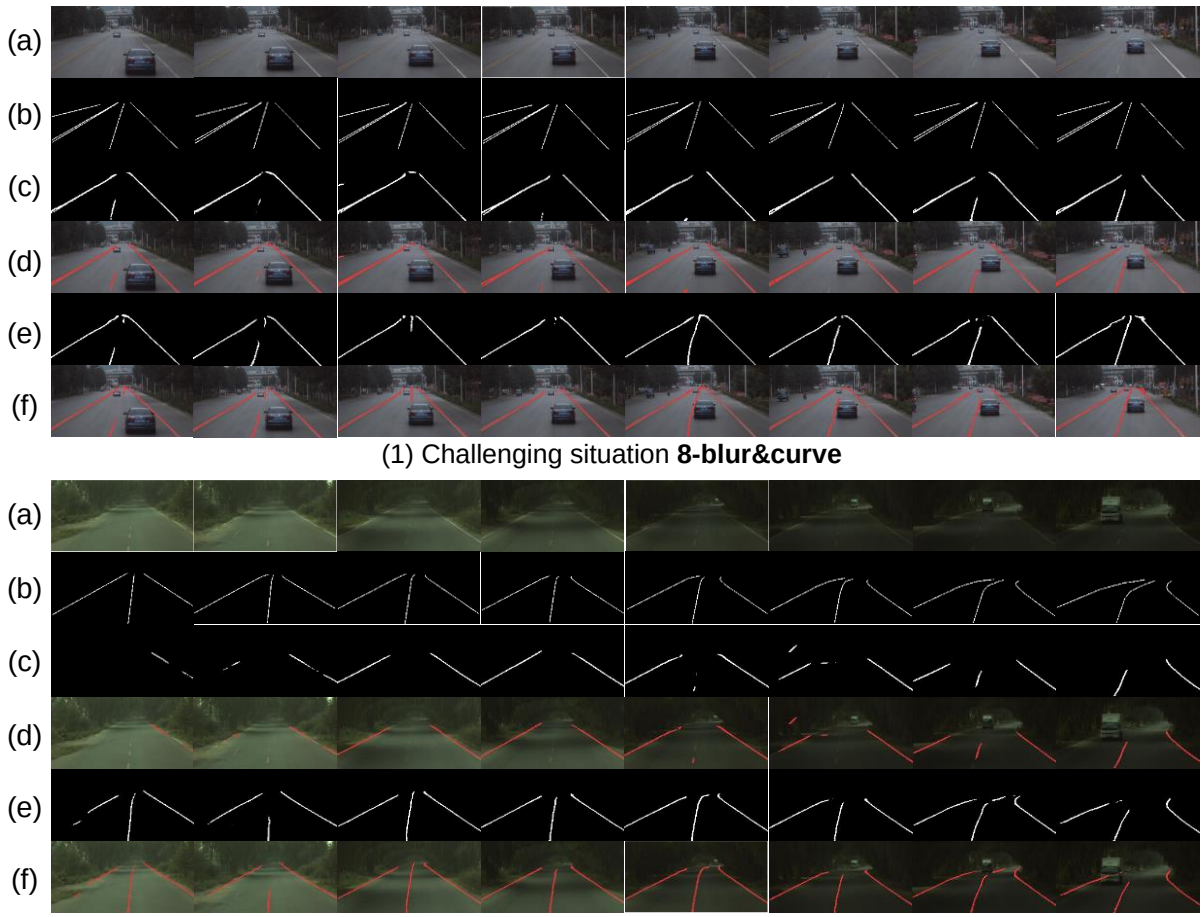


Figure 2-4. Visual comparison of the lane-detection results on challenging driving situations for UNet_ConvLSTM and the proposed model SCNN_SegNet_ConvLSTM2

All the results are not post-processed.

(a) Input images. (b) Ground truth. (c) Detection results of UNet_ConvLSTM. (d) Detection results of UNet_ConvLSTM overlapping on the original images. (e) Detection results of SCNN_SegNet_ConvLSTM2. (f) Detection results of SCNN_SegNet_ConvLSTM2 overlapping on the original images.

*The upper part (1) is for challenging situation **8-blur&curve**, while the down part (2) is for situation **10-shadow-dark**.*

Table 2-3. Model performance comparison on tvtLANE testset #2 (12 types of challenging scenes)

Precision

Model \ Challenging Scenes	1-curve & occlude	2-shadow - bright	3-bright	4-occlude	5-curve	6-dirty & occlude	7-urban	8-blur & curve	9-blur	10-shadow - dark	11-tunnel	12-dim & occlude	overall
UNet	0.7018	0.7441	0.6717	0.6517	0.7443	0.3994	0.4422	0.7612	0.8523	0.7881	0.7009	0.5968	0.6754
SegNet	0.6810	0.7067	0.5987	0.5132	0.7738	0.2431	0.3195	0.6642	0.7091	0.7499	0.6225	0.6463	0.6080
UNet_ConvLSTM	0.7591	0.8292	0.7971	0.6509	0.8845	0.4513	0.5148	**0.8290**	**0.9484**	**0.9358**	0.7926	0.8402	**0.7784**
SegNet_ConvLSTM	0.8176	0.8020	0.7200	0.6688	0.8645	0.5724	0.4861	0.7988	0.8378	0.8832	0.7733	0.8052	0.7563
SCNN_SegNet_ConvGRU1	0.8107	0.7951	0.7225	0.6830	0.8503	0.4640	0.5071	0.6699	0.8481	0.8994	0.7804	0.8429	0.7477
SCNN_SegNet_ConvGRU2	0.7952	0.8087	0.7770	0.6444	0.8689	0.5067	0.5171	0.7147	0.8423	0.8744	0.7979	**0.8757**	0.7572
SCNN_SegNet_ConvLSTM1	0.7945	0.8078	0.7600	0.6417	0.8525	0.5252	0.3686	0.7582	0.7715	0.8702	0.7778	0.8517	0.7348
SCNN_SegNet_ConvLSTM2	0.8326	0.7497	0.7470	0.7369	0.8647	**0.6196**	0.4333	0.7371	0.8566	0.9125	0.8153	0.8466	0.7673
SCNN_UNet_ConvGRU1	0.8492	0.8306	0.8163	**0.7845**	0.8819	0.4025	0.4493	0.7378	0.8291	0.8928	**0.8198**	0.8040	0.7639
SCNN_UNet_ConvGRU2	**0.8678**	0.7873	**0.8548**	0.7654	0.8805	0.5319	0.4735	0.8064	0.8765	0.8431	0.7112	0.7388	0.7640
SCNN_UNet_ConvLSTM1	0.8602	0.7844	0.8119	0.7807	**0.8871**	0.4066	0.4652	0.7445	0.8321	0.8972	0.7507	0.7068	0.7531
SCNN_UNet_ConvLSTM2	0.8182	**0.8362**	0.8189	0.7359	0.8365	0.5872	**0.5377**	0.8046	0.8770	0.8722	0.7952	0.7817	**0.7784**
SCNN_UNetLight_ConvGRU1	0.8212	0.7454	0.7189	0.6996	0.8521	0.3499	0.3999	0.7851	0.7282	0.8686	0.6940	0.6289	0.7011
SCNN_UNetLight_ConvGRU2	0.8147	0.8349	0.7390	0.7004	0.8591	0.4039	0.3360	0.6811	0.8300	0.8533	0.8125	0.7996	0.7238
SCNN_UNetLight_ConvLSTM1	0.7222	0.7450	0.6533	0.6203	0.8039	0.2635	0.2716	0.7341	0.7546	0.7319	0.6298	0.7406	0.6377
SCNN_UNetLight_ConvLSTM2	0.7618	0.7416	0.7067	0.6537	0.8096	0.1921	0.2639	0.6857	0.6830	0.6931	0.6391	0.6022	0.6190

F1-measure

Model \ Challenging Scenes	1-curve & occlude	2-shadow - bright	3-bright	4-occlude	5-curve	6-dirty & occlude	7-urban	8-blur & curve	9-blur	10-shadow - dark	11-tunnel	12-dim & occlude	overall
UNet	0.8200	0.8408	0.7946	0.7337	0.7827	0.3698	0.5658	0.8147	0.7715	0.6619	0.5740	0.4646	0.6985
SegNet	0.8042	0.7900	0.7023	0.6127	0.8639	0.2110	0.4267	0.7396	0.7286	0.7675	0.6935	0.5822	0.6727
UNet_ConvLSTM	0.8465	**0.8891**	**0.8411**	0.7245	0.8662	0.2417	0.5682	0.8323	0.7852	0.6404	0.4741	0.5718	0.7143
SegNet_ConvLSTM	0.8852	0.8544	0.7688	0.6878	0.9069	0.4128	0.5317	0.7873	0.7575	0.8503	0.7865	0.7947	0.7609
SCNN_SegNet_ConvGRU1	0.8821	0.8626	0.7734	0.7185	0.9039	0.3027	0.5288	0.7229	**0.7866**	0.8658	0.7759	0.7763	0.7547
SCNN_SegNet_ConvGRU2	0.8710	0.8630	0.8094	0.6989	0.9005	0.3963	0.5497	0.7470	0.7637	0.8525	0.7798	0.7396	0.7591
SCNN_SegNet_ConvLSTM1	0.8768	0.8801	0.8185	0.7166	0.9083	0.3750	0.4516	0.7806	0.7320	0.8622	**0.8029**	**0.8245**	0.7629
SCNN_SegNet_ConvLSTM2	0.8956	0.8237	0.7909	0.7468	**0.9108**	**0.4398**	0.4858	0.7379	0.7546	**0.8729**	0.7963	0.8074	**0.7666**
SCNN_UNet_ConvGRU1	0.8608	0.8745	0.8393	**0.7802**	0.9005	0.3181	0.5143	0.7833	0.7567	0.5554	0.3503	0.3703	0.6839
SCNN_UNet_ConvGRU2	0.8706	0.8556	0.8304	0.7647	0.8532	0.3515	0.5253	**0.8345**	0.7399	0.5405	0.3567	0.2855	0.6722
SCNN_UNet_ConvLSTM1	**0.8971**	0.8493	0.8234	0.7633	0.8997	0.3054	0.5307	0.7424	0.7436	0.6243	0.5568	0.5366	0.6992
SCNN_UNet_ConvLSTM2	0.8670	0.8866	0.8405	0.7565	0.7955	0.4179	**0.5933**	0.7880	0.7285	0.6296	0.4747	0.4134	0.7024
SCNN_UNetLight_ConvGRU1	0.8896	0.8212	0.7819	0.7517	0.8913	0.3043	0.4961	0.8133	0.7000	0.5635	0.3086	0.2733	0.6637
SCNN_UNetLight_ConvGRU2	0.8593	0.8730	0.7878	0.7406	0.8889	0.3335	0.4266	0.7263	0.7782	0.6498	0.5280	0.5257	0.6910
SCNN_UNetLight_ConvLSTM1	0.8115	0.8056	0.7168	0.6882	0.8179	0.2613	0.3681	0.7834	0.7576	0.5701	0.5281	0.5081	0.6418
SCNN_UNetLight_ConvLSTM2	0.8377	0.8158	0.7620	0.6971	0.8365	0.2209	0.3577	0.7551	0.6594	0.4597	0.3545	0.3559	0.6079

2.3.4 Parameter analysis and ablation study

1) The added value of SCNN

Regarding the neural network architecture, the effects of SCNN were investigated by evaluating the performances of the model variants with and without SCNN layers. As demonstrated in **Figure 2-3** and **Figure 2-4**, together with the quantitative results in **Table 2-2** and **Table 2-3**, the proposed SegNet and UNet-based models with SCNN embedded encoder, i.e., SCNN_SegNet_ConvLSTM, SCNN_SegNet_ConvGRU, SCNN_UNet_ConvLSTM, and SCNN_UNet_ConvGRU, outperform SegNet_ConvLSTM and UNet_ConvLSTM, which are also SegNet or UNet-based sequential models using multiple continuous image frames as inputs but without SCNN. Especially, SCNN_UNet_ConvLSTM2 obtains the best result in normal testing, while SCNN_SegNet_ConvLSTM2 delivers the best performance in challenging situations.

For testing on normal cases in tvtLANE testset #1, as shown in **Table 2-2**, by adding SCNN layer in the encoder, almost all the proposed models with SCNN embedded encoder outperform the baseline models with better F1-measure. To be specific, SCNN_SegNet_ConvLSTM2 improves the lane detection accuracy by around 0.3% and F1-measure by around 1%, and these improvements are from the already very good results obtained by SegNet_ConvLSTM. Similarly, SCNN_UNet_ConvLSTM2 overperforms UNet_ConvLSTM with even larger margins regarding Accuracy, Precision, and F1-measure.

For challenging situations, adding the SCNN layer also helps the proposed model, SCNN_SegNet_ConvLSTM2, beat other baseline models, and deliver the best F1-measure as indicated in **Table 2-3**.

Figure 2-5 visualises the extracted features at the *Down_ConvBlock_1* layer for UNet-based models, with and without SCNN. Clearly, vast differences can be witnessed between the baseline model UNet_ConvLSTM and the proposed model SCNN_UNet_ConvLSTM2. In **Figure 2-5 (b)**, the CNN-based UNet layers identify the low-level features in the images regarding the target lane lines. However, the extracted features are not so clear, i.e., there are some interference signals, especially as visualised in the third image of row (b), which is supposed to affect the model training (i.e., updating weight parameters of the neural networks) and thus affect the model's performance regarding the marking detection results. It might further influence the final detection results. In contrast, with SCNN layers, the extracted features of the lanes are more inerratic, clear, and evident as shown in **Figure 2-5 (c)**. There are fewer interferences surrounding the detected lane features. This verifies SCNN's powerful strength in detecting the spatial relations in every single image with its message passing mechanism.

All the above results demonstrate that adding the SCNN layer embedded in the encoder does contribute to the spatial feature extraction, with which the model could better make the utmost use of the spatial-temporal information among the continuous image frames.

2) Different locations of the SCNN layer

Results of testing different locations of the SCNN layer in the proposed model architecture are shown in **Table 2-4**. The results reveal that: (a) Compared with baseline models without SCNN

layers, the embedding of SCNN layers really helps to improve the models' performance, which further verifies the added value of SCNN and accords with the aforementioned results in *1)*; (b) In terms of the main evaluation metric F1-measure, embedding SCNN layer after the *Conv1_1* (in SegNet-based model) or *In_Conv_1* (in UNet-based model) layer delivers better results compared with embedding it at the very beginning or early layers of the encoder; (c) For UNet-based model, embedding SCNN layer at the very beginning delivers quite good Precision and Accuracy, but worse Recall, which means there are fewer False Positives but more False Negatives. This should be related to the properties of the UNet-style neural network. These results further confirm the effectiveness of the proposed model architecture.

Figure 2-5. Visualisation of the extracted low-level features at *Down_ConvBlock_1* for UNet-based models

(a) Original image. (b) Results of UNet_ConvLSTM (without SCNN layers). (c) Results of the SCNN_UNet_ConvLSTM2 (with SCNN layers).

Table 2-4. Model performance comparison with different locations of the SCNN layer on tvtLANE testsets #1 and #2

Model \ Testing Datasets		Testset #1 (Normal Situations)				Testset #2 (Challenging Scenes)			
	Location of SCNN	Test_ Acc (%)	Precision	Recall	F1-measure	Test_ Acc (%)	Precision	Recall	F1-measure
SegNet_ConvLSTM	Without	97.92	0.874	0.931	0.901	97.83	0.756	0.765	0.761
SCNN_SegNet_Conv LSTM2	Conv1_1	98.00	0.884	0.921	0.902	97.92	0.757	0.757	0.757
	Conv2_1	98.07	0.893	0.928	0.910	97.90	0.767	0.766	0.767
UNet_Conv LSTM	Without	98.00	0.857	0.957	0.904	97.93	0.778	0.660	0.714
SCNN_UNet_Conv LSTM2	In_Conv_1	98.28	0.896	0.939	0.917	98.08	0.776	0.593	0.672
	Conv1_1	98.19	0.889	0.950	0.918	97.95	0.778	0.640	0.702

3) Type and number of ST-RNN layers

As described in *Section 2.3*, in the proposed model architecture, two types of RNNs, that is, ConvLSTM and ConvGRU, are employed to serve in the ST-RNN block, to capture and make use of the ST dependencies and correlations among the continuous image sequences. The number of hidden ConvLSTM and ConvGRU layers was also tested from 1 to 2. The quantitative results are demonstrated in **Tables 2-2** and **2-3**, while some intuitive qualitative insights could be drawn from **Figures 2-3** and **2-4**.

From **Table 2-2**, it is illustrated that, in general, models adopting ConvLSTM layers in the ST-RNN block perform better than those adopting ConvGRU layers with improved F1-measure, except for the UNetLight-based models. This could be explained by ConvLSTM's better properties in extracting ST features and capturing time dependencies by more control gates and thus more parameters, compared with ConvGRU. Furthermore, from **Tables 2-2** and **2-3**, it is observed that models with two hidden ST-RNN layers, for both ConvLSTM and ConvGRU, generally perform better than those with only one hidden ST-RNN layer. This could be speculated that with two hidden ST-RNN layers, one layer can serve for sequential feature extraction, and the other can achieve ST feature integration. The improvements of two ST-RNN layers over one are not that significant, which might be due to (a) models employing one ST-RNN layer already obtaining good results; (b) since the length of the continuous image frames is only five, one ST-RNN layer might be already enough to do the ST feature extraction, so when incorporating longer image sequences, the superiorities of two ST-RNN layers could be promoted. However, longer image sequences require more computational resources and longer training time, which could not be afforded at the present stage in this study. This could be the future research direction.

4) Number of parameters and real-time capability

As shown in **Table 2-2**, the two proposed candidate models, that is, SCNN_SegNet_ConvLSTM2 and SCNN_UNet_ConvLSTM2, possess a bit more parameters compared with the baseline SegNet_ConvLSTM and UNet_ConvLSTM, respectively. However, almost all of the proposed model variants with different types and numbers of ST-RNN layers outperform the baselines, and some of them are even with low parameter sizes, for example, SCNN_SegNet_ConvGRU1, SCNN_SegNet_ConvLSTM1, SCNN_UNet_ConvGRU1, SCNN_UNet_ConvLSTM1. Generally speaking, lower numbers of model parameters mean better real-time capability.

In addition, four model variants were implemented with a modified light version of UNet, that is, UNetLight, serving as the network backbone to reduce the total parameter size and improve the model's ability to operate in real time. The UNetLight backbone has a similar network design to UNet, whose parameter settings are demonstrated in Appendix **Table 2-A2**. The only difference is that all the numbers of kernels in the ConvBlocks are reduced to half, except for the *Input* in *In_ConvBlock* (with the input channel of three unchanged) and the *Output* in *Out_ConvBlock* (with the output channel of two unchanged). From the testing results in **Table 2-2**, it is shown that the model named SCNN_UNetLight_ConvGRU2, with fewer parameters than all the baseline models, beats the baselines, exhibiting better performance regarding both accuracy and F1-measure. To be specific, compared with the best baseline model, that is, UNet_ConvLSTM, SCNN_UNetLight_ConvGRU2 only uses less than one-fifth of the

parameter size but delivers better evaluation metrics in testing accuracy, precision, and F1-measure.

Regarding UNetLight-based models, models using ConvGRU layers in the ST-RNN block perform better than those adopting ConvLSTM. The reason could be that the light version UNet cannot implement high-quality feature extraction, which does not feed enough information for ConvLSTM, while ConvGRU, with fewer control gates, is more robust when low-level features are not that fully extracted.

All these results further verify the proposed network architecture's effectiveness and strength.

2.4 Conclusion

In this study, a novel ST sequence-to-one model framework with a hybrid neural network architecture is proposed for robust lane detection under various normal and challenging driving scenes. This architecture integrates a single image feature extraction module with SCNN, an ST feature integration module with ST-RNN, together with the encoder-decoder structure. The proposed architecture achieved significantly better results in comparison to baseline models that use a single frame (e.g., UNet, SegNet, and LaneNet), as well as the state-of-the-art models adopting "CNN+RNN" structures (e.g., UNet_ConvLSTM, SegNet_ConvLSTM), with the best testing accuracy, precision, and F1-measure on the normal driving dataset (i.e., tvtLANE testset #1) and the best F1-measure on the 12 challenging driving scenarios dataset (tvtLANE testset #2). The results demonstrate the effectiveness of strengthening spatial relation abstraction in every single image with SCNN layer, plus the employment of multiple continuous image sequences as inputs. The results also demonstrate the proposed model architecture's ability in making the best of the ST information in continuous image frames. Extensive experimental results show the superiorities of the sequence-to-one "SCNN + ConvLSTM" over "SCNN + ConvGRU" and ordinary "CNN + ConvLSTM" regarding sequential ST feature extracting and learning, together with target-information classification for robust lane detection. In addition, testing results of the model variants with the modified light version of UNet (i.e., UNetLight) as the backbone demonstrate the proposed model architecture's potential regarding real-time capability.

To the best of the authors' knowledge, the proposed model is the first attempt that tries to strengthen both spatial relations regarding feature extraction in every image frame together with the ST correlations and dependencies among image frames for lane detection, and the extensive evaluation experiments demonstrate the strength of this proposed architecture. Therefore, it is recommended in future research to incorporate both aspects to obtain better performance.

In this study, the challenging cases do not include night driving, rainy, or wet road conditions, nor do they include situations in which the input images are defective (e.g., partly masked or blurred). There are demands to build larger test sets with comprehensive challenging situations to further validate the model's robustness. Since a large amount of unlabelled driving scene data involving various challenging cases was collected within the research group, a future research direction might be to develop semi-supervised learning methods and employ domain adaptation to label the collected data, and then open-source them for boosting the research in the field of robust lane detection. Furthermore, to further enhance the lane-detection model, customised loss functions, pre-trained techniques adopted in image-inpainting tasks, for example, masked

autoencoders, plus sequential attention mechanisms, could be introduced and integrated into the proposed framework.

Acknowledgements

This work was supported by the Applied and Technical Sciences (TTW), a subdomain of the Dutch Institute for Scientific Research (NWO) through the Project Safe and Efficient Operation of Automated and Human-Driven Vehicles in Mixed Traffic (SAMEN) under Contract 17187. The authors thank Dr. Qin Zou, Hanwen Jiang, and Qiyu Dai from Wuhan University, as well as Jiyong Zhang from Southwest Jiaotong University, for their tips in using the tvtLANE dataset.

References

Aly, M. (2008). Real time detection of lane markers in urban streets. 2008 IEEE Intelligent Vehicles Symposium. Eindhoven, the Netherlands (pp. 7–12). https://doi.org/10.1109/IVS.2008.4621152

Andrade, D. C., Bueno, F., Franco, F. R., Silva, R. A., Neme, J. H. Z., Margraf, E., Omoto, W. T., Farinelli, F. A., Tusset, A. M., Okida, S., Santos, M. M. D., Ventura, A., Carvalho, S., & Amaral, R. D. S. (2019). A novel strategy for road lane detection and tracking based on a vehicle's forward monocular camera. IEEE Transactions on Intelligent Transportation Systems, 20, 1497–1507. https://doi.org/ 10.1109/TITS.2018.2856361

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Ballas, N., Yao, L., Pal, C., & Courville, A. (2016). Delving deeper into convolutional networks for learning video representations. 4th International Conference on Learning Representations, ICLR 2016–Conference Track Proceedings, San Juan, Puerto Rico.

Bar Hillel, A., Lerner, R., Levi, D., & Raz, G. (2014). Recent progress in road and lane detection: A survey. Machine Vision and Applications, 25, 727–745. https://doi.org/10.1007/s00138-011-0404-2

Berriel, R. F., de Aguiar, E., de Souza, A. F., & Oliveira-Santos, T. (2017). Ego-Lane Analysis System (ELAS): Dataset and algorithms. Image and Vision Computing, 68, 64–75. https://doi.org/10.1016/j. imavis.2017.07.005

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT 2010–19th International Conference on Computational Statistics, Keynote, Invited and Contributed Papers, Paris, France. https://doi.org/10.1007/978-3-7908- 2604-3_16

Chen, S., Leng, Y., & Labi, S. (2020). A deep learning algorithm for simulating autonomous driving considering prior knowledge and temporal information. Computer-Aided Civil and Infrastructure Engineering, 35, 305–321. https://doi.org/10.1111/mice.12495

Chen, W., Wang, W., Wang, K., Li, Z., Li, H., & Liu, S. (2020). Lane departure warning systems and lane line detection methods based on image processing and semantic segmentation–a review. Journal of Traffic and Transportation Engineering (English Edition), 7(6), 748–774. https://doi.org/10.1016/j.jtte.2020.10.002

Chen, Z., Liu, Q., & Lian, C. (2019). PointLaneNet: Efficient end-toend CNNs for accurate real-time lane detection. IEEE Intelligent Vehicles Symposium 2019, Paris, France (pp. 2563–2568). https:// doi.org/10.1109/IVS.2019.8813778

Choi, Y., Park, J. H., & Jung, H. (2018). Lane detection using labeling based RANSAC algorithm. International Journal of Computer and Information Engineering, 12(4), 245–248.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on computer vision and Pattern Recognition, Miami, FL (pp. 248–255). https://doi.org/10.1109/cvprw.2009.5206848

Du, H., Xu, Z., & Ding, Y. (2018). The fast lane detection of road using RANSAC algorithm. In J. Abawajy, K. K. Choo, & R. Islam (Eds.), Advances in Intelligent Systems and Computing (pp. 1–7). Springer. https://doi.org/10.1007/978-3-319-67071-3_1

Guo, J., Wei, Z., & Miao, D. (2015). Lane detection method based on improved RANSAC algorithm. IEEE 12th International Symposium on Autonomous Decentralized Systems, ISADS 2015, Taichung, Taiwan (pp. 285–288). https://doi.org/10.1109/ISADS. 2015.24

Haris, M., & Glowacz, A. (2021). Lane line detection based on object feature distillation. Electron, 10(9), 1102. https://doi.org/10.3390/ electronics10091102

Hochreiter, S., & Schmidhuber, J. (1997). Long short term memory. Neural Computation. Neural Computation, 9(8), 1735–1780.

Hou, Y., Ma, Z., Liu, C., Hui, T. W., & Loy, C. C. (2020). Interregion affinity distillation for road marking segmentation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA (pp. 12483–125492). https://doi.org/10.1109/ CVPR42600.2020.01250

Jiao, X., Yang, D., Jiang, K., Yu, C., Wen, T., & Yan, R. (2019). Realtime lane detection and tracking for autonomous vehicle applications. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 233(9), 2301–2311. https:// doi.org/10.1177/0954407019866989

Kim, J., & Park, C. (2017). End-to-end ego lane estimation based on sequential transfer learning for self-driving cars. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, pp. 1194–1202. https://doi.org/10.1109/ CVPRW.2017.158

Ko, Y., Lee, Y., Azam, S., Munir, F., Jeon, M., & Pedrycz, W. (2020). Key points estimation and point instance segmentation approach for lane detection. IEEE Transactions on Intelligent Transportation Systems, 23(7), 8949-8958. https://doi.org/10.1109/tits.2021.3088488

Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Kolter, J. Z., Langer, D., Pink, O., Pratt, V., Sokolsky, M., Stanek, G., Stavens, D., Teichman, A., Werling, M., & Thrun, S. (2011). Towards fully autonomous driving: Systems and algorithms. 2011 IEEE Intelligent Vehicles Symposium, Baden-Baden, Germany (pp. 163–168). https://doi.org/10.1109/IVS.2011.5940 562

Li, X., Li, J., Hu, X., & Yang, J. (2020). Line-CNN: End-to-end traffic line detection with line proposal unit. IEEE Transactions on Intelligent Transportation Systems, 21, 248–258. https://doi.org/10.1109/ TITS.2019.2890870

Liang, D., Guo, Y. C., Zhang, S. K., Mu, T. J., & Huang, X. (2020). Lane detection: A Survey with new results. Journal of Computer Science and Technology, 35, 493–505. https://doi.org/10.1007/s11390- 020-0476-4

Lin, C., Li, L., Cai, Z., Wang, K. C. P., Xiao, D., Luo, W., & Guo, J. G. (2020). Deep learning-based lane marking detection using A2-LMDet. Transportation Research Record, 2674(11), 625–635. https://doi.org/10.1177/0361198120948508

Liu, L., Chen, X., Zhu, S., & Tan, P. (2021). CondLaneNet: A top-to-down lane detection framework based on conditional convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3773–3782).

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2019). On the variance of the adaptive learning rate and beyond. In International Conference on Learning Representations.

Liu, R., Yuan, Z., Liu, T., & Xiong, Z. (2020). End-to-end lane shape prediction with transformers. 2021 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI (pp. 3694–3702). https://doi.org/10.1109/wacv48630.2021.00374

Liu, T., Chen, Z., Yang, Y., Wu, Z., & Li, H. (2020). Lane detection in low-light conditions using an efficient data enhancement: Light conditions style transfer. 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV (pp. 1394–1399) https://doi.org/10.1109/IV47402.2020.9304613

Lu, Z., Xu, Y., Shan, X., Liu, L., Wang, X., & Shen, J. (2019). A lane detection method based on a ridge detector and regional GRANSAC. Sensors (Switzerland), 19(18), 4028. https://doi.org/10. 3390/s19184028

Neven, D., De Brabandere, B., Georgoulis, S., Proesmans, M., & Van Gool, L. (2017). Fast scene understanding for autonomous driving. arXiv preprint arXiv:1708.02550.

Neven, D., De Brabandere, B., Georgoulis, S., Proesmans, M., & Van Gool, L. (2018). Towards end-to-end lane detection: An Instance segmentation approach. 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China (pp. 286–291). https://doi.org/10.1109/IVS.2018.8500547

Pan, X., Shi, J., Luo, P., Wang, X., & Tang, X. (2018). Spatial as deep: Spatial CNN for traffic scene understanding. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018. New Orleans, LA (pp. 7276–7283).

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA.

Philion, J. (2019). FastDraw: Addressing the long tail of lane detection by adapting a sequential prediction network. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA (pp. 11574–11583). https://doi.org/10.1109/CVPR.2019.01185

Qin, Z., Wang, H., & Li, X. (2020). Ultra fast structure-aware deep lane detection. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16 (pp. 276-291). Springer International Publishing.

Ribeiro, A. H., Tiels, K., Aguirre, L. A., & Schön, T. (2020). Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness. International Conference on Artificial Intelligence and Statistics, Palermo, Sicily, Italy (pp. 2370–2380).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015 (Vol. 9351, pp. 234–241). https://doi.org/10.1007/978-3-319-24574- 4_28

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in Neural Information Processing Systems, Montreal, Canada.

Sivaraman, S., & Trivedi, M. M. (2013). Integrated lane and vehicle detection, localization, and tracking: A synergistic approach. IEEE Transactions on Intelligent Transportation Systems, 14, 906–917. https://doi.org/10.1109/TITS.2013.2246835

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, Montreal, Canada (pp. 3104–3112).

Tabelini, L., Berriel, R., Paixão, T. M., Badue, C., de Souza, A. F., & Oliveira-Santos, T. (2021a). PolyLaneNet: Lane estimation via deep polynomial regression. 2020 25th International Conference on Pattern Recognition (ICPR), (2021, pp. 6150–6156). https://doi.org/10.1109/ICPR48806.2021.9412265

Tabelini, L., Berriel, R., Paixão, T. M., Badue, C., De Souza, A. F., & Olivera-Santos, T. (2021b). Keep your eyes on the lane: Real-time attention-guided lane detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 294– 302).

Wang, B. F., Qi, Z. Q., & Ma, G. C. (2014). Robust lane recognition for structured road based on monocular vision. Journal of Beijing Institute of Technology (English Edition), 23, 345–351.

Wang, S., Hou, X., & Zhao, X. (2020). Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block. IEEE Access, 8, 7313–7322. https://doi.org/10.1109/ACCESS.2020.2964043

Wang, Y., Dahnoun, N., & Achim, A. (2012). A novel system for robust lane detection and tracking. Signal Processing, 92(2), 319–334. https://doi.org/10.1016/j.sigpro.2011.07.019

Wu, B., Li, K., Ge, F., Huang, Z., Yang, M., Siniscalchi, S. M., & Lee, C. H. L. (2017). An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1289–1300. https://doi.org/10.1109/JSTSP. 2017.2756439

Xing, Y., Lv, C., Chen, L., Wang, H., Wang, H., Cao, D., Velenis, E., & Wang, F. Y. (2018). Advances in vision-based lane detection: Algorithms, integration, assessment, and perspectives on ACP-based parallel vision. IEEE/CAA Journal of Automatica Sinica, 5, 645– 661. https://doi.org/10.1109/JAS.2018.7511063

Xu, H., Wang, S., Cai, X., Zhang, W., Liang, X., & Li, Z. (2020). CurveLane-NAS: Unifying lane-sensitive architecture search and adaptive point blending. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, (vol. 12360). Springer, Cham. https://doi.org/10.1007/978-3-030-58555-6_41

Yasrab, R., Gu, N., & Zhang, X. (2017). An encoder-decoder based Convolution Neural Network (CNN) for future Advanced Driver Assistance System (ADAS). Applied Science, 7(4), 312. https://doi. org/10.3390/app7040312

Yoo, S., Lee, H. S., Myeong, H., Yun, S., Park, H., Cho, J., & Kim, D. H. (2020). End-to-end lane marker detection via row-wise classification. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA (pp. 4335–4343). https://doi.org/10.1109/CVPRW50498.2020.00511

Zhang, J., Deng, T., Yan, F., & Liu, W. (2021). Lane detection model based on spatio-temporal network with double convolutional gated recurrent units. IEEE Transactions on Intelligent Transportation Systems, 1–13. https://doi.org/10.1109/TITS.2021.3060258

Zheng, F., Luo, S., Song, K., Yan, C. W., & Wang, M. C. (2018). Improved lane line detection algorithm based on Hough transform. Pattern Recognition and Image Analysis, 28, 254–260. https://doi.org/10.1134/S1054661818020049

Zou, Q., Jiang, H., Dai, Q., Yue, Y., Chen, L., & Wang, Q. (2020). Robust lane detection from continuous driving scenes using deep neural networks. IEEE Transactions on Vehicular Technology, 69, 41–54. https://doi.org/10.1109/TVT.2019.2949603

Zou, Q., Ni, L., Wang, Q., Li, Q., & Wang, S. (2017). Robust gait recognition by integrating inertial and RGBD sensors. IEEE Transactions on Cybernetics, 48(4), 1136–1150. https://doi.org/10.1109/ TCYB.2017.2682280

Appendix

Table 2-A1. Parameter settings for each layer of the SegNet-based neural network

Layer		Input (channel×hight×width)	Output (channel×hight×width)	Kernel	Padding	Stride	Activation
Down_Conv Block_1	Conv_1_1	3×128×256	64×128×256	3×3	(1,1)	1	ReLU
	Conv_1_2	64×128×256	64×128×256	3×3	(1,1)	1	ReLU
	Maxpool1	64×128×256	64×64×128	2×2	(0,0)	2	---
SCNN	SCNN_Down	64×1×128	64×1×128	1×9	(0,4)	1	ReLU
	SCNN_Up	64×1×128	64×1×128	1×9	(0,4)	1	ReLU
	SCNN_Right	64×64×1	64×64×1	9×1	(4,0)	1	ReLU
	SCNN_Left	64×64×1	64×64×1	9×1	(4,0)	1	ReLU
Down_Conv Block_2	Conv_2_1	64×64×128	128×64×128	3×3	(1,1)	1	ReLU
	Conv_2_2	128×64×128	128×64×128	3×3	(1,1)	1	ReLU
	Maxpool2	128×64×128	128×32×64	2×2	(0,0)	2	---
Down_Conv Block_3	Conv_3_1	128×32×64	256×32×64	3×3	(1,1)	1	ReLU
	Conv_3_2	256×32×64	256×32×64	3×3	(1,1)	1	ReLU
	Conv_3_3	256×32×64	256×32×64	3×3	(1,1)	1	ReLU
	Maxpool3	256×64×128	256×16×32	2×2	(0,0)	2	---
Down_Conv Block_4	Conv_4_1	256×16×32	512×16×32	3×3	(1,1)	1	ReLU
	Conv_4_2	512×16×32	512×16×32	3×3	(1,1)	1	ReLU
	Conv_4_3	512×16×32	512×16×32	3×3	(1,1)	1	ReLU
	Maxpool4	512×16×32	512×8×16	2×2	(0,0)	2	---
Down_Conv Block_5	Conv_5_1	512×8×16	512×8×16	3×3	(1,1)	1	ReLU
	Conv_5_2	512×8×16	512×8×16	3×3	(1,1)	1	ReLU
	Conv_5_3	512×8×16	512×8×16	3×3	(1,1)	1	ReLU
	Maxpool5	512×8×16	512×4×8	2×2	(0,0)	2	---
ST-RNN Layer1*	5 × ConvLSTMCell(input=(512×4×8), kernel=(3,3), stride=(1,1), padding=(1,1)) ***Or*** 5 × ConvGRUCell(input=(512×4×8), kernel=(3,3), stride=(1,1), padding=(1,1), dropout(0.5))						
ST-RNN Layer2**	5 × ConvLSTMCell(input=(512×4×8), kernel=(3,3), stride=(1,1), padding=(1,1)) ***Or*** 5 × ConvGRUCell(input=(512×4×8), kernel=(3,3), stride=(1,1), padding=(1,1), dropout(0.5))						
Up_Conv Block_5	MaxUnpool1	512×4×8	512×8×16	2×2	(0,0)	2	---
	Up_Conv_5_1	512×8×16	512×8×16	3×3	(1,1)	1	ReLU
	Up_Conv_5_2	512×8×16	512×8×16	3×3	(1,1)	1	ReLU
	Up_Conv_5_3	512×8×16	512×8×16	3×3	(1,1)	1	ReLU
Up_Conv Block_4	MaxUnpool2	512×8×16	512×16×32	2×2	(0,0)	2	---
	Up_Conv_4_1	512×16×32	512×16×32	3×3	(1,1)	1	ReLU
	Up_Conv_4_2	512×16×32	512×16×32	3×3	(1,1)	1	ReLU
	Up_Conv_4_3	512×16×32	256×16×32	3×3	(1,1)	1	ReLU
Up_Conv Block_3	MaxUnpool3	256×16×32	256×32×64	2×2	(0,0)	2	---
	Up_Conv_3_1	256×32×64	256×32×64	3×3	(1,1)	1	ReLU
	Up_Conv_3_2	256×32×64	256×32×64	3×3	(1,1)	1	ReLU
	Up_Conv_3_3	256×32×64	128×32×64	3×3	(1,1)	1	ReLU
Up_Conv Block_2	MaxUnpool4	128×32×64	128×64×128	2×2	(0,0)	2	---
	Up_Conv_2_1	128×64×128	128×64×128	3×3	(1,1)	1	ReLU
	Up_Conv_2_2	128×64×128	64×64×128	3×3	(1,1)	1	ReLU
Up_Conv Block_1	MaxUnpool5	64×64×128	64×128×256	2×2	(0,0)	2	---
	Up_Conv_1_1	64×128×256	64×128×256	3×3	(1,1)	1	ReLU
	Up_Conv_1_2	64×128×256	2×128×256	3×3	(1,1)	1	LogSoftmax

***Abbreviations**: ConvGRU, convolutional gated recurrent unit; ConvLSTM, convolutional long short-term memory; SCNN, spatial convolutional neural network; ST-RNN, spatial-temporal recurrent neural network; ReLU, Rectified Linear Unit.*

** Two types of ST-RNN, i.e., ConvLSTM and ConvGRU are tested;*

*** ST-RNN blocks are tested with 1 hidden layer or 2 hidden layers.*

Table 2-A2. Parameter settings for each layer of the UNet-based neural network

Layer		Input (channel×hight×width)	Output (channel×hight×width)	Kernel	Padding	Stride	Activation
In_Conv Block	In_Conv_1	3×128×256	64×128×256	3×3	(1,1)	1	ReLU
	In_Conv_2	64×128×256	64×128×256	3×3	(1,1)	1	ReLU
SCNN	SCNN_Down	64×1×256	64×1×256	1×9	(0,4)	1	ReLU
	SCNN_Up	64×1×256	64×1×256	1×9	(0,4)	1	ReLU
	SCNN_Right	64×128×1	64×128×1	9×1	(4,0)	1	ReLU
	SCNN_Left	64×128×1	64×128×1	9×1	(4,0)	1	ReLU
Down_Conv Block_1	Maxpool1	64×128×256	64×64×128	2×2	(0,0)	2	---
	Conv_1_1	64×64×128	128×64×128	3×3	(1,1)	1	ReLU
	Conv_1_2	128×64×128	128×64×128	3×3	(1,1)	1	ReLU
Down_Conv Block_2	Maxpool2	128×64×128	128×32×64	2×2	(0,0)	2	---
	Conv_2_1	128×32×64	256×32×64	3×3	(1,1)	1	ReLU
	Conv_2_2	256×32×64	256×32×64	3×3	(1,1)	1	ReLU
Down_Conv Block_3	Maxpool3	256×32×64	256×16×32	2×2	(0,0)	2	---
	Conv_3_1	256×16×32	512×16×32	3×3	(1,1)	1	ReLU
	Conv_3_2	512×16×32	512×16×32	3×3	(1,1)	1	ReLU
Down_Conv Block_4	Maxpool4	512×16×32	512×8×16	2×2	(0,0)	2	---
	Conv_4_1	512×8×16	512×8×16	3×3	(1,1)	1	ReLU
	Conv_4_2	512×8×16	512×8×16	3×3	(1,1)	1	ReLU
ST-RNN Layer1*	5 × ConvLSTMCell(input=(512×8×16), kernel=(3,3), stride=(1,1), padding=(1,1)) ***Or*** 5 × ConvGRUCell(input=(512×8×16), kernel=(3,3), stride=(1,1), padding=(1,1), dropout(0.5))						
ST-RNN Layer2**	5 × ConvLSTMCell(input=(512×8×16), kernel=(3,3), stride=(1,1), padding=(1,1)) ***Or*** 5 × ConvGRUCell(input=(512×8×16), kernel=(3,3), stride=(1,1), padding=(1,1), dropout(0.5))						
Up_Conv Block_4	Upsampling Bilinear2D_1	512×8×16	512×16×32	2×2	(0,0)	2	---
	Up_Conv_4_1	1024×16×32	256×16×32	3×3	(1,1)	1	ReLU
	Up_Conv_4_2	256×16×32	256×16×32	3×3	(1,1)	1	ReLU
Up_Conv Block_3	Upsampling Bilinear2D_2	256×16×32	256×32×64	2×2	(0,0)	2	---
	Up_Conv_3_1	512×32×64	128×32×64	3×3	(1,1)	1	ReLU
	Up_Conv_3_2	128×32×64	128×32×64	3×3	(1,1)	1	ReLU
Up_Conv Block_2	Upsampling Bilinear2D_3	128×32×64	128×64×128	2×2	(0,0)	2	---
	Up_Conv_2_1	156×64×128	64×64×128	3×3	(1,1)	1	ReLU
	Up_Conv_2_2	64×64×128	64×64×128	3×3	(1,1)	1	ReLU
Up_Conv Block_1	Upsampling Bilinear2D_4	64×64×128	64×128×256	2×2	(0,0)	2	---
	Up_Conv_1_1	128×128×256	64×128×256	3×3	(1,1)	1	ReLU
	Up_Conv_1_2	64×128×256	64×128×256	3×3	(1,1)	1	ReLU
Out_Conv Block	Out_Conv	64×128×256	2×128×256	1×1	(0,0)	1	---

Abbreviations: *ConvGRU, convolutional gated recurrent unit; ConvLSTM, convolutional long short-term memory; SCNN, spatial convolutional neural network; ST-RNN, spatial-temporal recurrent neural network; ReLU, Rectified Linear Unit.*

** Similar to the SegNet-based network architecture, two types of ST-RNN, i.e., ConvLSTM and ConvGRU, are tested;*

*** ST-RNN blocks are tested with one hidden layer or two hidden layers.*

3 Efficient sequential neural network based on spatial-temporal attention and linear LSTM for robust lane detection using multi-frame images

Abstract

Lane detection serves as a fundamental task for automated vehicles and Advanced Driver Assistance Systems. However, existing lane detection methods often fail to deliver the versatility of accurate, robust, and real-time compatible lane detection, especially under challenging driving scenes. Available vision-based methods in the literature frequently overlook critical regions of the image and their spatial-temporal salience regarding the detection results, leading to poor performance in peculiar difficult circumstances (e.g., serious occlusion, dazzle lighting). To address these limitations, this study introduces a novel spatial-temporal attention mechanism that can focus on key features of lane lines and exploit salient spatial-temporal correlations among continuous image frames to enhance the accuracy and robustness of lane detection. Under the standard encoder-decoder structure and with the implementation using common neural network backbones, efficient sequential neural network models are developed incorporating the proposed spatial-temporal attention mechanism. The developed models are trained and evaluated on three large-scale open-source datasets. Extensive experiments demonstrate the strength and robustness of the developed model outperforming available state-of-the-art methods across various testing scenarios. Furthermore, with the spatial-temporal attention mechanism, the developed sequential neural network models exhibit fewer parameters and reduced Multiply-Accumulate Operations (MACs) compared to baseline sequential models, highlighting their computational efficiency and real-world applicability. Relevant data, code, and models are released at https://doi.org/10.4121/4619cab6-ae4a-40d5-af77-582a77f3d821.

This chapter is currently under review for journal publication, and it has been pre-printed on TechRxiv.

Patil, S., Dong, Y., Farah, H., & Hellendoorn, H. (2025). Efficient Sequential Neural Network based on Spatial-Temporal Attention and Linear LSTM for Robust Lane Detection Using Multi-frame Images. https://doi.org/10.36227/techrxiv.174195585.50092304/v1 (Co-first authors and corresponding author)*

3.1 Introduction

The objective of lane detection is to assist vehicles in locating and positioning themselves within the lane by identifying and predicting the positions of marked lane lines. While various sensors, e.g., mono-camera, stereo-camera, radar, and LiDAR, could be applied in the process of detecting the lane boundaries for accurate localisation (Bai et al., 2018; Bar Hillel et al., 2014), the most common, feasible, and successful approach is vision-based lane marking detection (Chetan et al., 2020; Yeong et al., 2021).

Conventional vision-based methods usually handle lane detection by utilising specialized hand-crafted low-level features with traditional computer vision techniques, e.g., Inverse Perspective Mapping applied in the image pre-processing stage (Aly, 2008; B. F. Wang et al., 2014); Hough transform applied for feature extraction (Berriel et al., 2017; Jiao et al., 2019; Satzoda et al., 2010; Zheng et al., 2018); Gaussian filters and Random Sample Consensus (RANSAC) employed in the post-processing process to smooth the lane detection results (Sivaraman & Trivedi, 2013; Y. Wang et al., 2012). These traditional methods suffer from many shortcomings, e.g., they require hand-crafted features which are always complex and time-consuming but not necessarily suitable or effective enough, and they usually use one single image to detect the lane, thus they cannot handle some extremely challenging driving scenarios.

Recent advances in computational hardware, along with rapid developments in neural network (NN) models, have enabled deep learning based lane detection methods to extract useful features automatically. They have been widely used to eliminate intermediate feature crafting, as well as enable end-to-end lane detection, outperforming traditional approaches (Hou et al., 2019; Neven et al., 2018; Pan et al., 2018; Tang et al., 2021).

Usually, deep Convolutional Neural Networks (CNNs) have been widely adopted for their superior abilities in image feature abstraction, demonstrating exceptional performance in lane detection tasks, e.g., in (Kim & Park, 2017; Pan et al., 2018). In addition to CNNs, other architectures like Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), and Vision Transformers (ViTs) have also been explored in the domain of lane detection research. RNNs, known for their capability to process sequential data, are adept at abstracting and predicting time-series features. Consequently, they are often employed to model sequential patterns within a single image (J. Li et al., 2017) or across frames in continuous image sequences for lane detection (Dong et al., 2023; Zou et al., 2020). GANs, which leverage two neural networks competing in a shared task have been used for data augmentation (e.g., generating synthetic lane images) and transfer learning applications in the lane detection task (T. Liu et al., 2020). Recently, ViTs, adapted from the original Transformer architecture, renowned for its success in natural language processing (NLP) tasks, have been applied to computer vision problems, including lane detection. Studies such as (Han et al., 2022; R. Liu et al., 2021; Yu et al., 2020; Zhao et al., 2024) utilise the self-attention mechanism inherent in Transformers to focus on salient regions in an image, improving lane detection accuracy. However, these approaches predominantly rely on single-image, overlooking temporal correlations and the varying importance of frames in continuous driving scenarios.

Furthermore, a few studies have attempted to combine CNNs and RNNs to detect lane markings through continuous driving scene image frames (Dong et al., 2023; R. Li & Dong, 2023; J. Zhang et al., 2022; Zou et al., 2020). However, these approaches fail to fully exploit the inherent properties of lanes and often overlook the salient spatial-temporal correlations and

dependencies among critical regions across sequential frames. As a result, their performance remains unsatisfactory under highly challenging driving conditions.

To address the aforementioned research gaps and improve the performance of vision-based lane detection, this study introduces a novel efficient sequential neural network architecture with the proposed spatial-temporal attention mechanisms. The developed model formulates lane detection as a segmentation task and takes multiple continuous image frames as input. By effectively extracting key features and leveraging salient correlations across these frames, the proposed approach strongly exploits the spatial-temporal information inherent in the driving scene. Built on a standard encoder-decoder framework and utilising labelled ground truth from the final image in the sequence, the model employs a supervised, end-to-end learning strategy. The primary contributions of this study are as follows:

1. Introduction of spatial-temporal attention mechanisms: Three attention model variants are proposed and implemented to improve feature extraction.
2. Strong exploitation of spatial-temporal correlations: The proposed spatial-temporal attention mechanism effectively captures and utilises salient spatial-temporal relationships among different regions in continuous image frames.
3. Superior performance: Extensive experiments demonstrate that the proposed model outperforms state-of-the-art baseline models in both normal and challenging driving scenarios.
4. Lightweight architecture: The proposed model is more compact compared to other sequential models designed for multi-frame input, making it computationally efficient.
5. Robustness to unseen data: Qualitative evaluations show that the model maintains high robustness on entirely new and unlabelled datasets, unseen during training.

3.2 Literature review

Existing studies in the field of lane detection and prediction using vision-based models can be broadly classified into two main categories: (1) classical image processing methods with traditional computer vision techniques, and (2) deep learning based methods with neural network models. This section briefly reviews and summarises some existing works in both categories, and to connect with the proposed method, it also introduces the available attention mechanism applied in vision tasks.

3.2.1 Vision-based lane detection through classical image processing

Before the widespread adoption of machine learning technologies, particularly deep learning, lane detection primarily relied on hand-crafted features and was typically conducted in a four-step process: preprocessing an image, capturing features, detecting lines and fitting them, and postprocessing the image (Bar Hillel et al., 2014; Narote et al., 2018).

The following are some notable studies that exemplify this pipeline. After pre-processing and image enhancement by undergoing grayscale transformation and temporal blurring, Borkar et al. (2009) implemented Inverse Perspective Mapping (IPM) during preprocessing to transform images from a camera perspective to a bird's-eye view. They then utilised the RANSAC

algorithm to eliminate outliers and employed a Kalman filter for lane prediction and smoothing. In a follow-up study (Borkar et al., 2012), they improved their pipeline by introducing a novel time-slicing method to generate ground truth data, together with techniques of integrating pixel remapping, outlier removal, and prediction with tracking. Guo et al. (2015) developed a real-time and efficient lane detection algorithm method based on an improved RANSAC algorithm. Their image preprocessing includes setting a region of interest (ROI) on the video frames, grayscaling the images, and denoising the grey-level images. Then, they adopted the Canny edge detection algorithm to extract significant feature points. Based on the extracted features, they proposed an improved RANSAC algorithm combined with the least-squares technique to estimate parameters for lane modelling using a generalised curve lane parameter model. Tan et al. (2014) proposed a robust curve lane detection method based on the integration of the Improved River Flow (IRF) and RANSAC method. Their approach grouped lane markings into two vision fields: a near-vision field for straight lines and a far-vision field for curved lines. Based on the hyperbola-pair model, the IRF was employed to search feature points in the far vision field, and the RANSAC was adopted to calculate the curvatures to fit curved lane lines during postprocessing. Their experimental results demonstrated that the proposed model can handle certain challenging scenarios, such as dashed lane markings and vehicle occlusion.

As stated, these traditional vision-based lane detection methods through classical image processing rely on hand-crafted features, line detection and fitting, and extensive post-processing, making them inherently time-consuming. Furthermore, the limits of hand-crafted features and the usage of single-image inputs restrict their ability to adapt to dynamic and challenging driving scenarios. Consequently, these methods often fail to deliver the performance and robustness necessary for real-world lane detection applications.

3.2.2 Vision-based lane detection using deep learning methods

Over the past decade, significant advancements in computational power, the availability of large-scale datasets, and the rapid evolution of neural network algorithms have enabled deep learning (DL) methods to achieve remarkable success across various domains, including computer vision, NLP, and speech recognition. As a typical segmentation task in the field of computer vision, lane detection has particularly benefited from these advancements. Numerous vision-based deep learning models have been developed, demonstrating excellent performance and elevating the research in this domain to a brand new level.

DL approaches for lane detection can generally be categorised into four dominant approaches: (1) segmentation-based pipeline (Dong et al., 2023; Kim & Park, 2017; Ko et al., 2022; T. Liu et al., 2020; Pan et al., 2018; J. Zhang et al., 2022; Zou et al., 2020), (2) row-based prediction (Qin et al., 2020; Yoo et al., 2020), (3) anchor-based approach (X. Li et al., 2020; Tabelini et al., 2021; H. Xu et al., 2020), and (4) parametric prediction methods (R. Liu et al., 2021; Tabelini et al., 2020). Despite these approaches varying in methodology, they all share the commonality of leveraging deep neural networks (DNNs). Therefore, the following subsection categorises and reviews these methods based on the principal or dominant neural network structure utilised.

(1) CNN only (or CNN dominant)

CNNs are highly effective at image feature extraction and have become a cornerstone of nearly all computer vision tasks. Treating lane detection as a semantic segmentation task, CNNs are often employed in an end-to-end encoder-decoder framework. For instance, early work by Huval et al. (2015) demonstrated how existing CNN architectures could be applied to lane detection and classification in an end-to-end manner during highway driving. Pan et al. (2018) developed a special convolutional neural network, Spatial CNN (SCNN), which generalises traditional deep layer-by-layer convolutions to slice-by-slice convolutions within feature maps. This architecture proved particularly effective at detecting long, continuous shapes such as traffic lanes, poles, and walls. SCNN achieved outstanding performance, securing first place in the TuSimple Lane Detection Challenge[2], a widely recognised benchmark. Kim and Park (2017) investigated lane detection through a transfer learning framework, constructing a CNN-based encoder-decoder network trained on ImageNet for lane segmentation tasks. Furthermore, utilising the widely adopted backbone ResNet (K. He et al., 2016), Tabelini et al. (2021) developed an anchor-based feature pooling mechanism combined with a novel anchor-based attention approach that aggregates global contextual information, further enhancing lane detection performance.

(2) CNN combined with RNN

Neural network models combining CNN and RNN have been explored to model time-series features in both a single image and sequences of image frames. J. Li et al. (2017) divided road images into a number of continuous slices and employed a convolutional neural network to extract features from each slice. To infer the lane structure from these feature maps, they incorporated an RNN. This combined approach outperformed the use of a CNN alone. However, in this framework, the RNN is employed to model sequence features within a single image, limiting its ability to fully capture temporal dependencies across multiple frames. Zou et al. (2020) examined multiple frames of a continuous driving scene instead of focusing solely on one image. They proposed a hybrid model combining an encoder-decoder CNN with a Convolutional Long Short-Term Memory (ConvLSTM) network, a specific type of RNN. In this architecture, the CNN extracts features from each image frame, while the ConvLSTM processes these CNN-extracted features across multiple consecutive frames, enabling the model to capture both spatial and temporal information for lane prediction. This hybrid CNN-ConvLSTM architecture demonstrated a significant improvement in performance compared to models using only a single image. J. Zhang et al. (2022) developed a similar pipeline but with Convolutional Gated Recurrent Units (ConvGRUs) instead of ConvLSTM. In this model, one ConvGRU block extracts low-level lane features, while another ConvGRU block processes the spatial-temporal information across multiple frames. This approach also showed enhanced results compared to single-image methods.

(3) Generative adversarial network (GAN)

Generative adversarial networks (GANs) (Goodfellow et al., 2014) are primarily used for tasks such as image-to-image translation (Isola et al., 2017; J. Y. Zhu et al., 2017), where unpaired

---

[2] TuSimple Lane Detection Challenge: https://github.com/TuSimple/tusimple-benchmark/tree/master/doc/lane_detection

samples from one image domain are translated to another, and for image data enhancement (Alqahtani et al., 2021; K. Li et al., 2024; Meng et al., 2019), such as improving image quality or augmenting datasets. These networks consist of two neural networks, a generator and a discriminator, that work in opposition: the generator creates images, and the discriminator evaluates them, leading to progressively better image generation through adversarial training. This setup has made GANs highly effective for generating realistic images under various conditions and for tasks like data augmentation in computer vision.

For lane detection, GANs have been utilised to improve environmental adaptability, e.g., in challenging low-light driving conditions. T. Liu et al. (2020) proposed a style-transfer-based data enhancement method, using GANs to generate realistic images that simulate low-light driving scenarios. This model includes three components: Better-CycleGAN, a light condition style transfer network, and a lane detection network. Importantly, this method does not require manual annotations or additional inference computations, offering a more scalable solution for training lane detection systems in diverse conditions. In a different approach, Ghafoorian et al. (2019) employed a GAN for semantic segmentation tasks in lane detection. They introduced an embedding-loss GAN, which consists of a generator that predicts lane structures from input images and a discriminator that evaluates the detection results. The generator and discriminator share weights, which significantly enhances the efficiency of the network while maintaining high performance in lane detection.

3.2.3 Attention mechanism applied in vision tasks

Inspired by the human visual attention mechanism, where humans quickly scan an image to focus on areas of interest while suppressing irrelevant details, recent deep learning models have incorporated artificial attention mechanisms. These models aim to mimic human attention, enabling the neural network to focus on the most task-relevant parts of the input for more effective processing. This approach has led to significant advancements in various machine translation and visual tasks (M. H. Guo et al., 2022; W. He et al., 2021; Luong et al., 2015). Guo et al. (2022) grouped attention methods into six categories, i.e., 1) channel attention which generates attention masks to select important channels (Q. Wang et al., 2020; Yang et al., 2020); 2) spatial attention which considers where to pay attention to by generating attention mask to pick out important spatial regions (Jaderberg et al., 2015; X. Zhu et al., 2019); 3) temporal attention which counts when to pay attention to using attention mask in time to screen out key frames (e.g., Xu et al., 2017; R. Zhang et al., 2019); 4) branch attention that considers which important branches to pay attention to (Y. Chen et al., 2020; X. Li et al., 2019); 5) hybrid combined channel and spatial attention (L. Chen et al., 2017; F. Wang et al., 2017); and 6) hybrid combined spatial and temporal attention (Fu et al., 2019; Gao et al., 2020). For more details, please refer to the survey paper (M. H. Guo et al., 2022).

As for the lane detection task, several studies (Han et al., 2022; R. Liu et al., 2021; Yu et al., 2020; Zhao et al., 2024) have employed Transformer models, which integrate the self-attention mechanism. These models allow the network to focus on important regions of the image, which is particularly useful for lane detection. However, these methods typically still rely on a single image for detection, ignoring the temporal correlations and varying importance of frames in continuous driving scenes.

To sum up, while numerous sophisticated methods have been developed for vision-based lane detection, most of them still rely on one single image for detection, which limits their performance. Even if a few of them had sought to make use of multiple images, key and salient spatial-temporal relevances among continuous image frames are not fully exploited. Consequently, there is still room for improvement, particularly when handling extremely challenging driving scenarios, such as serious occlusion, shape curve, and marking degradation.

3.3 Proposed method

In this section, the overall architecture of the proposed pipeline is first presented in *subsection 3.3.1*, then the elaborated spatial-temporal attention mechanism is described in *subsection 3.3.2*, and lastly, the implementation details are introduced in *subsection 3.3.3*.

3.3.1 Overall architecture description

Inspired by the human visual attention mechanism and considering that traffic lanes are of long thin line structures with strong spatial correlation, for vision-based lane detection, certain regions of the images and certain frames in the continuous driving scenes deserve more attention than other areas and frames. Moreover, it is witnessed that fusing CNN and RNN with hybrid DNN architectures can make use of multiple continuous image frames to further improve lane detection performance (Dong et al., 2023; R. Li & Dong, 2023; J. Zhang et al., 2022; Zou et al., 2020). With all these clues, this study develops a novel dedicated spatial-temporal attention mechanism for the lane detection task to fill the aforementioned research gaps reviewed in Section 2. With the proposed spatial-temporal attention mechanism, three model variants are implemented under hybrid sequential deep end-to-end neural network structures fusing CNN-based encoder-decoder and temporal RNN (e.g., Long Short-term Memory (LSTM) neural network). On the whole, regarding vision-based lane detection as a segmentation task, the proposed model adopts a sequence-to-one architecture, i.e., it takes a sequence of multiple continuous image frames as inputs and outputs the detection result of the final image frame. UNet (Ronneberger et al., 2015), as the standard CNN-based encoder-decoder neural network, serves as the network backbone. In UNet, the encoder module and the decoder module both contain four convolutional blocks (details can be found in **Table 3-1**). The proposed attention module is embedded between the encoder and decoder. The encoder module extracts useful features from the input continuous frames and feeds them to the attention module for further spatial-temporal feature integration. The attention module can detect salient spatial-temporal relevances and dependencies among the extracted feature maps of the consecutive frames and pass these integrated features to the decoder. Lastly, the decoder module upsamples and decodes the integrated feature maps to the same size of the input image and outputs the detected lane lines. Note that, in the adopted UNet backbone, similar to (Ronneberger et al., 2015), a skip connection is applied between the encoder and decoder with a concatenating operation so that the decoder can reuse the extracted features and retained information from the encoder.

An architecture overview of the proposed method pipeline is illustrated in **Figure 3-1**, and detailed implementation is further elaborated in the following sections.

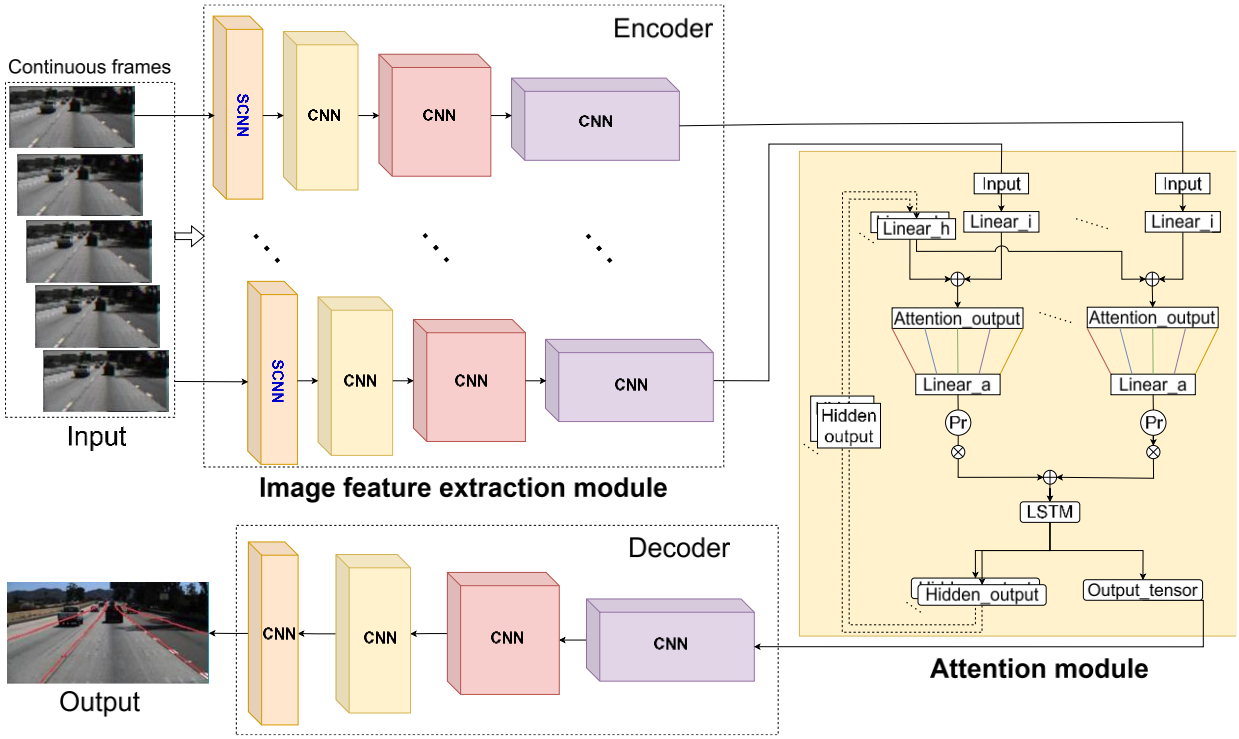


Figure 3-1. The architecture of the proposed pipeline

3.3.2 Spatial-temporal attention mechanism

The proposed attention module is developed to mimic human visual cognitive attention, which demonstrates the ability to focus on important parts and ignore minor parts. The attention module helps the neural network learn to focus on important frames and salient regions of each frame by assigning different weights to each image frame and particular regions of each frame. With the help of the embedded temporal feature extractor, e.g., LSTM or Gated Recurrent Unit (GRU), the attention module can also extract important temporal dependencies over the input consecutive image frames.

As illustrated in **Figure 3-1**, the attention module is applied when the input image sequences are downsized and the features are extracted by a series of convolution layers in the encoder. The attention module integrates the extracted features from the encoder and the hidden outputs produced by the embedded temporal feature extractor, e.g., LSTM/GRU. The LSTM/GRUs' hidden outputs of the very last previous time step and the input feature maps at the current time step are combined using a set of attention weights. Activation of these weights can then be obtained to learn which image frames and which specific regions are important for the lane detection task. The weighted sum of input feature maps highlights the salient features, which are then processed by the temporal feature extractor to produce the output at the current time step and the updated hidden state. All the attention weights can be trained simultaneously together with other neural network layer weights using the backpropagation mechanism. Equations (3-1)-(3-12) provide a formal mathematical description of the attention mechanism as described above.

The output of the final downsized convolutional block at time t for the n-th frame (i.e., timestep n within the image sequence) is denoted as $x_{down4}^{(t-N+n)}$, where $n = \{1, 2, , .., N\}$, and N is the number of frames in the sequence, (in this implementation $N = 5$). The input sequence for the

attention module is therefore defined as $\{x_{down4}^{(t-N+1)}, x_{down4}^{(t-N+2)}, .., x_{down4}^{(t)}\}$. Please note there are two distinct temporal increments. The increment in n corresponds to processing the subsequent image in the input sequence; where an increment in time t reflects the real-world temporal progression, i.e., moving to the next input sequence. Then, within a certain selected sequence, the following computations are performed:

$$x^{(t-N+n)} = Conv\left(x_{down4}^{(t-N+n)},\ k_{in}\right) \tag{3-1}$$

$$z^{(t-N+n)} = \left(U \odot x^{(t-N+n)}\right) + \left(H \odot h^{(t-N+n-1)}\right) \tag{3-2}$$

$$w^{(t-N+n)} = softmax\left(W \odot z^{(t-N+n)}\right) \tag{3-3}$$

$$\overline{x^{(t-N+n)}} = w^{(t-N+n)} \odot x^{(t-N+n)} \tag{3-4}$$

Here, "*Conv*" denotes the convolution operation, $\odot$ is the Hadamard (element-wise) multiplication, while "+" represents the element-wise addition operation. k_{in} is a convolution layer with a kernel of size 1 × 1 and 1 channel (as indicated by *In_Attention_Conv_5_1* in **Table 3-1**). The matrices U, H, W are the learnable weights that can be configured as trainable vectors of size 1 × 1 or 1 × 128, or as a trainable fully connected layer of size 1 × 128. $x^{(t-N+n)}$ and $z^{(t-N+n)}$ represent the intermediate outputs. $h^{(t-N+n-1)}$ is the hidden state vector of the previous step in the temporal feature extractor. $w^{(t-N+n)}$ denotes attention weights obtained from the softmax operation, and $\overline{x^{(t-N+n)}}$ is the attention-based weighted output.

After processing the *N* images and getting the weighted outputs $\{\overline{x^{(t-N+1)}}, \overline{x^{(t-N+2)}}, .., \overline{x^{(t)}}\}$ across the sequence, the following computations are carried out:

$$h^{(t)} = F(\{\overline{x^{(t-N+1)}}, \overline{x^{(t-N+2)}}, .., \overline{x^{(t)}}\},\ h^{(t-N+n-1)}) \tag{3-5}$$

$$x_{out} = Conv(h^{(t)},\ k_{out}) \tag{3-6}$$

where *F* stands for an embedded temporal feature extractor; $h^{(t)}$ is the hidden state vector initialised as $h^{(0)} = \mathbf{0}$ (zero vector) when $t = 0$ and $h^{(t)}$ will be updated with its new inheritor after the selected sequence is fully processed as in equation (3-5); $h^{(t)}$ is also the output from *F* after processing *N* frames, i.e., $h^{(t-N+N)} = h^{(t)}$, which is then expanded to 512 channels by the *outconv* layer k_{out}; k_{out} has a kernel size of 1 × 1 and 512 channels (indicated by *Out_Attention_Conv_5_2* in **Table 3-1**); x_{out} is the final output of the attention module which is then transferred to the decoder module.

The temporal feature extractor *F* can be LSTM or GRU. Take LSTM for example, an LSTM unit is visualised in **Figure 3-2**. Here, *C* is the memory cell, while *i*, *f*, and *o* stand for input gate, forget gate, and output gate, respectively, which regulate the flow of information. The key formulations of the LSTM are shown by equations (3-7)-(3-12):

$$f^{(t-N+n)} = \sigma\left(b^f + P^f\overline{x^{(t-N+n)}} + Q^f h^{(t-N+n-1)}\right) \tag{3-7}$$

$$i^{(t-N+n)} = \sigma\left(b^i + P^i\overline{x^{(t-N+n)}} + Q^i h^{(t-N+n-1)}\right) \tag{3-8}$$

$$\tilde{c}^{(t-N+n)} = g\left(b^c + P^c\overline{x^{(t-N+n)}} + Q^c h^{(t-N+n-1)}\right) \tag{3-9}$$

$$c^{(t-N+n)} = f^{(t-N+n)} \odot c^{(t-N+n-1)} + i^{(t-N+n)} \odot \tilde{c}^{(t-N+n)} \quad (3\text{-}10)$$

$$o^{(t-N+n)} = \sigma\left(b^o + P^o\overline{x^{(t-N+n)}} + Q^o h^{(t-N+n-1)}\right) \quad (3\text{-}11)$$

$$h^{(t-N+n)} = o^{(t-N+n)} \odot g(c^{(t-N+n)}) \quad (3\text{-}12)$$

where $g(\cdot)$ is typically the hyperbolic tangent function, and σ is the activation function. b^f, P^f, Q^f are biases, input weights and recurrent weights for the forget gates; b^i, P^i, Q^i are biases, input weights and recurrent weights for the input gate; b^c, P^c, Q^c are biases, input weights and recurrent weights for the current state of the memory cell; b^o, P^o, Q^o are biases, input weights and recurrent weights for the output gate.

At each time step, the current input vector $\overline{x^{(t-N+n)}}$ and the previous hidden state $h^{(t-N+n-1)}$ are combined and processed to produce the states of the forget gate $f^{(t-N+n)}$ (3-7), input gate $i^{(t-N+n)}$ (3-8), output gate $o^{(t-N+n)}$ (3-11), and the candidate memory update $\tilde{c}^{(t-N+n)}$ (3-9). The forget gate then determines which components of the previous cell state $c^{(t-N+n-1)}$ to retain, while the input gate modulates the contribution of the candidate cell state $\tilde{c}^{(t-N+n)}$, to yield the new cell state $c^{(t-N+n)}$ (3-10). Finally, the output gate filters the activated cell state through a nonlinear function $g(\cdot)$ to produce the hidden state $h^{(t-N+n)}$ (3-12), which is propagated as the recurrent input to the subsequent time step.

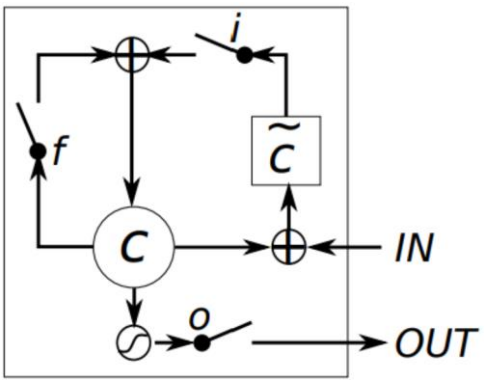


Figure 3-2. An illustration of the Long Short Term Memory unit (Chung et al., 2014)

In the implementation, depending on different settings of the learnable weights U, H, W in Equations (3-2)-(3-3), three variants of the proposed attention module are developed and tested. They are temporal attention (Tem_Att, for short), spatial-temporal attention (ST_Att, for short), and spatial-temporal attention model with fully connected layers (STFC_Att, for short).

The attention model is implemented after the encoder module (to be specific, the fourth down-sampling convolutional block, i.e., *Down_ConvBlock_4* in **Table 3-1**) and before the decoder module (to be specific, the first upsampling convolutional block, i.e., *Up_ConvBlock_4* in **Table 3-1**). Moreover, one should notice that the proposed attention model is modular in nature and can be adopted with any network backbone, not only UNet but also backbones such as SegNet (Badrinarayanan et al., 2017) and fully convolutional networks (Shelhamer et al., 2017).

(1) Temporal attention

The design of the spatial-temporal attention mechanism began with assessing the significance of each frame in a sequence for detecting lane markings in the current frame, which is

implemented through the temporal attention (Tem_Att) module. In this module, the learnable weights of *U*, *H*, and *W* in equations (3-2)-(3-3) are trainable vectors and are illustrated by V_i, V_h and V_a in **Figure 3-3**, respectively. The three trainable vectors, each of size 1×1, dynamically adjust the contributions of input features, hidden state output, and the attention output based on the learned weights. Specifically, the input features $x^{(t-N+n)}$ are modulated by the weight vector V_i and combined with the hidden output multiplied by V_h through element-wise addition to construct a summed intermediate attention signal $z^{(t-N+n)}$, as described in equation (3-2). This attention signal is subsequently passed through a softmax activation function shown as "Pr" in **Figure 3-3** to compute the attention weights $w^{(t-N+n)}$, as defined in equation (3-3). These weights effectively prioritise the significance of each frame in the sequence.

Leveraging the LSTM unit (detailed in equations (3-7)-(3-12)), the hidden state $h^{(t)}$ contextualises the input features by incorporating information from the entire sequence within the selected time window. The attention output $\overline{x^{(t-N+n)}}$ computed as a weighted combination of the input features $x^{(t-N+n)}$ and their respective attention weights $w^{(t-N+n)}$ (see equation (3-4)), captures these temporal dependencies. This output is processed through the LSTM and a convolutional layer (as outlined in equation (3-6)) to generate the module's final output x_{out}, which is subsequently passed to the decoder. When the three trainable vectors V_i, V_h and V_a are of size 1×1, this approach ensures that the model dynamically adapts its focus to relevant temporal features in the image sequence.

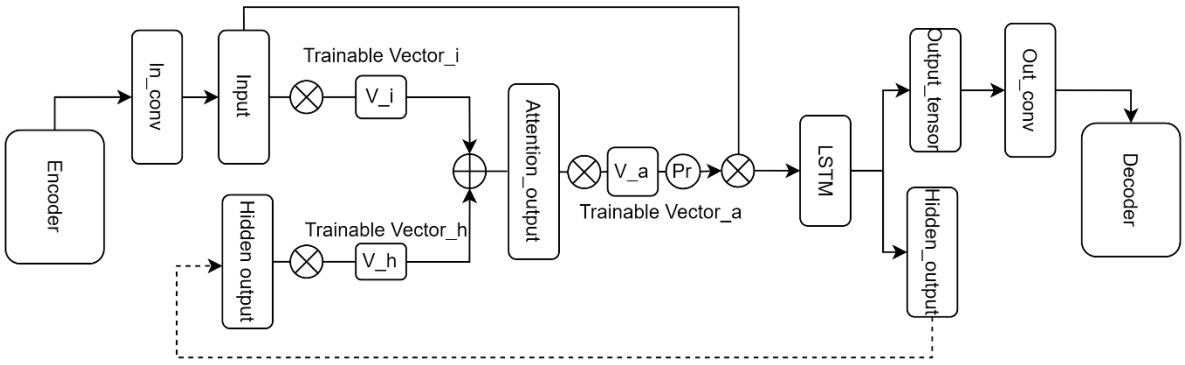


Figure 3-3. An illustration of the temporal attention module (Tem_Att)

(2) Spatial-temporal attention

Observations show that lane lines typically appear in specific regions within image frames, and certain features hold greater significance for accurate detection. To account for this, a spatial attention operation is applied to each frame, upgrading the Tem_Att module into the spatial-temporal attention (ST_Att) module. The ST_Att module introduces three learnable weight vectors, each of size 1×128, which are applied to the input feature matrix, the hidden state output, and the attention output at the current step. These weights, with the size of 1×128, enable the module to prioritise important features within each frame. However, the ST_Att module does not account for the spatial relationships and dependencies between neighbouring feature maps, which are addressed in the subsequent STFC_Att module.

The structure of the ST_Att module is depicted in **Figure 3-4**. While the workflow of ST_Att is similar to Tem_Att, the connections between the input features, hidden state outputs, and

attention outputs, along with their respective weight matrices, follow a one-to-one mapping. These connections are illustrated with colour-coded lines in **Figure 3-4**. Similar to Tem_Att, the attention weights are normalised to a range of 0 to 1 using a softmax activation function (denoted as "Pr" in **Figure 3-4**). The final attention output is processed through a convolutional layer before being passed to the decoder module. This mechanism ensures that the model emphasises the most critical spatial features in each frame. When combined with the temporal modelling capabilities of the LSTM, it effectively leverages spatial-temporal information across image frames in the sequence.

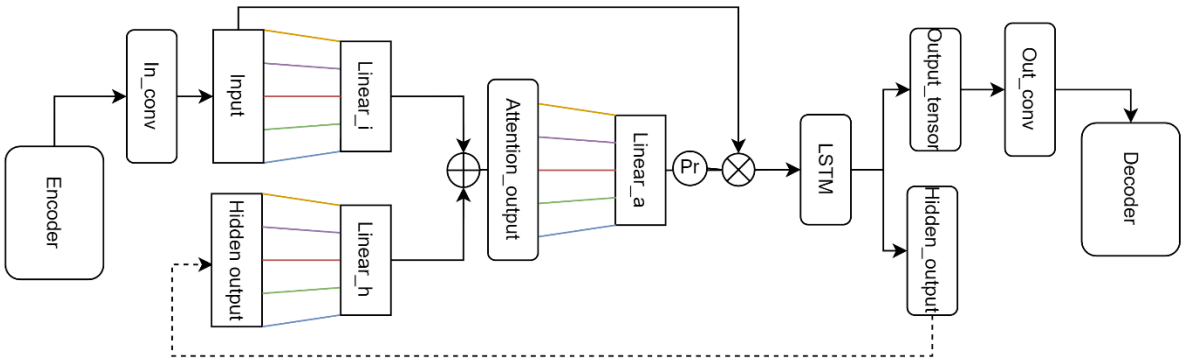


Figure 3-4. An illustration of the spatial-temporal attention module (ST_Att)

(3) Spatial-temporal attention with fully connected layers

The Spatial-Temporal Attention with Fully Connected Layers (STFC_Att) module builds upon the ST_Att module by incorporating a fully connected mechanism to enhance feature learning. Unlike the one-to-one connections in ST_Att, the STFC_Att module employs many-to-many connections, where each learnable weight matrix is multiplied with all values of the input feature map, as illustrated in **Figure 3-5**. This many-to-many connection allows the model to extract spatial dependencies between feature maps within the same image frame while concurrently capturing temporal features and correlations across consecutive frames with the assistance of the LSTM's hidden outputs.

The structure of the STFC_Att module is depicted in **Figure 3-5**, where different coloured lines represent the many-to-many connections between the input feature matrix, the hidden state output, and the attention output, along with their corresponding learnable weight matrices U, H, and W, denoted in **Figure 3-5** as Linear_i, Linear_h, and Linear_a, respectively. Each of these matrices is implemented as a trainable fully connected layer of size 1×128. These weight matrices dynamically adjust the importance of both spatial and temporal features, ensuring a robust and comprehensive feature extraction.

Similar to Tem_Att and ST_Att, in the STFC_Att module, the attention output, denoted as $\overline{x^{(t-N+n)}}$ is calculated using equation (3-4) as a weighted combination of the input features $x^{(t-N+n)}$ and their corresponding attention weights $w^{(t-N+n)}$. Subsequently, the attention output is processed through a linear layer of size 1×128 with many-to-many connections. This output is then scaled to a range of 0 to 1 using the softmax function (denoted as "Pr" in **Figure 3-5**). Finally, the processed attention outputs are passed through a convolutional layer, as outlined in equation (3-6), before being transferred to the decoder.

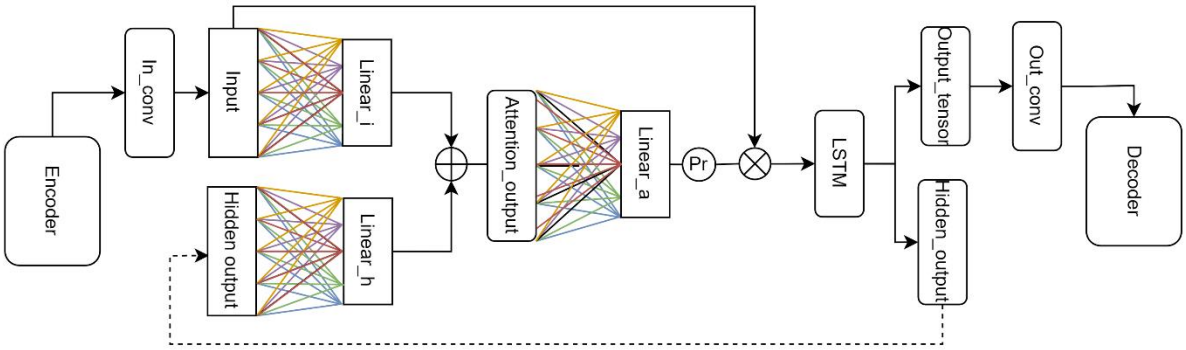


Figure 3-5. An illustration of the spatial-temporal attention module with fully connected layers (STFC_Att)

The key distinction between ST_Att and STFC_Att lies in their ability to capture spatial dependencies. While the ST_Att module focuses on weighting local spatial features within each frame, the fully connected mechanism in STFC_Att extends the network's capability by establishing interrelations between spatial features across the entire frame and throughout the input image sequence. This enhancement allows the model to distinguish among spatial-temporal features more effectively and to focus its attention on the most relevant patterns. Integrating the fully connected spatial-temporal attention mechanism, the STFC_Att module significantly enhances the model's ability to detect lane markings in diverse driving scenarios by leveraging both spatial and temporal interdependencies.

3.3.3 Implementation details

(1) Deep Neural Network Details

On the whole, as illustrated in **Figure 3-1**, the proposed method adopts an "encoder-attention module-decoder"-based sequence-to-one architecture. UNet (Ronneberger et al., 2015) is used as the neural network backbone, in which there are one *In_ConvBlock* and four consecutive down-sampling convolutional blocks (i.e., *Down_ConvBlock_1*, *Down_ConvBlock_2*, *Down_ConvBlock_3*, and *Down_ConvBlock_4*) in the encoder part, and four symmetrical upsampling convolutional blocks (i.e., *Up_ConvBlock_4*, *Up_ConvBlock_3*, *Up_ConvBlock_2*, and *Up_ConvBlock_1*) in the decoder part. Between the encoder and the decoder, there is the attention module with a temporal feature extractor (e.g., LSTM) embedded. **Table 3-1** illustrates in detail the input and output sizes, as well as the parameters of each layer in the entire DNN.

(2) Loss function

Vision-based lane detection can be considered as the pixel-wise binary classification problem, for which cross-entropy is a suitable loss function (Ho & Wookey, 2020). It is important to note that, in most cases, the pixels classified as "*lanes*" are far fewer than those classified as "*not lanes*" (i.e., the background), which makes it an unbalanced discriminative binary classification problem. Therefore, this study adopts the weighted cross-entropy as the loss function with two rescaling weights given to each class. The two weights for lane class and background class are set to the inverse proportion of the number of pixels in the two classes, i.e., there are fewer lane

Table 3-1. Architecture and layer-specific parameter settings of the neural network

Layer		Input (channel×hight×width)	Output (channel×hight×width)	Kernel	Padding	Stride	Activation
In_ConvBlock	In_Conv_1	3×128×256	64×128×256	3×3	(1,1)	1	ReLU
	In_Conv_2	64×128×256	64×128×256	3×3	(1,1)	1	ReLU
Down_Conv Block_1	Maxpool_1	64×128×256	64×64×128	2×2	(0,0)	2	---
	Down_Conv_1_1	64×64×128	128×64×128	3×3	(1,1)	1	ReLU
	Down_Conv_1_2	128×64×128	128×64×128	3×3	(1,1)	1	ReLU
Down_Conv Block_2	Maxpool_2	128×64×128	128×32×64	2×2	(0,0)	2	---
	Down_Conv_2_1	128×32×64	256×32×64	3×3	(1,1)	1	ReLU
	Down_Conv_2_2	256×32×64	256×32×64	3×3	(1,1)	1	ReLU
Down_Conv Block_3	Maxpool_3	256×32×64	256×16×32	2×2	(0,0)	2	---
	Down_Conv_3_1	256×16×32	512×16×32	3×3	(1,1)	1	ReLU
	Down_Conv_3_2	512×16×32	512×16×32	3×3	(1,1)	1	ReLU
Down_Conv Block_4	Maxpool_4	512×16×32	512×8×16	2×2	(0,0)	2	---
	Down_Conv_4_1	512×8×16	512×8×16	3×3	(1,1)	1	ReLU
	Down_Conv_4_2	512×8×16	512×8×16	3×3	(1,1)	1	ReLU
Attention Module	In_Attention_ Conv_5_1	512×8×16	1×8×16	1×1	---	1	---
	AttentionLayer_1	1×128*	1×128*	---	---	---	---
	AttentionLayer_2	1×128*	1×128*	---	---	---	---
	AttentionLayer_3	1×128*	1×128*	---	---	---	---
	LSTM	128	128	---	---	---	---
	Out_Attention_ Conv_5_2	1×8×16	512×8×16	1×1	---	1	---
Up_Conv Block_4	Upsampling Bilinear2D_1	512×8×16	512×16×32	2×2	(0,0)	2	---
	Up_Conv_4_1	1024×16×32	256×16×32	3×3	(1,1)	1	ReLU
	Up_Conv_4_2	256×16×32	256×16×32	3×3	(1,1)	1	ReLU
Up_Conv Block_3	Upsampling Bilinear2D_2	256×16×32	256×32×64	2×2	(0,0)	2	---
	Up_Conv_3_1	512×32×64	128×32×64	3×3	(1,1)	1	ReLU
	Up_Conv_3_2	128×32×64	128×32×64	3×3	(1,1)	1	ReLU
Up_Conv Block_2	Upsampling Bilinear2D_3	128×32×64	128×64×128	2×2	(0,0)	2	---
	Up_Conv_2_1	256×64×128	64×64×128	3×3	(1,1)	1	ReLU
	Up_Conv_2_2	64×64×128	64×64×128	3×3	(1,1)	1	ReLU
Up_Conv Block_1	Upsampling Bilinear2D_4	64×64×128	64×128×256	2×2	(0,0)	2	---
	Up_Conv_1_1	128×128×256	64×128×256	3×3	(1,1)	1	ReLU
	Up_Conv_1_2	64×128×256	64×128×256	3×3	(1,1)	1	ReLU
Out_Conv Block	Out_Conv	64×128×256	2×128×256	1×1	(0,0)	1	---

**This is an example of the spatial-temporal attention (ST_Att) module. Corresponding to three attention variants, parameters in AttentionLayer_1, AttentionLayer_2, and AttentionLayer_3 will be learnable vectors of size 1 × 1 for Tem_Att, learnable vectors of size 1 × 128 for ST_Att, or learnable vectors with many to many connections of size 1 × 128 for STFC_Att, respectively.*

pixels than the background, so the weight of the lane class is larger. The adopted weighted binary cross-entropy loss function is illustrated by equation (3-13).

$$Loss = -\frac{1}{M}\sum_{m=1}^{M}\left[w_l * y_m * log\left(h_\theta(x_m)\right) + w_{nl} * (1\text{-}y_m) * log\left(1 - h_\theta(x_m)\right)\right] \quad (3\text{-}13)$$

where M is the number of training examples; w_l stands for the weight of the lane class, while w_{nl} for the background (not lane) class; y_m is the true target label for the training example m; x_m is the input for the training example m; and h_θ is the neural network model with weights θ.

(3) Training details

Various variants of the developed neural network model, as well as selected baseline models, had been trained and tested on the Dutch national high-performance supercomputer cluster Lisa using four Titan RTX GPUs with the data trained parallelly using *torch.nn.DataParallel()* in the PyTorch library. The input image size is set as 128 × 256, and the training batch size is set as 64. The learning rate is initially set to 0.01 with decay applied after each epoch. The Adam (Kingma & Ba, 2015), RAdam optimiser (Liyuan Liu et al., 2020), and Stochastic Gradient Descent (SGD) (Bottou, 2010) optimisers were tested. Experiments demonstrated that SGD delivered the smallest loss in this study. Thus, the SGD optimiser was chosen, and the momentum term was applied.

3.4 Experiments and results

To verify the effectiveness and robustness of the proposed model with the designed attention module, extensive experiments were carried out on three commonly used large-scale open-source datasets, i.e., TuSimple, tvtLANE (Zou et al., 2020), and LLAMAS (Behrendt & Soussan, 2019) datasets. Several DNN-based lane detection models, e.g., LaneNet (Neven et al., 2018), SCNN (Pan et al., 2018), Seg-Net (Badrinarayanan et al., 2017), UNet (Ronneberger et al., 2015), SegNet_ConvLSTM (Zou et al., 2020), and UNet_ConvLSTM (Zou et al., 2020), were selected as the baselines.

3.4.1 Test on tvtLANE and TuSimple datasets

(1) Dataset description

The original dataset of the *TuSimple Lane Detection Challenge* consists of 3,626 training and 2,782 testing one-second clips that are collected under different driving conditions. Each clip is extracted into 20 continuous frames, and only the last frame, i.e., the 20th frame, is labelled with the ground truth. Additionally, Zou et al. (2020) added the label of the 13th frame and augmented the TuSimple dataset with 1,148 additional clips (with also the 13th and 20th frames labelled) regarding rural road driving scenes collected in China. Moreover, rotation, flip, and crop operations are employed for data augmentation, and finally, a total number of (3,626+1,148)×4=19,096 sequences were produced, among which 38,192 frames are labelled with ground truth.

For testing, there are 2,782 testing clips in the original TuSimple dataset. While in tvtLANE, there are two different testing sets, namely, testset #1 which is based on the original TuSimple test set for normal driving scene testing, as well as testset #2 which contains 12 challenging driving scenarios for testing challenging scenes and assessing the model robustness.

In the training phase, three different sampling strides, with an interval of 1, 2, and 3 frames respectively, were adopted to adapt to different driving speeds which also augment the training samples by three times, whereas in the test phase, the sampling stride was set as a fixed interval of 1 frame.

Detailed descriptions of the two datasets and sampling settings can be found in (Dong et al., 2023; Zou et al., 2020).

(2) Qualitative evaluation

As the intuitive evaluation approach with visualisation, in this subsection, qualitative lane detection results of different models are demonstrated in the figure visualisations. The figure demonstrations help identify the strengths and weaknesses of the evaluated models and provide insights.

1) Results on tvtLANE testset #1: normal driving scene testing

Samples of the results from lane detection segmentation on tvtLANE testset #1 are shown in **Figure 3-6**. The lane lines are segmented into white pixels, while the background is displayed in black pixels. Three proposed attention-based model variants and the baseline deep learning models were tested. Here in **Figure 3-6**, all of the results are without post-processing, which also applies to all the visualisations and quantitative evaluations discussed later in this study.

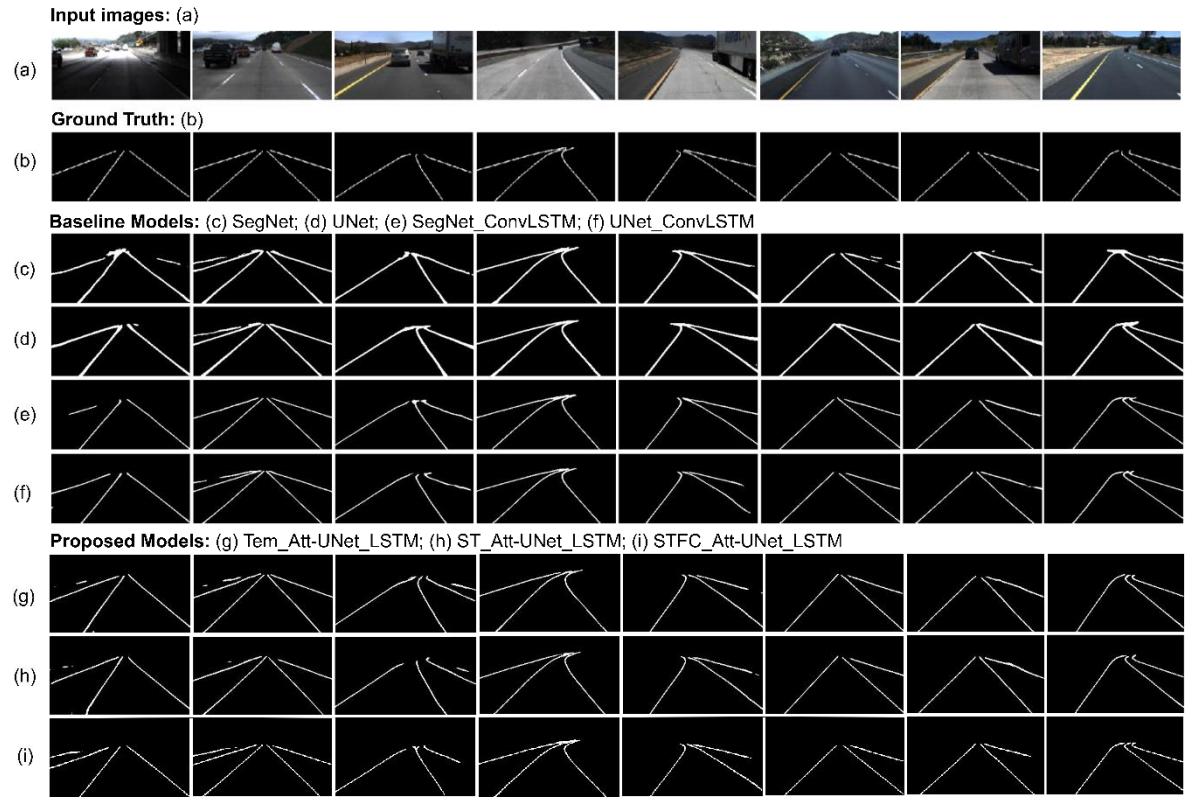


Figure 3-6. Qualitative evaluation 1: Comparison of the results of lane detection on tvtLANE testset #1 (normal situations)

Qualitatively, the models should be able to a) correctly predict the number of lanes; b) accurately locate the lane lines in the segmentation image; c) segment the lanes in thin lines without blurs; and d) keep proper continuity without unexpected breaks in continuous lanes. Regarding these aspects, the proposed models with attention mechanisms all deliver good

results, especially the STFC_Att-based model, indicated in the last row (i), which outputs the thinner lane lines with good continuity and fewer blurs. One may argue that it does not detect the correct number of lanes in the first two columns from the left. However, when zooming in for details, one can identify that the model correctly detects the left road boundary lanes which are too difficult and even not labelled in the ground truth. This defect with the labelled ground truth in the dataset is also discussed in (J. Zhang et al., 2022).

Furthermore, in accordance with previous studies (Dong et al., 2023; J. Zhang et al., 2022; Zou et al., 2020), models using multi-continuous image frames generally outperform models using a single frame, as the latter output thick lines with heavy blurs.

2) Results on tvtLANE testset #2: challenging scenes

According to **Figure 3-7**, the proposed model is compared qualitatively with the baseline models when faced with some extremely challenging driving scenarios (tested on the tvtLANE testset #2). Involving a broad range of challenging situations, the testset #2 is a separate new dataset which is unseen during the training phase. It is observed that all the models do not perform well, especially regarding the 3rd column where there are vehicle occlusions and dirt road surfaces simultaneously. However, similar to norm scenes, the proposed attention-based models overall surpass baselines with thinner continuous lines and more correct locations and lane numbers. Typically, shown in the 4th column of **Figure 3-7**, the STFC_Att-UNet_LSTM model demonstrates superior results in detecting smooth clear lines with the correct number of lanes in the serious vehicle occlusion case, in which almost all the other models are defeated. This can be inferred by its capability of exploring spatial-temporal correlations among neighbouring pixels.

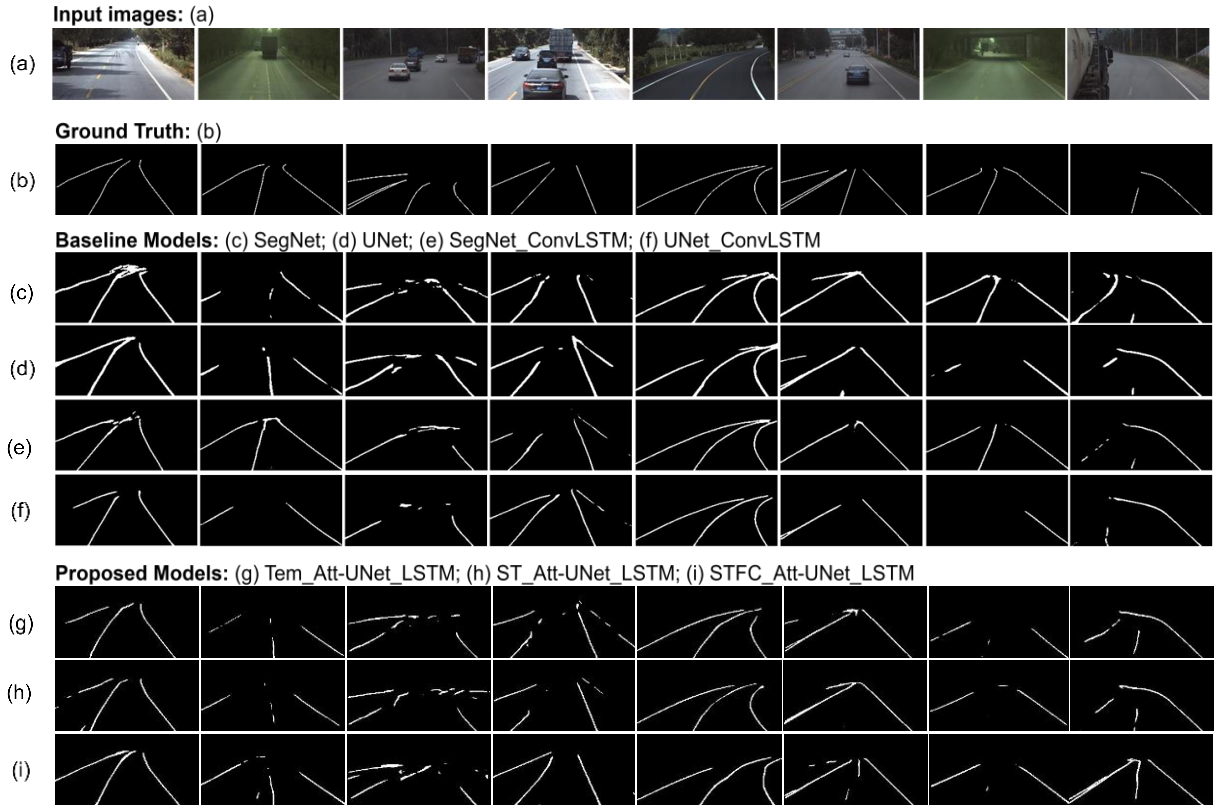


Figure 3-7. Qualitative evaluation 2: Comparison of the lane detection results on tvtLANE testset #2 (challenging situations)

3) Results on TuSimple testing set

As mentioned before, the TuSimple testing set is similar to the tvtLANE testset#1, thus, similar patterns are observed in **Figure 3-8**. Compared with the baseline model UNet_ConvLSTM, the proposed models can detect more correct lane lines with fewer blurs.

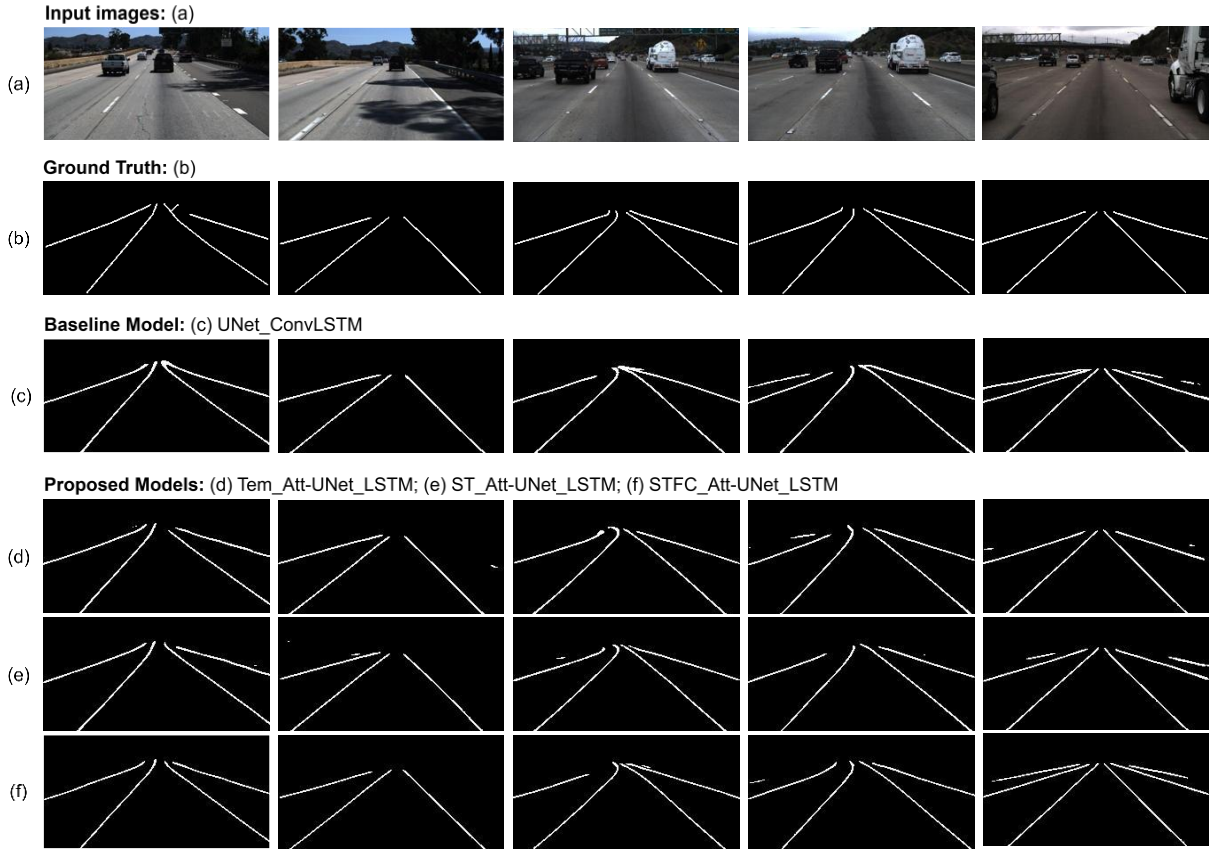


Figure 3-8. Qualitative evaluation 3: Comparison of the lane detection results on the TuSimple testing set

(3) Quantitative evaluation

1) Evaluation metrics

Treating the vision-based lane detection as a pixel-wise unbalanced two-class classification and discriminative segmentation task, and following the convention in previous studies (Dong et al., 2023; Lizhe Liu et al., 2021; Pan et al., 2018; H. Xu et al., 2020; J. Zhang et al., 2022; Zou et al., 2020), this study utilises four commonly adopted evaluation criteria, i.e., accuracy, precision, recall, and F1-measure, to quantitatively verify the proposed models. The four criteria are illustrated in equations (3-14)-(3-17):

$$\text{Accuracy} = \frac{Truly\ Classified\ Pixels}{Total\ Number\ of\ Pixels} \tag{3-14}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \tag{3-15}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \tag{3-16}$$

$$\text{F1-measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{3-17}$$

where true positive correlates to the pixels that are accurately identified as lanes, false-positive indicates the number of background pixels that are incorrectly categorised as lanes, and false negative is for the number of lane pixels that were incorrectly categorised as background.

This study also provides the size of the model parameter, referred to as Params (M), as well as multiply-accumulate (MAC) operations, referred to as MACs (G), as indicators for estimating the computational complexity and capabilities of the models for real-time performance.

2) Quantitative performance comparison on tvtLANE testset #1 (normal situations)

As shown in **Table 3-2**, when testing on tvtLANE testset#1, all the developed attention-based models perform better than the baselines regarding F1-measure, accuracy, and precision. This verifies the effectiveness of the proposed attention mechanism. The developed model STFC_Att-SCNN_UNet_LSTM (which will be discussed in detail in the ablation study) performs the best with the highest F1-measure, accuracy, and precision. Furthermore, compared to the other two baseline models, i.e., UNet_ConvLSTM and SegNet_ConvLSTM, which also adopt multiple frames as inputs, the developed models are all smaller in parameter size and with fewer MACs. This means that the developed models can deliver better results while using lower computational resources and with higher processing speed.

Table 3-2. Model quantitative performance comparison on tvtLANE testset #1 (normal situations)

Model		Test_Acc (%)	Precision	Recall	F1-measure	MACs (G)	Params (M)
	Baseline Models						
Using single image	UNet	96.54	0.790	**0.985**	0.877	15.5	13.4
	SegNet	96.93	0.796	0.962	0.871	50.2	29.4
	SCNN*	96.79	0.654	0.808	0.722	77.7	19.2
	LaneNet*	97.94	0.875	0.927	0.901	44.5	19.7
Using continuous image frames	SegNet_ConvLSTM	97.92	0.874	0.931	0.901	217.0	67.2
	UNet_ConvLSTM	98.00	0.857	0.958	0.904	69.0	51.1
	Proposed Models						
	Tem_Att-UNet_LSTM	98.08	0.877	0.936	0.906	44.7	13.5
	ST_Att-UNet_LSTM	98.09	0.879	0.941	0.909	44.8	13.5
	STFC_Att-UNet_LSTM	**98.14**	**0.887**	0.941	**0.911**	44.9	13.5
	*STFC_Att-SCNN_UNet_LSTM***	***98.20***	***0.906***	*0.936*	***0.921***	*68.9*	*13.7*

** Results reported in (J. Zhang et al., 2022).*

*** Model variant used for ablation study.*

Tem_Att-UNet_LSTM means the temporal attention based model using the UNet_LSTM network backbone. This naming rule also applies to other models.

3) Quantitative performance comparison on tvtLANE testset #2 (challenging scenes)

For testing model robustness, the developed models were also evaluated and verified on the brand-new dataset, namely the tvtLANE testset #2, which contains 12 challenging scenes.

As shown in **Table 3-3**, in terms of precision, ST_Att-UNet_LSTM performs the best in *bright*, *curve*, and *urban* scenes, while STFC_Att-UNet_LSTM performs the best in *occluded*, *shadow* and *tunnel* scenes. Therefore, they dominate half of the 12 challenging scenes.

Table 3-3. Model quantitative performance comparison on tvtLANE testset #2 (12 challenging scenes)

Precision												
Model \ Challenging Scenes	1-curve & occlude	2-shadow - bright	3-bright	4-occlude	5-curve	6-dirty & occlude	7-urban	8-blur & curve	9-blur	10-shadow - dark	11-tunnel	12-dim & occlude
UNet	0.7018	0.7441	0.6717	0.6517	0.7443	0.3994	0.4422	0.7612	0.8523	0.7881	0.7009	0.5968
SegNet	0.6810	0.7067	0.5987	0.5132	0.7738	0.2431	0.3195	0.6642	0.7091	0.7499	0.6225	0.6463
UNet_ConvLSTM	0.7591	0.8292	0.7971	0.6509	0.8845	0.4513	**0.5148**	0.8290	**0.9484**	0.9358	0.7926	**0.8402**
SegNet_ConvLSTM	0.8176	0.8020	0.7200	0.6688	0.8645	**0.5724**	0.4861	0.7988	0.8378	0.8832	0.7733	0.8052
Tem_Att-UNet_LSTM	**0.8430**	**0.8909**	0.7732	0.5740	0.8322	0.4692	0.4567	**0.8358**	0.8090	0.9244	0.7893	0.8046
ST_Att-UNet_LSTM	0.7938	0.8743	**0.8013**	0.7014	**0.8894**	0.5215	0.4935	0.8290	0.8517	0.9286	0.7516	0.8218
STFC_Att-UNet_LSTM	0.8239	0.8782	0.7646	**0.7031**	0.8871	0.5295	0.4848	0.7354	0.9023	**0.9395**	**0.8794**	0.7542
F1-measure												
UNet	0.8200	0.8408	0.7946	0.7337	0.7827	0.3698	0.5658	0.8147	0.7715	0.6619	0.5740	0.4646
SegNet	0.8042	0.7900	0.7023	0.6127	0.8639	0.2110	0.4267	0.7396	0.7286	0.7675	0.6935	0.5822
UNet_ConvLSTM	0.8465	**0.8891**	**0.8411**	0.7245	0.8662	0.2417	**0.5682**	0.8323	**0.7852**	0.6404	0.4741	0.5718
SegNet_ConvLSTM	0.8852	0.8544	0.7688	0.6878	**0.9069**	**0.4128**	0.5317	0.7873	0.7575	**0.8503**	**0.7865**	**0.7947**
Tem_Att-UNet_LSTM	**0.8933**	0.8657	0.8123	0.6513	0.8306	0.3530	0.5263	0.8290	0.7039	0.5338	0.5225	0.5226
ST_Att-UNet_LSTM	0.8548	0.8977	0.8253	0.7293	0.8254	0.3627	0.5543	**0.8369**	0.7480	0.6197	0.5522	0.5363
STFC_Att-UNet_LSTM	0.8690	0.9059	0.8314	**0.7456**	0.8086	0.3660	0.5277	0.7715	0.7329	0.6543	0.6471	0.5852

High precision means the model is more strict for the pixels to be classified as lane lines, i.e., fewer False Positives. This is crucial for the vehicles' localising lanes. However, being too strict might result in more False Negatives, then a lower recall ratio, and then a worse F1-measure. This is why the developed models are not good in terms of F1-measure. Furthermore, it is witnessed that during the training process, all the models obtained higher recalls and lower precisions at the beginning. Then, as the training went on, the recalls decreased while the precisions rose. This general pattern applies to all models. With this, one can infer that a higher precision is more important. All these demonstrate the developed models' robustness over challenging scenes.

4) Performance and comparisons on the TuSimple testing set

The aforementioned TuSimple testing set has similar but more testing samples compared to the tvtLANE testset#1. Regarding the quantitative results on the TuSimple testing set, as demonstrated in **Table 3-4**, the proposed STFC_Att-UNet_LSTM obtains the best F1-measure, the best precision, and the second-best accuracy (i.e., 98.20%, only a bit lower than the best of 98.22%). Although UNet_ConvLSTM shows the best accuracy, it is worth noting that its MACs and parameter size are much larger than the proposed models. In this case, one can conclude that the developed models with lower computational complexities are robust with competitive results on the TuSimple testing set.

Table 3-4. Model quantitative performance comparison on TuSimple testing set

Model		Test_Acc (%)	Precision	Recall	F1-measure	MACs (G)	Params (M)
Using continuous image frames	**Baseline Models**						
	SegNet_ConvLSTM*	97.96	0.852	0.964	0.901	217.0	67.2
	UNet_ConvLSTM*	**98.22**	0.857	0.958	0.904	69.0	51.1
	UNet_DoubleConvGRU*	98.04	0.875	0.953	0.912	---	13.4
	Proposed Models						
	Tem_Att-UNet_LSTM	98.05	0.876	0.923	0.899	44.7	13.5
	ST_Att-UNet_LSTM	98.14	0.881	0.925	0.902	44.8	13.5
	STFC_Att-UNet_LSTM	***98.20***	**0.886**	0.950	**0.917**	44.9	13.5

** Results reported in (J. Zhang et al., 2022).*

3.4.2 Test on LLAMAS dataset

(1) Dataset description

To further verify the robustness of the proposed method, the LLAMAS dataset (Behrendt & Soussan, 2019) is adopted to train, validate, and test different models. Consisting of a total of 100,042 images, LLAMAS is one of the largest open-source lane marker datasets. Among the 100,042 images, 79,113 of them are used for training with labelled ground truth, while 20,929 of them were originally used for testing with no corresponding labels. To still follow the proposed end-to-end supervised learning pipeline and make it comparable with the previous work (J. Zhang et al., 2022), this study follows the processes described in (J. Zhang et al., 2022), utilising only the labelled 79,113 images and dividing them into two groups. To be specific, 58,269 images were used for training, and 20,844 images were used for testing. More details about the LLAMAS dataset can be found in (Behrendt & Soussan, 2019; J. Zhang et al., 2022).

(2) Qualitative evaluation

Limited by computational resources and time, this study only trained ST_Att-UNet_LSTM and STFC_Att-UNet_LSTM models on the LLAMAS dataset. **Figure 3-9** provides the qualitative visualisation results of ST_Att-UNet_LSTM for testing on the LLAMAS dataset. In the top row, the predicted lane lines are shown in red colour, and in the bottom row, the predicted lane lines are segmented with white pixels under black background. As shown, the lane lines in LLAMAS are labelled in a different way using dashed lines, which makes it much more challenging. Qualitatively, from the visualisation, one can observe that there are very few false positives and the lane lines are generally predicted accurately.

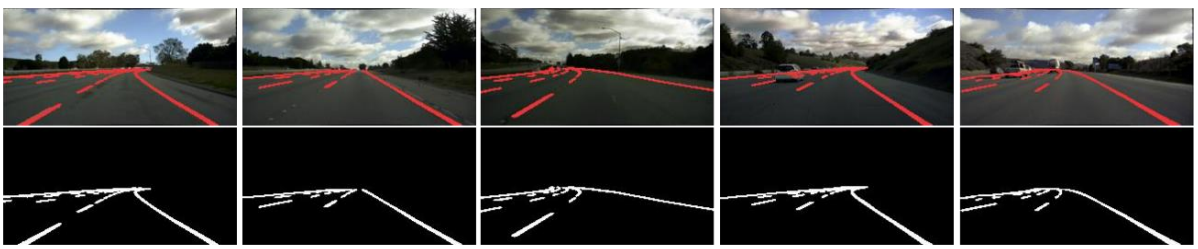


Figure 3-9. Qualitative evaluation 4: Lane detection results on the LLAMAS dataset

(3) Quantitative evaluation

To quantitatively evaluate the model performances on the LLAMAS dataset, except for the aforementioned precision and recall, similar to (Behrendt & Soussan, 2019; J. Zhang et al., 2022), average precision (AP) was also adopted. AP is the mean of the weighted precision scores at different thresholds. The weights are the differences in recalls from the prior tested thresholds. To be clear, AP is illustrated in equation (3-18)

$$\mathrm{AP} = \sum_{p=1}^{T} \sum_{q=1}^{V+1} (Precision_p * \Delta Recall_q) \tag{3-18}$$

where T means the total number of the tested image frames; V means the number of pixel samples for a single image; $\Delta Recall$ represents the difference between the *Recall* values of two consecutive samples. The variables p and q are subscripts to number the samples. In the implementation, similar to (J. Zhang et al., 2022), this study sets $Recall_0$ to 0, $Precision_0$ to 1, and the variable V to 100.

The quantitative results are demonstrated in **Table 3-5**.

Table 3-5. Model quantitative performance comparison on the LLAMAS dataset

Model	Average Precision (AP)	Precision	Recall
UNet_Double_ConvGRU*	0.**8519**	0.6162	0.6163
SegNet ConvLSTM*	0.8500	0.5487	**0.6839**
UNet_ConvLSTM*	0.8510	0.5857	0.6558
ST_Att-UNet_LSTM	0.7106	0.6253	0.6584
STFC_Att-UNet_LSTM	0.7141	**0.6317**	0.6413

** Results reported in (J. Zhang et al., 2022).*

As shown in **Table 3-5**, the STFC_Att-UNet_LSTM model provides the best corner precision when testing on the LLAMAS dataset. This is an indication that the model delivers a lower number of false positives, which, as discussed before, is more crucial for lane localisation. It also obtains a comparable corner recall. Furthermore, it is worth noting that both the proposed models maintain a better balance among the three evaluation metrics, although they do not perform well in average precision. To sum up, the developed models' robustness on the LLAMAS dataset is demonstrated with competitive quantitative and qualitative detection results.

3.4.3 Qualitative test on unlabelled Netherlands lane dataset

To further verify the developed models' robustness in handling new and challenging driving scenes, the unlabelled Netherlands lane dataset was adopted for qualitative testing. This dataset covers a wide range of driving situations in the Netherlands, some of which are very challenging.

Figure 3-10 shows the lane detection results of ST_Att-UNet_LSTM, which is only trained on the LLAMAS dataset. Even without any supervised training on the unlabelled Netherlands lane dataset, the proposed model demonstrates excellent transfer capabilities by clearly detecting lane line numbers and locations. Furthermore, the model can correctly identify whether the lanes are continuous or dashed lanes. The good performance can be attributed to that the

developed ST_Att-UNet_LSTM with spatial-temporal attention module can aggregate rich valuable context information to focus on generalised information and salient regions in both one image and the continuous image frames. This qualitative testing further verifies the robustness of the developed model.

Figure 3-10. Qualitative evaluation 5: Lane detection results on the unlabelled Netherlands lane dataset

3.5 Ablation study and discussion

3.5.1 Post-explanation of the attention mechanism by visualisation

To elucidate the functionality of the proposed spatial-temporal attention mechanism, this subsection presents a case study using feature map visualisations. Consider a scenario where a vehicle is travelling under a bridge, as depicted in **Figure 3-11**, shadows cast by the bridge obscure portions of the road, rendering lane markings indiscernible (even to human observers). Furthermore, on the right side, lane markings are partially occluded by a preceding vehicle.

The top row (a) displays the original sequence of continuous image frames, illustrating the vehicle's gradual movement beneath the bridge from frame 1 to frame 5 (left to right). Rows (b), (c), and (d) compare the feature map activations at *Up_ConvBlock_4* (the first upsampling block as shown in **Table 3-1**) for UNet, UNet-ConvLSTM, and STFC_Att_UNet_LSTM, respectively. Since all three models share the *Up_ConvBlock_4* structure, which immediately follows the attention module in STFC_Att_UNet_LSTM and the ConvLSTM module in UNet-ConvLSTM, this comparison provides a meaningful evaluation of their performance.

The baseline UNet model exhibits strong activation primarily along the leftmost lane, with detection on the right appearing fragmented. Critically, in distant regions, the detected left and right lanes converge erroneously, accompanied by blurred and spurious activations. This limitation reflects the model's insufficient contextual reasoning for maintaining coherent lane structures in occluded and distant areas. By integrating ConvLSTM, UNet_ConvLSTM demonstrates an improved ability to detect lane markings in occluded regions through temporal dependency modelling. However, its overall activation intensity remains subdued relative to the baseline UNet, suggesting constraints in its spatial feature extraction and a suboptimal spatial-temporal correlation. In contrast, STFC_Att_UNet_LSTM outperforms both models by maintaining consistent, non-converging activation patterns, particularly excelling in scenarios with partial or complete occlusions. Its spatial-temporal attention mechanism can dynamically weigh the importance of each frame based on lane visibility, enabling robust inference of lane positions even in challenging scenarios. This enhanced performance is attributed to the model's capacity to establish strong interrelations among spatial features across sequential frames, effectively "memorising" lane positions from previous observations.

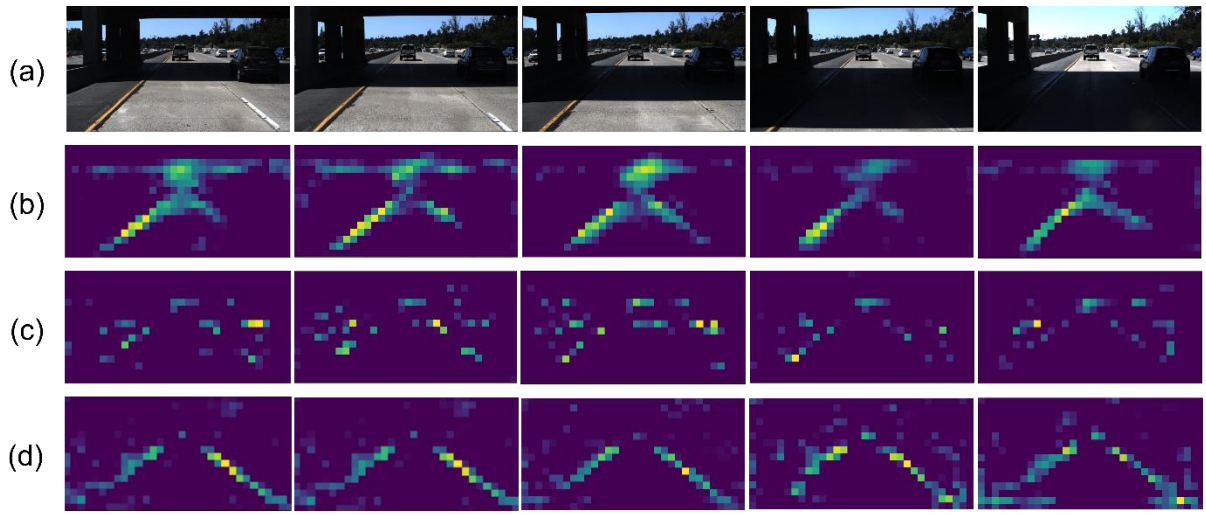


Figure 3-11. Post-explanation visualisation of the case study - under a bridge with shadows and occlusion: (a) input images; and feature map visualisations of (b) UNet, (c) UNet_ConvLSTM, (d) STFC_Att_UNet_LSTM

Ultimately, the spatial-temporal attention-based neural network leverages information from prior frames to predict lane locations, even when they are entirely obscured in the current frame (e.g., the 5th frame in row (a)). As shown in frames 1-4, lane markings are partially visible, allowing the model to detect salient regions of interest (highlighted by pronounced bright activations in **Figure 3-11**). By leveraging this accumulated spatial-temporal information, the model accurately predicts lane positions in frame 5. This ability to retain and utilise learned spatial-temporal correlations across frames enhances robustness in adverse driving conditions.

3.5.2 The comparisons between the three model variants

Comparing the three proposed model variants' results in **Tables 3-2**, **3-3**, **3-4**, and **3-5**, it is demonstrated that STFC_Att-UNet_LSTM outperforms Tem_Att-UNet_LSTM and ST_Att-UNet_LSTM in various situations and regarding different metrics; while ST_Att-UNet_LSTM is also generally better than Tem_Att-UNet_LSTM. This can be explained by the fact that Tem_Att-UNet_LSTM only gets the temporal attention mechanism which does not consider the interrelationship among the pixels and different regions; while, in the ST_Att-UNet_LSTM, with the one-to-one connection, it can learn the importance of the individual pixel with the weights and hidden layer but without the knowledge of neighbouring pixels; and finally, in the STFC_Att-UNet_LSTM, using the many-to-many connection, the spatial dependencies between the pixels are incorporated, along with the temporal correlations among the continuous frames. Thus, the STFC_Att-UNet_LSTM is the real "spatial-temporal" attention, and in this way, the verification of the strengths of the proposed spatial-temporal attention mechanism is further enhanced.

3.5.3 Cooperation with other model structures and methods

This study also investigated the compatibility of the proposed model working with other mechanisms, such as incorporating the SCNN layer to further enhance feature extraction and spatial correlation within individual images (Dong et al., 2023; Pan et al., 2018). The last row

of **Table 3-2** shows the results for STFC_Att-SCNN_UNet_LSTM**, which incorporates the SCNN layer into the proposed spatial-temporal attention mechanism. Compared to all other models, including those with the attention mechanism but without SCNN layers, STFC_Att-SCNN_UNet_LSTM** achieves the highest accuracy, precision, and F1-measure, with only minor increases in parameter size and slight increases in MACs. These findings demonstrate that embedding SCNN layers further strengthens the model's performance, confirming the compatibility of the proposed spatial-temporal attention mechanism. As the developed spatial-temporal is modular in nature, it should be able to cooperate with any other mechanisms. Additionally, the proposed encoder-decoder-based pipeline allows all the developed models to integrate seamlessly with other methodologies, such as self-supervised pre-training approaches. In particular, the employment of pre-training using the masked sequential autoencoders (R. Li & Dong, 2023) was shown to significantly improve the performance of the STFC_Att-SCNN_UNet_LSTM model. Incorporating self-supervised pre-training not only enhances the model's accuracy but also substantially reduces total training time (R. Li & Dong, 2023), further demonstrating the flexibility and adaptability of the proposed approach.

3.5.4 Model size and real-time capability

As illustrated in **Table 3-2**, all the developed models with the proposed attention mechanism possess fewer parameters and lower MACs compared with the baseline SegNet_ConvLSTM and UNet_ConvLSTM, which also use continuous frames as input. Fewer parameters and lower MACs mean the models get better performance regarding processing time and real-time capability, which would be advantageous when deployed in real-world applications.

Within the developed model variants, Tem_Att-UNet_LSTM, ST_Att-UNet_LSTM, and STFC_Att-UNet_LSTM have nearly identical parameter sizes (the little difference can not be visible in one decimal), while ST_Att-UNet_LSTM and STFC_Att-UNet_LSTM get slightly larger MACs, with STFC_Att-UNet_LSTM getting the largest among the three variants. These variations arise from differences in model architecture. Within the group of the developed models, as the MACs increase, the model's performance generally gets better (demonstrated in **Tables 3-2**, **3-3**, **3-4**, and **3-5**), making the slight computational cost increase acceptable.

The results from all these ablation experiments provide robust evidence of the effectiveness and reliability of the proposed spatial-temporal attention mechanism for lane detection. The mechanism strikes a balance between model complexity and real-time capability, ensuring practical viability in real-world diverse driving scenarios.

3.6 Conclusion

Previous vision-based methods for lane detection often fail to account for critical image regions and their spatial-temporal salience across continuous frames, leading to poor performance under challenging driving scenarios. In this study, a novel spatial-temporal attention mechanism embedded within a hybrid sequence-to-one encoder-decoder neural network architecture is proposed and implemented for accurate and robust lane detection in a variety of normal and challenging driving scenarios. The proposed spatial-temporal attention mechanism can focus on key features of lane lines and exploit salient spatial-temporal correlations among continuous frames to enhance the accuracy and robustness of lane detection. Extensive experiments

conducted on three large-scale open-source datasets demonstrate the robustness and superiority of the proposed model, outperforming available state-of-the-art methods in various testing scenarios. In addition, ablation studies confirm the developed spatial-temporal attention mechanism's capabilities of cooperating with other architectures and model mechanisms. Last but not least, the sequential neural network models implemented by the proposed spatial-temporal attention mechanism possess fewer parameters and smaller multiply-accumulate operations compared with other sequential baseline models, highlighting their computational efficiency.

However, it is observed that the proposed models struggle in certain challenging cases. These cases are underrepresented in the training datasets and, in some instances, include mislabelled ground truth data (as noted in (J. Zhang et al., 2022)), hindering the models' ability to learn their patterns. Furthermore, initial tests of the models' transferability between datasets revealed that the models trained on tvtLANE underperformed on the LLAMAS dataset due to differences in lane structures and labelling formats, and vice versa. In contrast, models trained on the LLAMAS dataset performed well on the unlabelled Netherlands lane dataset due to similar lane structures.

For real-world deployment of lane detection algorithms, adaptability to diverse lane structures and types across different countries and regions is crucial. With these findings in mind, it is recommended that future research should focus on building integrated, comprehensive lane datasets and exploring domain adaptation and transfer learning methods to improve lane detection performance across diverse datasets and driving conditions.

Acknowledgements

This study was supported by the Dutch Institute for Scientific Research (NWO) through its subdomain in the Applied and Technical Sciences (TTW) under the project Safe and Efficient Operation of Automated and Human-Driven Vehicles in Mixed Traffic (SAMEN) with the contract number of 17187.

References

Alqahtani, H., Kavakli-Thorne, M., & Kumar, G. (2021). Applications of Generative Adversarial Networks (GANs): An updated review. Archives of Computational Methods in Engineering. https://doi.org/10.1007/s11831-019-09388-y

Aly, M. (2008). Real time detection of lane markers in urban streets. IEEE Intelligent Vehicles Symposium, Proceedings, 7–12. https://doi.org/10.1109/IVS.2008.4621152

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Bai, M., Mattyus, G., Homayounfar, N., Wang, S., Lakshmikanth, S. K., & Urtasun, R. (2018). Deep multi-sensor lane detection. IEEE International Conference on Intelligent Robots and Systems, 3102–3109. https://doi.org/10.1109/IROS.2018.8594388

Bar Hillel, A., Lerner, R., Levi, D., & Raz, G. (2014). Recent progress in road and lane detection: A survey. Machine Vision and Applications, 25(3), 727–745. https://doi.org/10.1007/s00138-011-0404-2

Behrendt, K., & Soussan, R. (2019). Unsupervised labeled lane markers using maps. Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019. https://doi.org/10.1109/ICCVW.2019.00111

Berriel, R. F., de Aguiar, E., de Souza, A. F., & Oliveira-Santos, T. (2017). Ego-Lane Analysis System (ELAS): Dataset and algorithms. Image and Vision Computing. https://doi.org/10.1016/j.imavis.2017.07.005

Borkar, A., Hayes, M., & Smith, M. T. (2009). Robust lane detection and tracking with RANSAC and Kalman filter. Proceedings - International Conference on Image Processing, ICIP. https://doi.org/10.1109/ICIP.2009.5413980

Borkar, A., Hayes, M., & Smith, M. T. (2012). A novel lane detection system with efficient ground truth generation. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2011.2173196

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT 2010 - 19th International Conference on Computational Statistics, Keynote, Invited and Contributed Papers. https://doi.org/10.1007/978-3-7908-2604-3_16

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2017). SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. https://doi.org/10.1109/CVPR.2017.667

Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., & Liu, Z. (2020). Dynamic convolution: Attention over convolution kernels. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR42600.2020.01104

Chetan, N. B., Gong, J., Zhou, H., Bi, D., Lan, J., & Qie, L. (2020). An overview of recent progress of lane detection for autonomous driving. Proceedings - 2019 6th International Conference on Dependable Systems and Their Applications, DSA 2019. https://doi.org/10.1109/DSA.2019.00052

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. 1–9. http://arxiv.org/abs/1412.3555

Dong, Y., Patil, S., van Arem, B., & Farah, H. (2023). A hybrid spatial-temporal deep learning architecture for lane detection. Computer-Aided Civil and Infrastructure Engineering, 38(1), 67–86. https://doi.org/10.1111/mice.12829

Fu, Y., Wang, X., Wei, Y., & Huang, T. (2019). STA: Spatial-temporal attention for large-scale video-based person re-identification. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 8287-8294. https://doi.org/10.1609/aaai.v33i01.33018287

Gao, L., Li, X., Song, J., & Shen, H. T. (2020). Hierarchical LSTMs with adaptive attention for visual captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2019.2894139

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems.

Guo, J., Wei, Z., & Miao, D. (2015). Lane detection method based on improved RANSAC algorithm. Proceedings - 2015 IEEE 12th International Symposium on Autonomous Decentralized Systems, ISADS 2015. https://doi.org/10.1109/ISADS.2015.24

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., Zhang, S. H., Martin, R. R., Cheng, M. M., & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. Computational Visual Media, 8(3), 331–368. https://doi.org/10.1007/s41095-022-0271-y

Han, J., Deng, X., Cai, X., Yang, Z., Xu, H., Xu, C., & Liang, X. (2022). Laneformer: Object-aware row-column transformers for lane detection. Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022. https://doi.org/10.1609/aaai.v36i1.19961

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 770–778. https://doi.org/10.1109/CVPR.2016.90

He, W., Wu, Y., & Li, X. (2021). Attention mechanism for neural machine translation: A survey. IEEE Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2021. https://doi.org/10.1109/ITNEC52019.2021.9586824

Ho, Y., & Wookey, S. (2020). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2962617

Hou, Y., Ma, Z., Liu, C., & Loy, C. C. (2019). Learning lightweight lane detection CNNS by self attention distillation. Proceedings of the IEEE International Conference on Computer Vision, 1013–1021. https://doi.org/10.1109/ICCV.2019.00110

Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., Andriluka, M., Rajpurkar, P., Migimatsu, T., Cheng-Yue, R., Mujica, F., Coates, A., & Ng, A. Y. (2015). An empirical evaluation of deep learning on highway driving. 1–7. http://arxiv.org/abs/1504.01716

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. https://doi.org/10.1109/CVPR.2017.632

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial Transformer networks. Advances in Neural Information Processing Systems.

Jiao, X., Yang, D., Jiang, K., Yu, C., Wen, T., & Yan, R. (2019). Real-time lane detection and tracking for autonomous vehicle applications. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering. https://doi.org/10.1177/0954407019866989

Kim, J., & Park, C. (2017). End-to-end ego lane estimation based on sequential transfer learning for self-driving cars. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. https://doi.org/10.1109/CVPRW.2017.158

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Ko, Y., Lee, Y., Azam, S., Munir, F., Jeon, M., & Pedrycz, W. (2022). Key points estimation and point instance segmentation approach for lane detection. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2021.3088488

Li, J., Mei, X., Prokhorov, D., & Tao, D. (2017). Deep neural network for structural prediction and lane detection in traffic scene. IEEE Transactions on Neural Networks and Learning Systems, 28(3), 690–703. https://doi.org/10.1109/TNNLS.2016.2522428

Li, K., Dai, Z., Wang, X., Song, Y., & Jeon, G. (2024). GAN-based controllable image data augmentation in low-visibility conditions for improved roadside traffic perception. IEEE Transactions on Consumer Electronics, 70(3), 6174–6188. https://doi.org/10.1109/TCE.2024.3387557

Li, R., & Dong, Y. (2023). Robust lane detection through self pre-training with masked sequential autoencoders and fine-tuning with customized PolyLoss. IEEE Transactions on Intelligent Transportation Systems, 24(12), 14121–14132. https://doi.org/10.1109/TITS.2023.3305015

Li, X., Li, J., Hu, X., & Yang, J. (2020). Line-CNN: End-to-end traffic line detection with line proposal unit. IEEE Transactions on Intelligent Transportation Systems, 21(1), 248–258. https://doi.org/10.1109/TITS.2019.2890870

Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2019.00060

Liu, Liyuan, He, P., Chen, W., & Tech, G. (2020). On the variance of the adaptive learning rate and beyond. International Conference on Learning Representations, 1–13.

Liu, Lizhe, Chen, X., Zhu, S., & Tan, P. (2021). CondLaneNet: A top-to-down lane detection framework based on conditional convolution. Proceedings of the IEEE International Conference on Computer Vision. https://doi.org/10.1109/ICCV48922.2021.00375

Liu, R., Yuan, Z., Liu, T., & Xiong, Z. (2021). End-to-end lane shape prediction with transformers. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 3694–3702.

Liu, T., Chen, Z., Yang, Y., Wu, Z., & Li, H. (2020). Lane detection in low-light conditions using an efficient data enhancement: Light conditions style transfer. IEEE Intelligent Vehicles Symposium, Proceedings. https://doi.org/10.1109/IV47402.2020.9304613

Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, 1412–1421. https://doi.org/10.18653/v1/d15-1166

Meng, Y., Kong, D., Zhu, Z., & Zhao, Y. (2019). From night to day: GANs based low quality image enhancement. Neural Processing Letters. https://doi.org/10.1007/s11063-018-09968-2

Narote, S. P., Bhujbal, P. N., Narote, A. S., & Dhane, D. M. (2018). A review of recent advances in lane detection and departure warning system. Pattern Recognition. https://doi.org/10.1016/j.patcog.2017.08.014

Neven, D., De Brabandere, B., Georgoulis, S., Proesmans, M., & Van Gool, L. (2018). Towards end-to-end lane detection: An instance segmentation approach. IEEE Intelligent Vehicles Symposium, Proceedings, 2018-June, 286–291. https://doi.org/10.1109/IVS.2018.8500547

Pan, X., Shi, J., Luo, P., Wang, X., & Tang, X. (2018). Spatial as deep: Spatial CNN for traffic scene understanding. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 7276–7283.

Qin, Z., Wang, H., & Li, X. (2020). Ultra fast structure-aware deep lane detection. Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, 12369 LNCS, 276–291. https://doi.org/10.1007/978-3-030-58586-0_17

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9351, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Satzoda, R. K., Sathyanarayana, S., Srikanthan, T., & Sathyanarayana, S. (2010). Hierarchical additive Hough transform for lane detection. IEEE Embedded Systems Letters. https://doi.org/10.1109/LES.2010.2051412

Shelhamer, E., Long, J., & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2016.2572683

Sivaraman, S., & Trivedi, M. M. (2013). Integrated lane and vehicle detection, localization, and tracking: A synergistic approach. IEEE Transactions on Intelligent Transportation Systems, 14(2), 906–917. https://doi.org/10.1109/TITS.2013.2246835

Tabelini, L., Berriel, R., Paixão, T. M., Badue, C., de Souza, A. F., & Oliveira-Santos, T. (2021). Keep your eyes on the lane: real-time attention-guided lane detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 294–302. https://doi.org/10.1109/CVPR46437.2021.00036

Tabelini, L., Berriel, R., Paixão, T. M., Badue, C., de Souza, A. F., & Oliveira-Santos, T. (2020). PolyLaneNet: Lane estimation via deep polynomial regression. Proceedings - International Conference on Pattern Recognition, 6150–6156. https://doi.org/10.1109/ICPR48806.2021.9412265

Tan, H., Zhou, Y., Zhu, Y., Yao, D., & Li, K. (2014). A novel curve lane detection based on Improved River Flow and RANSA. 2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014. https://doi.org/10.1109/ITSC.2014.6957679

Tang, J., Li, S., & Liu, P. (2021). A review of lane detection methods based on deep learning. Pattern Recognition, 111, 107623. https://doi.org/10.1016/j.patcog.2020.107623

Wang, B. F., Qi, Z. Q., & Ma, G. C. (2014). Robust lane recognition for structured road based on monocular vision. Journal of Beijing Institute of Technology (English Edition), 23(3), 345–351.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., & Tang, X. (2017). Residual attention network for image classification. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. https://doi.org/10.1109/CVPR.2017.683

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR42600.2020.01155

Wang, Y., Dahnoun, N., & Achim, A. (2012). A novel system for robust lane detection and tracking. Signal Processing, 92(2), 319–334. https://doi.org/10.1016/j.sigpro.2011.07.019

Xu, H., Wang, S., Cai, X., Zhang, W., Liang, X., & Li, Z. (2020). CurveLane-NAS: Unifying lane-sensitive architecture search and adaptive point blending. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16 (pp. 689-704). Springer International Publishing. https://doi.org/10.1007/978-3-030-58555-6_41

Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., & Zhou, P. (2017). Jointly attentive spatial-temporal pooling networks for video-based person re-identification. Proceedings of the IEEE International Conference on Computer Vision. https://doi.org/10.1109/ICCV.2017.507

Yang, Z., Zhu, L., Wu, Y., & Yang, Y. (2020). Gated channel transformation for visual recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR42600.2020.01181

Yeong, D. J., Velasco-hernandez, G., Barry, J., & Walsh, J. (2021). Sensor and sensor fusion technology in autonomous vehicles: A review. In Sensors. https://doi.org/10.3390/s21062140

Yoo, S., Seok Lee, H., Myeong, H., Yun, S., Park, H., Cho, J., & Hoon Kim, D. (2020). End-to-end lane marker detection via row-wise classification. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2020-June, 4335–4343. https://doi.org/10.1109/CVPRW50498.2020.00511

Yu, Z., Ren, X., Huang, Y., Tian, W., & Zhao, J. (2020). Detecting lane and road markings at a distance with perspective Transformer layers. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC 2020. https://doi.org/10.1109/ITSC45102.2020.9294383

Zhang, J., Deng, T., Yan, F., & Liu, W. (2022). Lane detection model based on spatio-temporal network with double convolutional gated recurrent units. IEEE Transactions on Intelligent Transportation Systems, 23(7), 6666–6678. https://doi.org/10.1109/TITS.2021.3060258

Zhang, R., Li, J., Sun, H., Ge, Y., Luo, P., Wang, X., & Lin, L. (2019). SCAN: Self-and-collaborative attention network for video person re-identification. IEEE Transactions on Image Processing. https://doi.org/10.1109/TIP.2019.2911488

Zhao, J., Qiu, Z., Hu, H., & Sun, S. (2024). HWLane: HW-Transformer for lane detection. IEEE Transactions on Intelligent Transportation Systems, 25(8), 9321–9331. https://doi.org/10.1109/TITS.2024.3386531

Zheng, F., Luo, S., Song, K., Yan, C. W., & Wang, M. C. (2018). Improved lane line detection algorithm based on Hough transform. Pattern Recognition and Image Analysis, 28(2), 254–260. https://doi.org/10.1134/S1054661818020049

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision. https://doi.org/10.1109/ICCV.2017.244

Zhu, X., Cheng, D., Zhang, Z., Lin, S., & Dai, J. (2019). An empirical study of spatial attention mechanisms in deep networks. Proceedings of the IEEE International Conference on Computer Vision. https://doi.org/10.1109/ICCV.2019.00679

Zou, Q., Jiang, H., Dai, Q., Yue, Y., Chen, L., & Wang, Q. (2020). Robust lane detection from continuous driving scenes using deep neural networks. IEEE Transactions on Vehicular Technology, 69(1), 41–54. https://doi.org/10.1109/TVT.2019.2949603

4 Robust lane detection through self pre-training with masked sequential autoencoders and fine-tuning with customised PolyLoss

Abstract

Lane detection is crucial for vehicle localisation which makes it the foundation for automated driving and many intelligent and advanced driving assistant systems. Available vision-based lane detection methods do not make full use of the valuable features and aggregate contextual information, especially the interrelationships between lane lines and other regions of the images in continuous frames. To fill this research gap and upgrade lane detection performance, this study proposes a pipeline consisting of self-supervised pre-training with masked sequential autoencoders (MSAEs) and fine-tuning with customised PolyLoss for the end-to-end neural network models using multi-continuous image frames. The MSAEs are adopted to pre-train the neural network models with reconstructing the missing pixels from a random masked image as the objective. Then, in the fine-tuning segmentation phase where lane detection segmentation is performed, the continuous image frames serve as the inputs, and the pre-trained model weights are transferred and further updated using the backpropagation mechanism with customised PolyLoss calculating the weighted errors between the output lane detection results and the labelled ground truth. Extensive experiment results demonstrate that, with the proposed pipeline, the lane detection model performance on both normal and challenging scenes can be advanced beyond the state-of-the-art, delivering the best testing accuracy (98.38%), precision (0.937), and F1-measure (0.924) on the normal scene testing set, together with the best overall accuracy (98.36%) and precision (0.844) in the challenging scene test set, while the training time can be substantially shortened.

This chapter is based on the journal publication:

Li, R., & Dong, Y. (2023). Robust Lane Detection Through Self Pre-Training with Masked Sequential Autoencoders and Fine-Tuning with Customized PolyLoss. IEEE Transactions on Intelligent Transportation Systems, 24(12), 14121-14132. https://doi.org/10.1109/TITS.2023.3305015 (Co-first authors and corresponding author)*

4.1 Introduction

Lane detection is one of the crucial parts of automated driving and is the foundation of many intelligent and advanced driving assistant systems. However, lane detection has always been a challenging task, for complex and variable realistic road conditions (these scenes are easily disturbed by factors including shadows, degraded road signs, blocking, poor lighting, and bad weather), and the curved and elongated features of lane lines (Zou et al., 2020).

In recent years, many deep learning models have been proposed for vision-based lane detection (Y. Zhang et al., 2022). Before the emergence of deep learning, traditional methods mainly utilise traditional computer vision techniques, which rely on manually manipulated operators to extract handcrafted features, including geometry (Borkar et al., 2011; Y. Wang et al., 2004), colour (Somawirata & Utaminingrum, 2017), etc., to do the detection, and then refine the results using a series of fitting methods, such as Hough transform (Zheng et al., 2018) and B-spline fitting (Cao et al., 2019). Although some progress had been made, traditional methods are not robust to complex and challenging traffic scenes. In contrast, deep learning based methods can extract more favourable features automatically and achieve superior performance in a variety of complex environments (Y. Zhang et al., 2022). Generally, deep learning approaches are currently developed from three main perspectives: segmentation-based (Dong et al., 2022; Feng et al., 2022; Lee & Liu, 2023; Li et al., 2021; Pan et al., 2018; Patil et al., 2022; Ren et al., 2022; H. Wang et al., 2022; Zang et al., 2018; J. Zhang et al., 2021; Zou et al., 2020), anchor-based (Huang et al., 2023; Jin et al., 2022; Qin et al., 2022; Tabelini et al., 2021), and parameter-based (Liu et al., 2021; Torres et al., 2020), among which the most commonly used approach is the segmentation-based method. The performance of segmentation-based methods for lane detection has been continuously improving with various neural network structures developed. Getting rid of dense layers, Fully Convolutional Networks (FCNs) (Long et al., 2015; Zang et al., 2018) employ solely locally connected layers, e.g., convolution, pooling, and upsampling, to enable efficient learning of input images with arbitrary sizes, which makes it well-suited for the varying input images of lane detection. Spatial convolutional neural network (SCNN) (Pan et al., 2018) adopts customised spatial convolutional layers using slice-by-slice convolutions for message passing to capture essential spatial information and correlation for lane detection. UNet-based (Dong et al., 2022; Lee & Liu, 2023; Patil et al., 2022; Ronneberger et al., 2015; J. Zhang et al., 2021; Zou et al., 2020) neural networks with symmetrical encoder-decoder structures can extract features at multiple scales, leading to accurate identification of lane markings of different sizes and shapes. Using similar symmetrical encoder-decoder structures, SegNet-based (Al Mamun et al., 2021; Badrinarayanan et al., 2017; Gad et al., 2020) models employ pooling indices for upsampling, reducing trainable parameters and memory requirements. Generative Adversarial Neural Network (GAN) (Ghafoorian et al., 2018) with embedding loss can preserve label-resembling qualities and improve the outputs' realism and structure preservation, reducing the need for complex post-processing in lane detection.

On the other hand, self-supervised learning has shown in recent studies (Bao et al., 2022; El-Nouby et al., 2021; He et al., 2022; Xie et al., 2022) that learning a generic feature representation by self-supervision can enable the downstream tasks to achieve highly desirable performance. The basic idea, masking and then reconstructing, is to input a masked set of image patches to the neural network model and then reconstruct the masked patches at the output, allowing the model to learn more valuable features and aggregate contextual information. When it comes to vision-based lane detection, self-supervised learning can provide stronger feature

characterisation by exploring interrelationships between lane lines and other regions of the images in the continuous frames for the downstream lane detection task. With self-supervised pre-training, it is also possible to accelerate the model convergence in the training phase reducing training time. Meanwhile, with the aggregated contextual information and valuable features by pre-training, the lane detection results can be further advanced.

In this study, a self-supervised pre-training paradigm is investigated for boosting the lane detection performance of the end-to-end encoder-decoder neural network using multi-continuous image frames. The masked sequential autoencoders (MSAEs) are adopted to pre-train the neural network model by reconstructing the missing pixels from a randomly masked image with mean squared error (MSE) as the loss function. The pre-trained model weights are then transferred to the fine-tuning segmentation phase of the per-pixel image segmentation task, in which the transferred model weights are further updated using backpropagation with a customised PolyLoss calculating the weighted errors between the output lane detection results and the labelled ground truth. With this proposed pipeline, the model performance for lane detection on both normal and challenging scenes is advanced beyond the state-of-the-art results by considerable margins.

The main contributions of this study are as follows:

1. This study proposes a robust lane detection pipeline through self-supervised pre-training with masked sequential autoencoders (MSAEs) and fine-tuning with customised PolyLoss, and verifies its effectiveness by extensive comparison experiments.
2. A customised PolyLoss is developed and adopted to further improve the capability of the neural network model. Without any extra parameter tuning, the customised PolyLoss can bring a significant improvement in the lane detection segmentation task while substantially accelerating model convergence speed and reducing the training time.
3. The whole pipeline is tested and verified using three deep neural network structures, i.e., UNet_ConvLSTM (Zou et al., 2020), SCNN_UNet_ConvLSTM (Dong et al., 2022), and SCNN_UNet_Attention (Patil et al., 2022), with the SCNN_UNet_Attention based model delivering the best detection results for normal testing scenes, while SCNN_UNet_ConvLSTM model delivering the best detection results for challenging scenes surpassing baseline models.

4.2 Proposed method

This study proposes a pipeline for lane detection through self-supervised with MSAEs and fine-tuning segmentation with customised PolyLoss. In the first stage, the images are randomly masked as the inputs, and the neural network model is pre-trained with reconstructing the complete images as the objective. In the second stage, the pre-trained neural network model weights are transferred to the segmentation neural network model with the same backbone, and only the structure of the output layer is adjusted. In this phase, continuous image frames without any masking are served as inputs. The neural network weights are further updated and fine-tuned by minimising PolyLoss with the backpropagation mechanism. In this study, three neural network models, i.e., UNet_ConvLSTM (Zou et al., 2020), SCNN_UNet_ConvLSTM (Dong et al., 2022), and SCNN_UNet_Attention (Patil et al., 2022), are tested. In the last stage, post-processing methods, e.g., Density-based spatial clustering of applications with noise (DBSCAN)

(Ester et al., 1996) for clustering the lane types and curve fitting to smooth the detected lines, are proposed to further improve the overall performance of the detection. However, due to time constraints and computational restrictions and following the convention in literature, e.g., (Dong et al., 2022; J. Zhang et al., 2021; Zou et al., 2020), post-processing is not specifically explored in this study. The framework of the proposed pipeline is illustrated in **Figure 4-1**. In the remaining parts of this section, each phase will be introduced in detail.

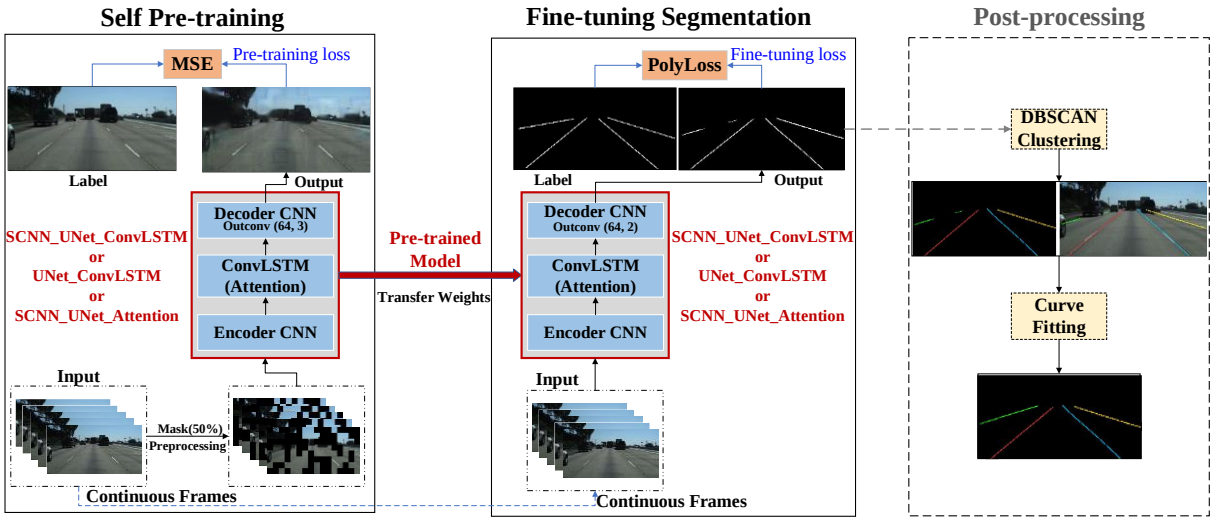


Figure 4-1. The framework of the proposed pipeline

4.2.1 Preliminary and network backbone

This study tests the proposed pipeline with three hybrid neural network models based on the UNet (Ronneberger et al., 2015) backbone, i.e, UNet_ConvLSTM (Zou et al., 2020), SCNN_UNet_ConvLSTM (Dong et al., 2022), and SCNN_UNet_Attention (Patil et al., 2022). The three models are in similar structures, composing three parts, i.e., encoder Convolutional Neural Network (CNN), Convolutional Long Short-Term Memory (ConvLSTM) block or Attention block, and decoder CNN, and they both work in an end-to-end approach.

Encoder-decoder is a widely used framework in the field of deep learning with various network structures. It is capable of mapping directly from the original input to the desired output in an end-to-end manner and keeping the input and output of the same size. Such a framework has demonstrated good performances in natural language processing tasks, e.g., machine translation, summary extraction, and computer vision tasks, e.g., target detection, scene perception, and image segmentation (Dong et al., 2022; Ronneberger et al., 2015; J. Zhang et al., 2021; Zou et al., 2020). Lane detection as a typical image semantic segmentation or instance segmentation task can surely be tackled with super results under the encoder-encoder structure, e.g., (Dong et al., 2022; Patil et al., 2022; J. Zhang et al., 2021; Zou et al., 2020).

A commonly used base neural network backbone for lane detection (and also other image segmentation tasks) is the UNet (Ronneberger et al., 2015), which is an improved FCN. UNet with a symmetric encoder-decoder structure was originally developed to solve the problem of medical image segmentation. In UNet, a block of its encoder contains two convolutional layers, and the feature map is downsampled using pooling layers to reduce the feature map size and increase the number of channels. The decoder, which is symmetric with the encoder, performs

deconvolution and upsampling operations for feature recovery and data reconstruction. The decoder CNNs have the same size and number of feature maps as in the encoder but are arranged in the opposite direction, and the feature maps are appended in a direct manner. With a symmetrical CNN-based encoder-decoder structure, UNet is widely used in various aspects of segmentation tasks, including lane detection, with outstanding performance.

However, the original pure UNet does not consider the slender spatial structure and the correlations and continuity of lane lines in continuous image frames. To tap the temporal continuity of the lane line detection, the ConvLSTM module is embedded between the encoder-decoder in the UNet_ConvLSTM model (Zou et al., 2020), which can integrate the time series features extracted from the input multi-continuous frames. To further improve lane detection results, SCNN_UNet_ConvLSTM (Dong et al., 2022) incorporates SCNN in its single image feature extraction module to make use of the spatial correlations of lane structure and achieves state-of-the-art performance. SCNN_UNet_Attention (Patil et al., 2022) which applies a spatial-temporal attention module with linear LSTM in the middle of the encoder and decoder rather than ConvLSTM, can further exploit spatial-temporal correlations and dependencies of different image regions among different frames in the continuous image sequence, and further advance the detection performance. This study implemented and tested UNet_ConvLSTM, SCNN_UNet_ConvLSTM, and SCNN_UNet_Attention models to verify the proposed pipeline.

4.2.2 Self pre-training with Masked Sequential Autoencoders (MSAEs)

For vision-based lane detection, in most of the driving scene image frames, lane lines only account for a small fraction of the whole image, which means there is more spatial redundancy compared to other segmentation tasks. It is vital but challenging to make full use of the valuable features and aggregate contextual information, especially the interrelationships between lane lines and other regions of the images in continuous frames.

He et al. (2022) show that taking advantage of a pre-training strategy by randomly masking a high proportion of input image and reconstructing the original image from the masked patches using the latent representations can improve accuracy and accelerate training speed for downstream tasks. That is, the images with a high masking rate are input into the designed model for reconstruction as a self-supervised learning task, and then the pre-trained model can be migrated to the downstream tasks for fine-tuning. With this pre-training method, the model can gain a better overall "understanding" of the images, since reconstructing the masked pixels in the pre-training phase facilitates the trained model with a good generalisation capability, which can serve for downstream tasks.

Inspired by and upgraded upon the idea of self-training by "random masking-reconstructing" with autoencoders (He et al., 2022), this study proposed to incorporate a pre-training phase with MSAEs to pre-train the lane detection models and facilitate their capabilities in aggregating contextual information for feature extraction through continuous frames. In the pre-training phase, S (for the experiments carried out in this study, $S = 5$) consecutive images are used as the inputs, with every image getting certain parts randomly masked. To implement the masking, each of the input images with the size of (128×256) is first divided into non-overlapping patches with the size of (16×16), and then random masking is applied to mask a certain ratio of the (8×16=128) patches in each image. The original last image within the input consecutive five image frames is set as the target of the reconstruction task. Using the mean squared error (MSE)

as the loss function, the image reconstruction task can be expressed as a minimisation problem by (4-1):

$$\min \quad \frac{1}{S}\sum_{k=1}^{S} d_2(M_k, P_k) \tag{4-1}$$

where S is the number of image samples; M_k is the pixel value matrix with a size of (128×256), containing all pixel values in the reconstructed image k reconstructed from the one with masked patches; P_k is the pixel value matrix with a size of (128×256), containing all pixel values in the original image k; $d_2(\cdot)$ means Euclidean norm, which calculates the Euclidean distance between the matrix M_k and P_k, and can be illustrated by (4-2):

$$d_2(M_k, P_k) = \frac{1}{h*w}\sum_{i=1}^{h}\sum_{j=1}^{w}(m_{i,j} - p_{i,j})^2 \tag{4-2}$$

where $m_{i,j}$ and $p_{i,j}$ are the pixel values on i^{th} row j^{th} column in the constructed image matrix M_k and the original image matrix P_k respectively; h is the height of the image with h = 128 in this study; w is the width of the image with w = 256 in this study.

Using UNet_ConvLSTM, SCNN_UNet_ConvLSTM, and SCNN_UNet_Attention models, the input continuous images with maskings are downsampled four times consecutively by the encoder, and the extracted time-series features of size (8×16×512) are then transferred to the ConvLSTM module (or Attention module) for spatial-temporal features integration. Finally, the decoder upsamples the integrated features four times into the same size as the input image and calculates the MSE loss between the reconstructed 5$^{\text{th}}$ image and the original 5$^{\text{th}}$ image of the input frames. Note that in the pre-training phase, the output layers of both UNet_ConvLSTM, SCNN_UNet_ConvLSTM, and SCNN_UNet_Attention, are adjusted from the original models reported in (Zou et al., 2020), (Dong et al., 2022), and (Patil et al., 2022), with the number of channels changed to 3 (check **Figure 4-2**).

Regarding the masking ratio, the results of ablation tests with ratios set at 25%, 50%, or 75% found that a 50% ratio delivers a balanced performance. Thus, in the pre-training phase, the random masking ratio is set at 50% for all models.

Different from the original masked autoencoders (He et al., 2022) implemented by the vision Transformer, the proposed upgrade version of masked sequential autoencoders for pre-training is implemented under the "CNN-ConvLSTM-CNN" or "CNN-Attention_LSTM-CNN" architecture, which can further aggregate valuable image contextual information and spatial-temporal features. By masking the whole continuous 5 image frames and only recovering the last frame, which is also the current frame for lane detection, the proposed upgraded MSAEs facilitate the model to learn not only correlations of different regions within one image but also the spatial-temporal interrelationships and dependencies between different regions of the images among continuous frames.

4.2.3 Fine-tuning with PolyLoss

Vision-based lane detection, as a typical segmentation task, aims to classify the image at the pixel level, labelling each pixel with its corresponding class, i.e., lane or background. Generally, for a segmentation task, the input is one image, but in the proposed pipeline, a continuous image sequence is used as input, and only the last image of the continuous sequence is segmented, check **Figure 4-1** for details.

By pre-training with reconstructing the masked patches, the pre-trained model should already get the aggregate contextual information and valuable spatial-temporal features, however, fine-tuning is required to further train the model to adapt it to the per-pixel segmentation task, making full use of those extracted features.

With the elongated structure, lane lines often occupy only a very small fraction of the overall pixels in an image, making lane detection a typical imbalanced two-class segmentation task. Usually, weighted cross-entropy (CE) loss is adopted for addressing this imbalanced two-class segmentation, which reshapes the standard CE loss by introducing weighting factors to reduce the weights of the background samples and focus more on the weights of lane pixels. However, literature (Jadon, 2020; Leng et al., 2022) revealed that weighted CE loss does not perform well under certain situations with severely imbalanced data. To further improve the performance of the lane detection models and improve the capabilities of handling the severe imbalance between lane line and background pixels, this study customises a PolyLoss (PL for short in the model names), and tests and verifies its effectiveness.

PloyLoss is based on the Taylor expansions of CE loss and focal loss (FL), which treats the loss functions as a linear combination of polynomial functions (Leng et al., 2022). The CE loss and FL loss can be expressed in (4-3) and (4-4):

$$L_{\mathrm{CE}} = -log(Q_t) \tag{4-3}$$

$$L_{\mathrm{FL}} = -\alpha(1-Q_t)^{\varepsilon} log(Q_t) \tag{4-4}$$

where L_{CE} and L_{FL} stand for the CE loss and FL loss, respectively; Q_t is the prediction probability of the target ground-truth class; α and ε are the tunable hyperparameters for L_{FL}.

The loss functions of both CE and FL can be decomposed into a series of weighted polynomial bases in the form of $\sum_{j=1}^{\infty}\alpha_j(1-Q_t)^j$ where$j \in \mathbb{Z}^+$, $\alpha_j \in R^+$ is the polynomial coefficient. Each polynomial basis $(1-Q_t)^j$ is weighted by the corresponding polynomial coefficients $\alpha_j \in R^+$, so that it is easy to adjust the different polynomial bases of PolyLoss. The Taylor expansion of FL, indicated by $L_{\mathrm{FL\text{-}T}}$, is given in (4-5):

$$\begin{aligned} L_{\mathrm{FL\text{-}T}} &= -(1-Q_t)^{\varepsilon} log(Q_t) = \sum_{j=1}^{\infty}\frac{(1-Q_t)^{j+\varepsilon}}{j} \\ &= (1-Q_t)^{1+\varepsilon} + \frac{(1-Q_t)^{2+\varepsilon}}{2} + \ldots + \frac{(1-Q_t)^{N+\varepsilon}}{N} + \frac{(1-Q_t)^{N+1+\varepsilon}}{N+1} + \ldots \end{aligned} \tag{4-5}$$

where $N \in \mathbb{Z}^+$; ε is a modulating factor, with which the FL can simply shift the power j by ε, i.e., shift all polynomial coefficients horizontally by ε (Leng et al., 2022).

To improve the model performance and robustness, dropping the higher-order polynomials and tuning the leading polynomials are applied in previous studies (Gonzalez & Miikkulainen, 2021; Leng et al., 2022). Similarly here, after truncating all higher order ($N+1 \to \infty$) polynomial terms and tuning the leading N polynomials using the perturbation term γ_j, $j = 1, 2, 3, \cdots, N$, the truncated $L_{\mathrm{PL\text{-}N}}$ is obtained and shown in (4-6):

$$\begin{aligned} L_{\mathrm{PL\text{-}N}} &= (\gamma_1+1)(1-Q_t)^{1+\varepsilon} + (\gamma_2+\frac{1}{2})(1-Q_t)^{2+\varepsilon} + \cdots + (\gamma_N+\frac{1}{N})(1-Q_t)^{N+\varepsilon} \\ &= -(1-Q_t)^{\varepsilon} log(Q_t) + \sum_{j=1}^{N}\gamma_j(1-Q_t)^{j+\varepsilon} \end{aligned} \tag{4-6}$$

To further simplify the $L_{\text{PL-N}}$ and render it applicable to be easily tuned for different tasks and data sets, Leng et.al (2022) carried out extensive experiments and observed that adjusting a single coefficient for the leading polynomial can achieve better performance than the original FL loss. With this, the general form of the finally simplified formula of PolyLoss L_{PL} (of FL) is illustrated by (4-7):

$$L_{\text{PL}} = -\alpha(1-Q_t)^{\varepsilon}\, log(\,Q_t) + \gamma(1-Q_t)^{\varepsilon+1} \tag{4-7}$$

where $\alpha, \gamma, \varepsilon$ are the tunable hyperparameters. Adapting it to the imbalanced two-class segmentation task of lane detection, this study further customised (4-7) into (4-9), which will be discussed in the following *subsection 4.2.5*.

More details about PolyLoss can be referred to in (Leng et al., 2022).

4.2.4 Post-processing phase

Since in real driving scenarios, it is necessary to identify the types and colours of the lane lines (e.g., dashed lines, continuous double yellow lines), the detected lane lines need to be grouped into different colours to indicate their different types, i.e., lane detection considered as an instance segmentation task. With the fine-tuning lane line segmentation outputs, the DBSCAN (Ester et al., 1996) algorithm is proposed to cluster the detected lane lines into different colours, indicating different types. Then, curve fitting is proposed at the end to smooth the detected lines, repairing the discontinuous broken ones (see the post-processing section in **Figure 4-1**). One needs to note that this study only presents here the idea of post-processing which can serve to upgrade the lane detection results, however, all the results in this study do not use post-processing which follows the general convention in literature, e.g., (Dong et al., 2022; Patil et al., 2022; J. Zhang et al., 2021; Zou et al., 2020).

4.2.5 Implementation details

Configuration details: In this study, to reduce the computational payload and save training time, the size of the images for both the training set and test set is set to a resolution of 128×256. In pre-training, the proportion of masked patches is set to 50%. Experiments were carried out on two NVIDIA Tesla V100 (32 GB memory) GPUs, using PyTorch version 1.9.0 with CUDA Deep Neural Network library (cuDNN) version 11.1. The batch size is set to be as large as possible, which is 60. The learning rate was initially set to 0.001 with decay applied after each epoch.

Network details: In network models of UNet_ConvLSTM, SCNN_UNet_ConvLSTM, and SCNN_UNet_Attention, most of the convolutional kernel size is 3×3, except for the SCNN block in SCNN_UNet_ConvLSTM and SCNN_UNet_Attention. The encoder part (see the left *Encoder* section in **Figure 4-2**) uses two convolutional layers as a downsampling block, in which the size of the feature map is reduced by half and the number of channels is doubled by the pooling layer. Four successive downsampling blocks are performed, and the last downsampling block does not change the number of output channels compared with its input. The final feature map of the encoder with a size of 8×16×512 is fed into the spatial-temporal ConvLSTM (or Attention) module.

The sequential features of the feature map are learned in the ConvLSTM/Attention module, which is equipped with 2 hidden layers of size 512 and outputs an 8×16 feature map of the same size as its input. The decoder network (check the *Decoder* part in **Figure 4-2**), is with the same size and number of feature maps as in the encoder but of a reverse-arranged symmetric structure that upsamples the extracted features to the original size of the input image. One needs to note that, in the pre-training task, to recover the image into original RGB pixels, the number of channels in the output layer of the decoder is set as 3; while in the fine-tuning segmentation phase, it is set as 2 for the two-class segmentation task. Therefore, for model weights transfer from the pre-training to the fine-tuning phase, the pre-trained model weights are transferred to the fine-tuning model except for the weights of the output layer. Both the pre-training and fine-tuning segmentation phases output images of the same size as the input one. Details can be checked in **Figure 4-2**.

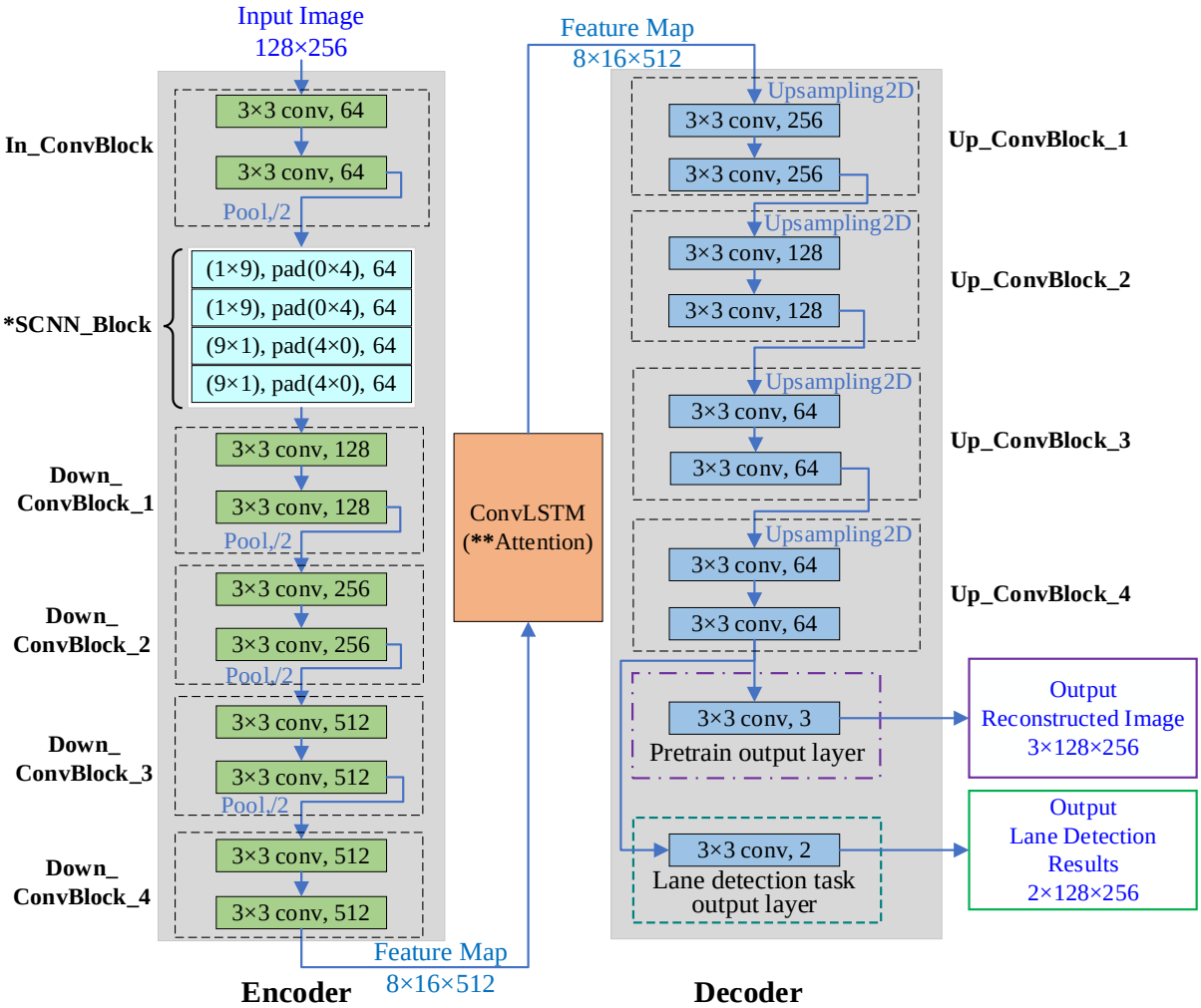


Figure 4-2. Pre-training network and lane line detection neural network structure.

**SCNN block is for SCNN_UNet_ConvLSTM and SCNN_UNet_Attention, UNet_ConvLSTM does not have it.*

***Attention block is only for SCNN_UNet_Attention model.*

Loss function details: As mentioned before, to make the proposed pipeline work, different loss functions are adopted accordingly in different phases. In the pre-training phase, the objective is

to reconstruct the masked images, and for that, the mean square error (MSE) is selected as the loss function.

While in the fine-tuning segmentation phase, the objective is to segment the pixels into lanes or backgrounds, which is a typical discriminative binary segmentation task. This study tested the weighted CE loss and the customised PolyLoss and compared their performances in tackling the imbalanced lane segmentation task. The two tailored losses applied in the fine-tuning segmentation phase are illustrated by (4-8) and (4-9).

$$\mathrm{CE} = -\frac{1}{T}\sum_{i=1}^{T}[\omega_1 y_i \, log(\, h_\theta(x_i)) + \omega_0(1-y_i)\, log(\, 1-h_\theta(x_i))] \tag{4-8}$$

$$\mathrm{PL} = -\frac{1}{T}\sum_{i=1}^{T}\begin{pmatrix}\alpha\big[y_i\big(1-h_\theta(x_i)\big)^{\varepsilon}\, log(\, h_\theta(x_i)\big) + (1-y_i)h_\theta\big(x_i\big)^{\varepsilon}\, log(\, 1-h_\theta(x_i))\big] - \\ \gamma[y_i(1-h_\theta(x_i))^{\varepsilon+1}) + (1-y_i)h_\theta(x_i)^{\varepsilon+1}]\end{pmatrix} \tag{4-9}$$

where CE and PL stand for the weighted cross-entropy loss and the customised PolyLoss, respectively; T is the number of training examples; y_i is the true segmentation label for the training example i; ω_1 and ω_0 stands for the weights for the lane class and the background class, respectively; x_i is the input training example i; $h_\theta(\cdot)$ represents the neural network model with trainable weights θ; and α, γ, ε are the tunable hyperparameters for the customised PolyLoss, which are determined by grid search method.

Optimiser details: To efficiently train and validate the proposed model pipeline, different optimisers were tested in different stages. Three optimisers, Stochastic Gradient Descent (SGD), Adaptive Moment Estimation (Adam), and Rectified Adaptive Moment Estimation (RAdam), were tested in the pre-training and fine-tuning segmentation phases. Compared to Adam, SGD requires more parameters, decreases more slowly, and may oscillate continuously on both sides of the gully. Through the tests, Adam performed better than SGD in both the pre-training task and the fine-tuning lane segmentation task. Furthermore, RAdam solves the problem of falling into local optimisation that is easily encountered by Adam, and is more robust to the changes of learning rate. Experiments verified that there was even a slight improvement in the model performance of RAdam over Adam. Therefore, the RAdam optimiser was finally chosen for both the pre-training and the fine-tuning segmentation phases.

4.3 Experiments and results

4.3.1 Datasets descriptions

To verify the proposed pipeline, a lane image dataset with continuous image frames is required. Although there are various open-sourced lane detection image datasets, e.g., CULane (Pan et al., 2018), CurveLane (Xu et al., 2020), seldom do they contain continuous frames. Therefore, this study adopted the tvtLANE (Zou et al., 2020) dataset, which is upgraded on the TuSimple lane detection challenge dataset, to train and verify the proposed method. There are one integrated training dataset and two testing sets in tvtLANE.

The tvtLANE dataset is mainly built based on the TuSimple lane detection challenge dataset. In the original TuSimple dataset, there are 3,626 training segments and 2,782 test segments with 20 continuous frames in each segment. The images are collected in different scenes at different times, and only the last frame of each segment, e.g., the 20th frame, is labelled with ground truth.

Zou et al. (2020) additionally labelled the 13th image in each segment and enlarged the dataset by adding 1,148 segments of rural driving scenes collected in China. Furthermore, data augmentation methods with cropping, flipping, and rotating operations are employed, and finally, a total number of 19,096 continuous segments are produced.

The tvtLANE consists of two test sets, i.e., testset #1 (normal) which is built on the original TuSimple test set for normal driving scenario testing, and testset #2 (challenging) which consists of 12 challenging driving scenarios for robustness evaluation. More details of tvtLANE can be found in (Dong et al., 2022; Zou et al., 2020).

In this study, 5 images are sampled from the continuous frames with a fixed stride. The sampling strides and frames used in the training and testing sets are elaborated in **Table 4-1**, and image samples are demonstrated in **Figure 4-3**.

Figure 4-3. Image samples in the tvtLANE training set and the test set

The first five images in each column are the inputs of consecutive frames, and the sixth one is the labelled ground truth of the last image in the consecutive frames. The first column is one sample in the training set, the second column is for the testset #1 (normal), and the third column is for testset #2 (challenging).

Table 4-1. Sample methods for the trainset and testset

Subset	Labeled Ground Truth	Sample Stride	Sample Frames
Trainset	13^{th}	3	1^{st},4^{th},7^{th},10^{th},13^{th}
		2	5^{th},7^{th},9^{th},11^{th},13^{th}
		1	9^{th},10^{th},11^{th},12^{th},13^{th}
	20^{th}	3	8^{th},11^{th},14^{th},17^{th},20^{th}
		2	12^{th},14^{th},16^{th},18^{th},20^{th}
		1	16^{th},17^{th},18^{th},19^{th},20^{th}
Testset #1 *Normal*	13^{th} 20^{th}	1	9^{th},10^{th},11^{th},12^{th},13^{th}
		1	16^{th},17^{th},18^{th},19^{th},20^{th}
Testset #2 *Challenging*	All	1	1^{st},2^{nd},3^{rd},4^{th},5^{th}
			2^{nd},3^{rd},4^{th},5^{th},6^{th}
			3^{rd},4^{th},5^{th},6^{th},7^{th}
			…

4.3.2 Evaluation metrics

Overall, the model performance is evaluated in terms of both visual qualitative examination with results demonstration and quantitative analysis with metrics. Considering the vision-based lane detection task as a pixel-level classification task, commonly used metrics, i.e., accuracy, precision, recall, and F1-measure (Dong et al., 2022; Patil et al., 2022; J. Zhang et al., 2021; Zou et al., 2020), are adopted. The calculations of these metrics are illustrated by (4-10)-(4-13).

$$\text{Accuracy} = \frac{Truly\ Classified\ Pixels}{Total\ Number\ of\ Pixels} \tag{4-10}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{4-11}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{4-12}$$

$$\text{F1-measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4-13}$$

In the above equations, true positive indicates the number of image pixels that are lane lines and are correctly identified; False positive indicates the number of image pixels that are background but incorrectly classified as lane lines; False negative indicates the number of image pixels that are lane lines but incorrectly classified as background.

Furthermore, for estimating the models' computational complexities, the model parameter size, i.e., Params (M), and the multiply-accumulate (MAC) operations, i.e., MACs (G), are provided.

4.3.3 Results

In this sub-section, reconstruction performance in the self pre-training phase will be visually demonstrated, while the lane detection testing results of various models on both tvtLANE testset#1 (normal) and tvtLANE testset#2 (challenging) will be evaluated qualitatively and quantitatively.

Self pre-training results: **Figure 4-4** shows the reconstructing results of the masked images in the pre-training phase. It can be seen that the masked patches in the images can be restored very

well. Although there are some minor blurs in certain images, the reconstructed images generally recover the main and critical patterns.

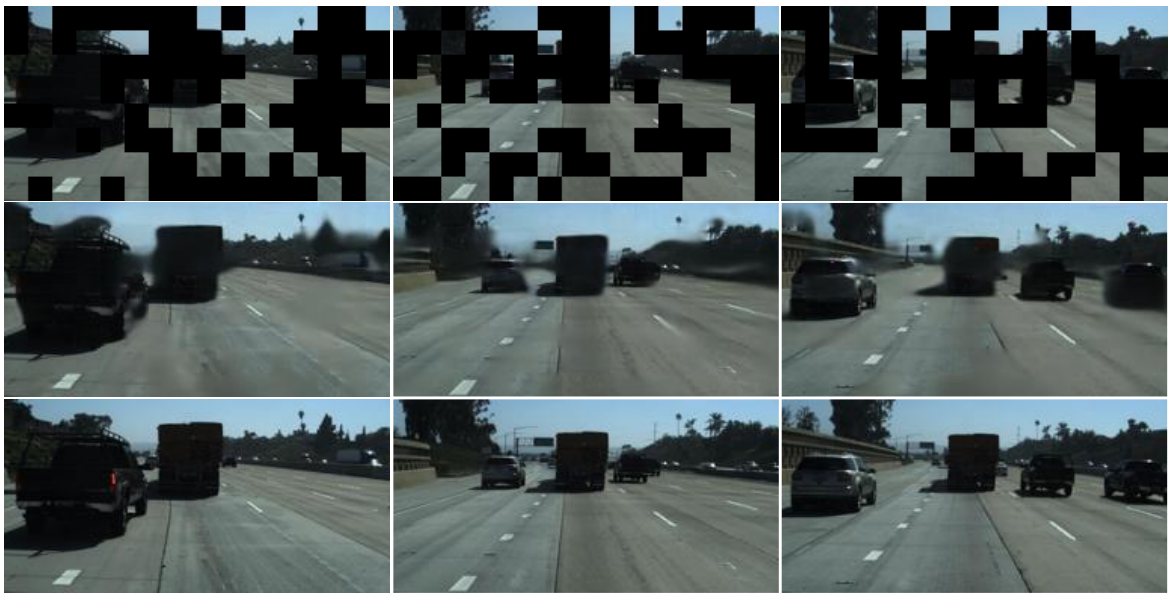


Figure 4-4. Visualisation of the reconstructing results in the pre-training phase

The first row shows images with 50% of the patches masked. The second row shows the reconstructed images after pre-training. The third row shows the original images.

Testing results on tvtLANE testset #1 (normal): **Figure 4-5**, **Figure 4-7 (A)**, and **Table 4-2 (a)** demonstrate the qualitative and quantitative testing results on tvtLANE testset #1 (normal).

Qualitatively, for the lane detection segmentation task, the model should be able to accurately predict the total number of lane lines, correctly detecting the location of the lane lines while avoiding unexpected broken lines and blurs. Visualisations of the lane detection results show that models using the proposed self-supervised pre-training method generally perform better than those without. Furthermore, models using the customised PolyLoss generally outperform those using weighted CE loss with thinner detected lane lines and fewer blurs. Aligning with previous studies (Dong et al., 2022; Patil et al., 2022; J. Zhang et al., 2021; Zou et al., 2020), models using multi-continuous image frames defeat those using one single image as indicated in rows (c) and (d) there are fatter lane lines, merged lanes, and blurred areas at the top boundary of the image, and even wrongly detected lane numbers (check the first column in **Figure 4-7 (A)**). One can also notice that even when vehicles or shadings of the vehicles are blocking the lane lines, the models with the proposed pretraining method and using the proposed PolyLoss can identify the lane lines completely and continuously with correct locations (check the first, fourth, and sixth columns in **Figure 4-7 (A)**), which is crucial for vehicle localisation.

Quantitatively, **Table 4-2 (a)** demonstrates that the proposed self-supervised pre-training method improves the lane detection results for both UNet_ConvLSTM and SCNN_UNet_ConvLSTM models, and the models using the customised PolyLoss all outperform those using the weighted CE loss regarding accuracy, precision, and F1-measure. To be specific, with the self-supervised pre-training pipeline and using the customised PolyLoss, UNet_ConvLSTM_PL** advances a lot from the baseline UNet_ConvLSTM with testing accuracy improved from 98.00% to 98.34%, precision improved from 0.857 to 0.921, and F1-measure improved from 0.904 to 0.915; while SCNN_UNet_ConvLSTM_PL** also improves

a lot from the baseline SCNN_UNet_ConvLSTM with testing accuracy improved from 98.19% to 98.38%, precision improved from 0.889 to 0.929, and F1-measure improved from 0.918 to 0.922. All the models' parameter sizes and MACs do not increase.

Figure 4-5. Lane detection results obtained by SCNN_UNet_Attention_PL on tvtLANE testset #1 (normal) without post-processing**

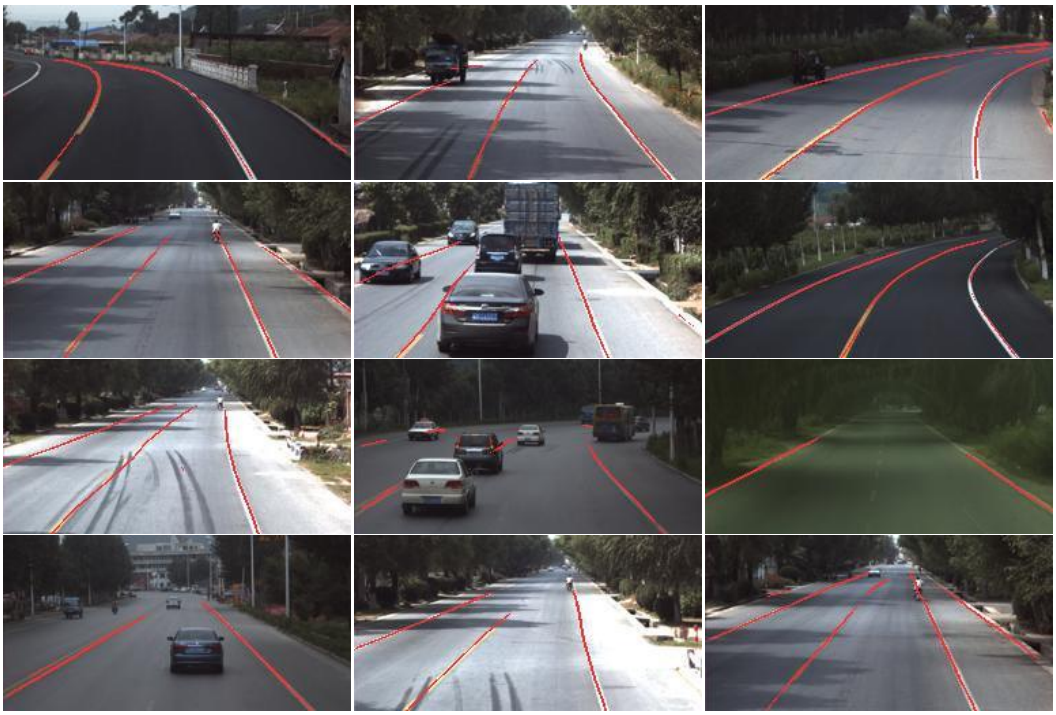


Figure 4-6. Lane detection results obtained by SCNN_UNet_ConvLSTM_PL on tvtLANE testset #2 (challenging) without post-processing**

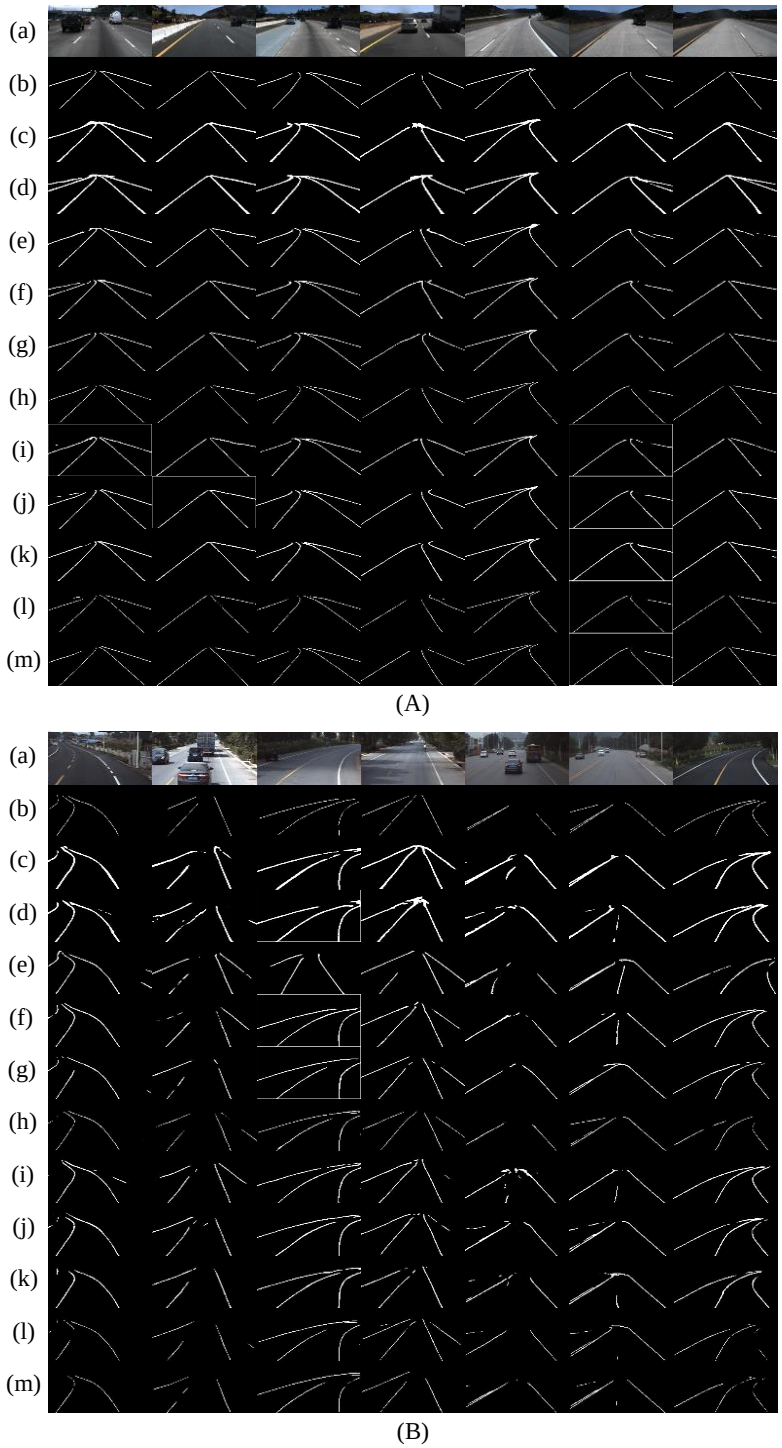


Figure 4-7. Qualitative visual comparison of the lane detection results testing on (A) tvtLANE testset #1 (normal) and (B) tvtLANE testset #2 (challenging)

*All results in the figure are without post-processing. (a) Original input images; (b) Ground truth; (c)~(l) are the lane detection results corresponding to the models: (c) SegNet, (d) UNet, (e) SegNet_ConvLSTM (Zou et al., 2020), (f) UNet_ConvLSTM, (g) UNet_ConvLSTM_CE**, (h) UNet_ConvLSTM_PL**, (i) SCNN_SegNet_ConvLSTM (Dong et al., 2022), (j) SCNN_UNet_ConvLSTM, (k) SCNN_UNet_ConvLSTM_CE**, (l) SCNN_UNet_ConvLSTM_PL**, (m) SCNN_UNet_Attention_PL**. (Note: CE and PL are short for weighted cross-entropy loss and PolyLoss, respectively, while ** means the model is pre-trained with the proposed self-supervised pre-training method.)*

Table 4-2. Model performance comparison

(a) tvtLANE testset #1 (normal)

Model		Test_Acc (%)	Precision	Recall	F1-measure	MACs (G)	Params (M)
Using single image	SegNet	96.93	0.796	0.962	0.871	50.2	29.4
	UNet	96.54	0.790	**0.985**	0.877	15.5	13.4
	SCNN*	96.79	0.654	0.808	0.722	77.7	19.2
	LaneNet*	97.94	0.875	0.927	0.901	44.5	19.7
Using multi-continuous images	SegNet_ConvLSTM	97.92	0.874	0.931	0.901	217.0	67.2
	UNet_ConvLSTM	98.00	0.857	0.958	0.904	69.0	51.1
	UNet_ConvLSTM_CE**	98.19	0.882	0.940	0.910	69.0	51.1
	UNet_ConvLSTM_PL**	98.34	0.921	0.909	0.915	69.0	51.1
	SCNN_SegNet_ConvLSTM	98.07	0.893	0.928	0.910	223.0	67.3
	SCNN_UNet_ConvLSTM	98.19	0.889	0.950	0.918	93.0	51.3
	SCNN_UNet_ConvLSTM_CE**	98.20	0.891	0.952	0.921	93.0	51.3
	SCNN_UNet_ConvLSTM_PL**	**98.38**	0.929	0.915	0.922	93.0	51.3
	SCNN_UNet_Attention_PL**	98.36	**0.937**	0.911	**0.924**	68.9	13.7

(b) tvtLANE testset #2 (challenging)

Model \ Challenging Scenes	overall	1-curve & occlude	2-shadow-bright	3-bright	4-occlude	5-curve	6-dirty & occlude	7-urban	8-blur & curve	9-blur	10-shadow & dark	11-tunnel	12-dim & occlude
Precision													
SegNet	0.6080	0.6810	0.7067	0.5987	0.5132	0.7738	0.2431	0.3195	0.6642	0.7091	0.7499	0.6225	0.6463
UNet	0.6754	0.7018	0.7441	0.6717	0.6517	0.7443	0.3994	0.4422	0.7612	0.8523	0.7881	0.7009	0.5968
SegNet_ConvLSTM	0.7563	0.8176	0.8020	0.7200	0.6688	0.8645	0.5724	0.4861	0.7988	0.8378	0.8832	0.7733	0.8052
UNet_ConvLSTM	0.7784	0.7591	0.8292	0.7971	0.6509	0.8845	0.4513	0.5148	0.8290	**0.9484**	0.9358	0.7926	0.8402
UNet_ConvLSTM_CE**	0.7932	0.8004	0.8312	0.8285	0.7661	0.8557	0.5242	0.5567	0.7545	0.9200	0.9312	0.8496	0.8026
UNet_ConvLSTM_PL**	0.8331	0.8429	**0.8824**	0.8691	0.8125	**0.9578**	0.5970	**0.5591**	0.8289	0.9247	**0.9634**	0.7688	0.9160
SCNN_SegNet_ConvLSTM	0.7673	0.8326	0.7497	0.7470	0.7369	0.8647	0.6196	0.4333	0.7371	0.8566	0.9125	0.8153	0.8466
SCNN_UNet_ConvLSTM	0.7784	0.8182	0.8362	0.8189	0.7359	0.8365	0.5872	0.5377	0.8046	0.8770	0.8722	0.7952	0.7817
SCNN_UNet_ConvLSTM_CE**	0.8001	0.8754	0.8672	0.8519	0.7763	0.8664	0.5523	0.5261	0.7396	0.8865	0.8974	0.8115	0.9101
SCNN_UNet_ConvLSTM_PL**	**0.8444**	0.9074	0.8757	0.8644	0.8464	0.9049	**0.7177**	0.4827	0.8157	0.9440	0.9606	**0.8736**	0.9220
SCNN_UNet_Attention_PL**	0.8413	**0.9189**	0.8763	**0.8838**	**0.8598**	0.9238	0.6210	0.5229	**0.8847**	0.9039	0.9229	0.8408	**0.9369**
F1-measure													
SegNet	0.6727	0.8042	0.7900	0.7023	0.6127	0.8639	0.2110	0.4267	0.7396	0.7286	0.7675	0.6935	0.5822
UNet	0.6985	0.8200	0.8408	0.7946	0.7337	0.7827	0.3698	0.5658	0.8147	0.7715	0.6619	0.5740	0.4646
SegNet_ConvLSTM	0.7609	0.8852	0.8544	0.7688	0.6878	0.9069	0.4128	0.5317	0.7873	0.7575	0.8503	0.7865	0.7947
UNet_ConvLSTM	0.7143	0.8465	0.**8891**	0.8411	0.7245	0.8662	0.2417	0.5682	0.8323	**0.7852**	0.6404	0.4741	0.5718
UNet_ConvLSTM_CE**	0.6537	0.8365	0.8697	0.8263	0.7614	0.8165	0.2440	0.5359	0.7618	0.7206	0.4832	0.3274	0.2595
UNet_ConvLSTM_PL**	0.6284	0.8220	0.8731	0.8300	0.7705	0.8295	0.1845	0.4426	0.7278	0.5712	0.4157	0.3545	0.4821
SCNN_SegNet_ConvLSTM	**0.7666**	**0.8956**	0.8237	0.7909	0.7468	**0.9108**	**0.4398**	0.4858	0.7379	0.7546	**0.8729**	**0.7963**	**0.8074**
SCNN_UNet_ConvLSTM	0.7024	0.8670	0.8866	0.8405	0.7565	0.7955	0.4179	**0.5933**	0.7880	0.7285	0.6296	0.4747	0.4134
SCNN_UNet_ConvLSTM_CE**	0.7327	0.8937	0.8690	**0.8426**	0.7656	0.8352	0.2493	0.5751	0.7756	0.7122	0.7661	0.6989	0.5420
SCNN_UNet_ConvLSTM_PL**	0.6711	0.8685	0.8796	0.8161	**0.7988**	0.7897	0.2853	0.4921	0.8258	0.7255	0.5244	0.3963	0.3255
SCNN_UNet_Attention_PL**	0.6772	0.8530	0.8771	0.8111	0.7579	0.7881	0.2926	0.5057	**0.8595**	0.7569	0.5857	0.3737	0.4565
Accuracy (%)													
SegNet	96.57	96.72	96.16	96.01	96.83	96.50	95.93	96.16	96.39	96.12	97.26	96.79	97.37
UNet	96.68	96.68	96.00	95.78	97.06	96.35	95.45	96.35	96.58	96.62	97.50	97.53	97.58
SegNet_ConvLSTM	97.83	98.10	97.38	97.52	98.17	97.72	96.98	97.92	97.61	97.08	98.39	98.07	98.26
UNet_ConvLSTM	97.93	97.83	97.48	97.70	97.94	97.73	97.27	97.86	97.75	97.65	98.49	98.37	98.38
UNet_ConvLSTM_CE**	98.13	98.19	97.72	98.04	98.47	97.77	97.41	98.30	97.67	97.69	98.58	98.54	98.57
UNet_ConvLSTM_PL**	**98.38**	98.60	**98.06**	**98.33**	98.75	**98.35**	97.66	**98.61**	98.09	97.77	98.63	**98.63**	98.63
SCNN_SegNet_ConvLSTM	97.90	98.24	97.21	97.68	98.39	97.73	97.11	97.80	97.48	97.29	98.50	98.28	98.34
SCNN_UNet_ConvLSTM	97.95	98.08	97.45	97.86	98.31	97.63	97.17	97.95	97.63	97.43	98.41	98.39	98.39
SCNN_UNet_ConvLSTM_CE**	98.03	98.33	97.64	98.05	98.45	97.69	97.42	97.95	97.54	97.57	98.38	98.23	98.56
SCNN_UNet_ConvLSTM_PL**	98.36	98.75	97.98	98.31	**98.78**	98.06	**97.69**	98.36	98.12	**97.92**	**98.65**	98.55	98.63
SCNN_UNet_Attention_PL**	98.35	**98.77**	97.98	98.30	98.70	98.17	97.57	98.56	**98.19**	97.74	98.61	98.51	**98.64**

** Results reported in (J. Zhang et al., 2021).*

*** Results of the models with the proposed self-supervised pre-training.*

"CE" is short for weighted cross-entropy loss, and "PL" is short for PolyLoss.

*Therefore, "UNet_ConvLSTM_CE**" means UNet_ConvLSTM model with self-supervised pre-training and using weighted cross-entropy loss in the fine-tuning phase. This naming rule applies to all other models.*

One can find that for both models, the most significant improvement was identified in precision (i.e., 0.857 to 0.921 and 0.889 to 0.929). The higher the precision the lower the false positive is (check (4-11)) which means the models become more strict on pixel samples to be classified as the lane line contributing to fewer wrong detected lane pixels, which is also illustrated by the thinner detected lane lines in **Figure 4-7 (A)**. However, this might increase the number of lane pixels that are incorrectly identified as background, i.e., higher false negatives, thus the recall ratio decreases. Therefore, the F1-measure, which balances precision and recall ratio, is a more reasonable evaluation measure to serve as the main benchmark (Dong et al., 2022; Pan et al., 2018; Patil et al., 2022; J. Zhang et al., 2021; Zou et al., 2020). Furthermore, SCNN_UNet_Attention, which was tested only under the best setting of using pre-training and customised PolyLoss, obtained the best precision (0.937) and F1-measure (0.924), beating all other state-of-the-art baseline models on this tvtLANE testset #1 (normal scene testing).

Testing results on tvtLANE testset #2 (challenging): **Figure 4-6**, **Figure 4-7 (B)**, and **Table 4-2 (b)** demonstrate the qualitative and quantitative testing results on tvtLANE testset #2 (challenging).

Qualitatively, as illustrated in **Figure 4-6** and **Figure 4-7 (B)**, when testing on the challenging driving scenes, all the models do not perform well. However, the results obtained by the models using the proposed self-supervised pre-training method are still better than those without pre-training. Especially models adopting the customised PolyLoss still output thinner lanes with less blur and more correct lane numbers.

Quantitatively, as shown in **Table 4-2 (b)**, models with pre-training generally outperform those without, regarding overall accuracy and precision. Typically, using the self-supervised pre-training method plus the customised PolyLoss, the developed UNet_ConvLSTM_PL** model obtains the best overall accuracy (98.38%), and together with other proposed models (with ** in their names), they take all the best accuracies in all 12 challenging scenes; SCNN_UNet_ConvLSTM_PL** obtains the best overall precision (0.8444) followed by SCNN_UNet_Attention_PL** (0.8413), and also together with other proposed models, they fill 11 best precisions out of all the 12 challenging scenes except for only scene 9 *blur*.

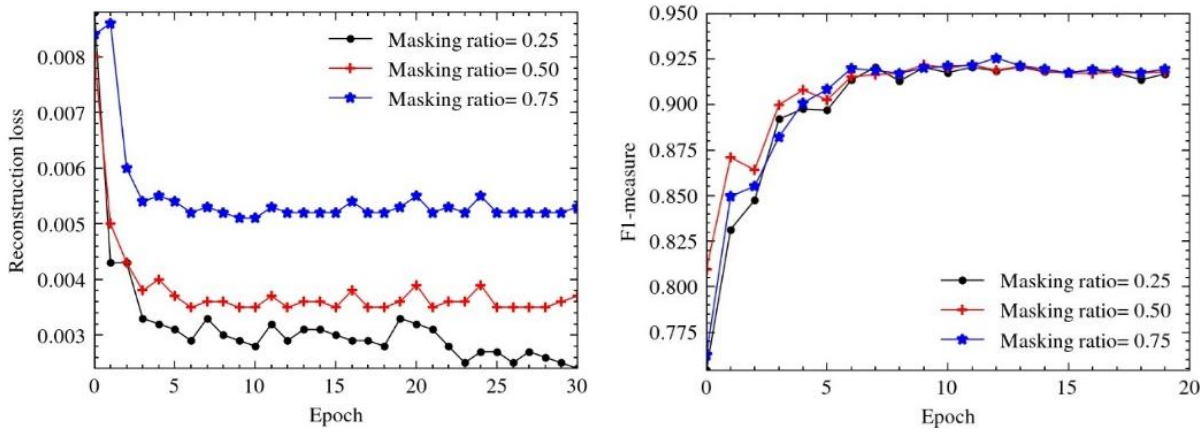
It is worth noting that the models using the proposed self-supervised pre-training deliver slightly worse F1-measures compared to those without pre-training. This is because the models are more strict with the pixels classified as the lane lines which might increase the number of lane pixels that are incorrectly identified as background, i.e., resulting in higher false negatives, thus the recall ratio decreases and the F1-measures get slightly worse (even if there are increases in precisions). From **Figure 4-7 (B)**, it is more intuitive to see that the developed models with the proposed pre-training and PolyLoss still show acceptable results, better than the baselines.

4.3.4 Ablation study and discussion

Masking ratio: Experimental results in the previous study (Xie et al., 2022) showed that the masking ratio needs to correspond to the mask patch, i.e., "for a small mask patch size of 8, the masking ratio needs to be as high as 80% to perform well", while "for a large masking patch size of 32, the approach can achieve competitive performance in a wide range of masking ratios (10%-70%)". In this study, the patch size is set as (16×16), and the experimental comparisons were carried out with ratios set as 25%, 50%, and 75%.

Testing on SCNN_UNet_ConvLSTM model, **Figure 4-8 (a)** shows the average normalised reconstruction loss indicated by the MSE of the image reconstruction task during the pre-training phase. It is observed that using a smaller masking ratio leads to lower reconstruction loss, which is easy to understand, as a smaller masking ratio means fewer pixels need to be reconstructed.

Figure 4-8 (b) shows the lane detection performance on the normal driving scene dataset regarding F1-measure with different masking ratios, and **Table 4-3** shows the detailed quantitative results.



(a) (b)

Figure 4-8. Model performance comparison with different masking ratio settings: (a) reconstruction loss in the pre-training phase, and (b) the F1-measure testing on tvtLANE testset #1

It is found that although the result of masking at a 75% ratio achieves the best F1-measure of 0.926 on the normal dataset, it does not perform particularly well on the challenge dataset, where it only achieves an F1-measure of 0.7162 worse than that of masking at a 50% ratio (F1-measure at 0.7327).

Furthermore, referring to the results of the pre-training phase, it is clear that masking at a 50% ratio delivers balanced results during both the pre-training phase and fine-tuning testing phases. It is more reasonable to adopt the balanced setting to verify the proposed lane detection pipeline and method, and thus, 50% was chosen as the masking ratio for all testing models.

Loss function: Earlier mentioned in this study, two loss functions (i.e., weighted CE loss and PolyLoss) were tested in the experiments under the proposed pipeline in the fine-tuning segmentation phase. The quantitative comparison results are shown in **Table 4-2**, and the qualitative results are intuitively demonstrated with visualisations in **Figure 4-7**.

As shown in **Table 4-2 (a)**, testing on tvtLANE testset #1 (normal scene), for both SCNN_UNet_ConvLSTM and UNet_ConvLSTM-based models, the overall performance of using PolyLoss outperforms that of weighted CE loss. To be specific, compared with UNet_ConvLSTM_CE**, the UNet_ConvLSTM_PL** model obtains an increase of 0.15% in accuracy; a significant increase of 0.039, i.e., around 4.4% improvement, in precision; while a bit decrease in recall ratio; and, overall, a better F1-measure of 0.915 over 0.910.

SCNN_UNet_ConvLSTM_PL** gets the same superiority patterns over SCNN_UNet_ConvLSTM_CE**, and SCNN_UNet_ConvLSTM_PL** obtains the second-best F1-measure (0.922), the second-best precision (0.929), and the best accuracy (98.38%), among all tested models. SCNN_UNet_Attention_PL** slightly beats SCNN_UNet_ConvLSTM_PL** in F1-measure (0.924) and precision (0.937). The superiority of the customised PolyLoss over weighted CE loss can be explained by that the PolyLoss function is designed as a linear combination of polynomial functions, so that the importance of polynomial bases can be adjusted according to the imbalanced dataset and regarding the segmentation task. With the fine-tuned hyperparameters $\alpha, \gamma, \varepsilon$ in (4-9), the customised PolyLoss is perfectly adjusted to the dedicated lane detection task.

The model using PolyLoss also performs better than the ones using weighted CE loss in almost all challenging scenes regarding accuracy and precision. In particular, testing on the challenging driving scenes dataset, UNet_ConvLSTM_PL** gets the highest overall accuracy at 98.38%, while SCNN_UNet_ConvLSTM_PL** obtains the best overall precision at 0.8444.

Table 4-3. Model performance with different masking ratios

(a) tvtLANE testset #1 (normal)

Mask Ratio	Test_Acc (%)	Precision	Recall	F1-measure
25%	98.36	0.927	0.915	0.921
50%	98.20	0.891	**0.952**	0.921
75%	**98.40**	**0.933**	0.918	**0.926**

(b) tvtLANE testset #2 (challenging)

Challenging Scenes	Precision			F1-measure			Accuracy (%)		
Mask ratio	25%	50%	75%	25%	50%	75%	25%	50%	75%
overall	0.8248	0.8001	0.8348	0.7196	0.7327	0.7162	98.31	98.03	98.36
1-crve&occlude	0.8083	0.8754	0.9433	0.8238	0.8937	0.9260	98.58	98.33	98.83
2-shadow-bright	0.8881	0.8672	0.9028	0.7953	0.869	0.8777	98.01	97.64	98.09
3-bright	0.8611	0.8519	0.8786	0.7944	0.8426	0.8111	98.30	98.05	98.34
4-occlude	0.8480	0.7763	0.8615	0.7703	0.7656	0.7438	98.80	98.45	98.83
5-curve	0.9327	0.8664	0.9187	0.7660	0.8352	0.8840	98.14	97.69	98.14
6-dirty&occlude	0.7052	0.5523	0.4813	0.3595	0.2493	0.2655	97.55	97.42	97.44
7-urban	0.5090	0.5261	0.5565	0.4939	0.5751	0.5150	98.46	97.95	98.52
8-blur&curve	0.7915	0.7396	0.7823	0.7933	0.7756	0.7426	98.01	97.54	98.08
9-blur	0.9473	0.8865	0.9462	0.7396	0.7122	0.7437	97.74	97.57	97.76
10-shadow&dark	0.9553	0.8974	0.9331	0.7942	0.7661	0.7180	98.71	98.38	98.65
11-tunnel	0.8427	0.8115	0.8956	0.7217	0.6989	0.6667	98.44	98.23	98.53
12-dim&occlude	0.7588	0.9101	0.8750	0.6173	0.542	0.5973	98.33	98.56	98.54

*All of the test results in **Table 4-3** were tested on the SCNN_UNet_ConvLSTM model.*

Training time and model complexity: In addition to the improvement regarding the evaluation metrics, the proposed self-supervised pre-training pipeline plus the customised PolyLoss can also reduce the training time, with the model convergence speed greatly improved. To be specific, tests revealed that for UNet_ConvLSTM-based models, UNet_ConvLSTM_PL** converged at the 10th epoch, while UNet_ConvLSTM_CE** converged at the 91st epoch, and

UNet_ConvLSTM without the proposed pertaining needed around 100 epochs to converge (Zou et al., 2020). Similarly, for SCNN_UNet_ConvLSTM-based models, SCNN_UNet_ConvLSTM_PL** converged at the 12th epoch, while SCNN_UNet_ConvLSTM_CE** converged at the 29th epoch, and SCNN_UNet_ConvLSTM without the proposed pre-training needed around 100 epochs to converge.

These results demonstrate that pre-training with MSAEs plus fine-tuning with PolyLoss can not only boost the models' overall performance regarding accuracy, precision, and F1-measure, but also speed up model convergence, greatly reducing the training time.

Furthermore, from the parameters and MACs illustrated in **Table 4-2 (a)**, it is demonstrated that, with the proposed pre-training and customised PolyLoss, the model size and complexity merely change.

In short, the proposed pipeline contributes to the improvement of model efficiency and detection accuracy simultaneously.

4.4 Conclusion

In this study, a novel deep learning pipeline integrating self-supervised pre-training with masked sequential autoencoders, fine-tuning segmentation with customised PolyLoss, and post-processing with clustering and curve-fitting, is proposed for the vision-based robust lane detection task. With the proposed self-supervised pre-training method by reconstructing the randomly masked image frames and the customised PolyLoss for the fine-tuning segmentation phase, the tested three neural network models (i.e., UNet_ConvLSTM, SCNN_UNet_ConvLSTM, and SCNN_UNet_Attention) all delivered significantly better performances in comparison to baselines. Through extensive experiments, the models under the proposed pipeline surpass other state-of-the-art models with the best testing accuracy, precision, and F1-measure on the normal driving dataset (i.e., tvtLANE testset #1) and the best overall accuracy and precision on the 12 challenging driving scenarios (i.e., tvtLANE testset #2). Furthermore, without changes in the model size and complexity, under the proposed pipeline, the test models converged faster, especially when adopting the customised PolyLoss in the fine-tuning segmentation phase, while performing better detection results. These findings demonstrate the effectiveness of the proposed lane detection pipeline which upgrades the model training efficiency and detection accuracy simultaneously.

It is witnessed that when testing with some brand new challenging samples, i.e., no similar samples are covered in the training phase, the model might be defeated with a low F1-measure. In practice, lane detection models trained on datasets from one certain country might not work well when testing on datasets with different lane structures from another country. To tackle this problem and further enhance the model's robustness, for future studies, it is suggested to investigate domain generalisation and adaptation methods to transfer the knowledge and patterns learned from available datasets to unseen domains and fields with brand new data.

Acknowledgements

This work was supported by the Applied and Technical Sciences (TTW), a subdomain of the Dutch Institute for Scientific Research (NWO) through the Project Safe and Efficient Operation

of Automated and Human-Driven Vehicles in Mixed Traffic (SAMEN) under Contract 17187. The authors would like to thank Dr. Haneen Farah for her valuable comments and suggestions on this work.

References

Al Mamun, A., Em, P. P., & Hossen, J. (2021). Lane marking detection using simple encode decode deep learning technique: SegNet. International Journal of Electrical and Computer Engineering, 11(4), 3032.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2016.2644615

Bao, H., Dong, L., Piao, S., & Wei, F. (2022). BEIT: Bert pre-training of image transformers. ICLR 2022 - 10th International Conference on Learning Representations.

Borkar, A., Hayes, M., & Smith, M. T. (2011). Polar randomized hough transform for lane detection using loose constraints of parallel lines. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 1037–1040. https://doi.org/10.1109/ICASSP.2011.5946584

Cao, J., Song, C., Song, S., Xiao, F., & Peng, S. (2019). Lane detection algorithm for intelligent vehicles in complex road conditions and dynamic environments. Sensors, 19(14), 3166.

Dong, Y., Patil, S., van Arem, B., & Farah, H. (2022). A hybrid spatial-temporal deep learning architecture for lane detection. Computer-Aided Civil and Infrastructure Engineering, 1–18. https://doi.org/10.1111/mice.12829

El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., & Grave, E. (2021). Are large-scale datasets necessary for self-supervised pre-training? ArXiv Preprint ArXiv:2112.10740.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd, 96(34), 226–231.

Feng, Z., Guo, Y., Liang, Q., Bhutta, M. U. M., Wang, H., Liu, M., & Sun, Y. (2022). MAFNet: Segmentation of road potholes with multimodal attention fusion network for autonomous vehicles. IEEE Transactions on Instrumentation and Measurement, 71. https://doi.org/10.1109/TIM.2022.3200100

Gad, G. M., Annaby, A. M., Negied, N. K., & Darweesh, M. S. (2020). Real-time lane instance segmentation using SegNet and image processing. 2nd Novel Intelligent and Leading Emerging Sciences Conference, NILES 2020. https://doi.org/10.1109/NILES50944.2020.9257977

Ghafoorian, M., Nugteren, C., Baka, N., Booij, O., & Hofmann, M. (2018). El-gan: Embedding loss driven generative adversarial networks for lane detection. Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 0.

Gonzalez, S., & Miikkulainen, R. (2021). Optimizing loss functions through multi-variate Taylor polynomial parameterization. GECCO 2021 - Proceedings of the 2021 Genetic and Evolutionary Computation Conference, 305–313. https://doi.org/10.1145/3449639.3459277

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 15979–15988. https://doi.org/10.1109/cvpr52688.2022.01553

Huang, S., Shen, Z., Huang, Z., Ding, Z., Dai, J., Han, J., Wang, N., & Liu, S. (2023). Anchor3DLane: learning to regress 3D anchors for monocular 3D lane detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17451-17460).

Jadon, S. (2020). A survey of loss functions for semantic segmentation. 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2020. https://doi.org/10.1109/CIBCB48159.2020.9277638

Jin, D., Park, W., Jeong, S.-G., Kwon, H., & Kim, C.-S. (2022). Eigenlanes: Data-driven lane descriptors for structurally diverse lanes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17163-17171).

Lee, D. H., & Liu, J. L. (2023). End-to-end deep learning of lane detection and path prediction for real-time autonomous driving. Signal, Image and Video Processing. https://doi.org/10.1007/s11760-022-02222-2

Leng, Z., Tan, M., Liu, C., Cubuk, E. D., Shi, X., Cheng, S., & Anguelov, D. (2022). PolyLoss: A polynomial expansion perspective of classification loss functions. In 10th International Conference on Learning Representations, ICLR 2022.

Li, J., Jiang, F., Yang, J., Kong, B., Gogate, M., Dashtipour, K., & Hussain, A. (2021). Lane-DeepLab: Lane semantic segmentation in automatic driving scenarios for high-definition maps. Neurocomputing, 465, 15–25.

Liu, R., Yuan, Z., Liu, T., & Xiong, Z. (2021). End-to-end lane shape prediction with transformers. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 3694–3702.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440.

Pan, X., Shi, J., Luo, P., Wang, X., & Tang, X. (2018). Spatial as deep: Spatial CNN for traffic scene understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

Patil, S., Dong, Y., Farah, H., & Hellendoorn, H. (2022). Sequential neural network model with spatial-temporal attention mechanism for robust lane detection using multi continuous image frames. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4273506

Qin, Z., Zhang, P., & Li, X. (2022). Ultra fast deep lane detection with hybrid anchor driven ordinal classification. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Ren, X., Li, M., Li, Z., Wu, W., Bai, L., & Zhang, W. (2022). Curiosity-driven attention for anomaly road obstacles segmentation in autonomous driving. IEEE Transactions on Intelligent Vehicles, 1–11. https://doi.org/10.1109/TIV.2022.3204714

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 234–241.

Somawirata, I. K., & Utaminingrum, F. (2017). Road detection based on the color space and cluster connecting. 2016 IEEE International Conference on Signal and Image Processing, ICSIP 2016, 118–122. https://doi.org/10.1109/SIPROCESS.2016.7888235

Tabelini, L., Berriel, R., Paixão, T. M., Badue, C., de Souza, A. F., & Oliveira-Santos, T. (2021). Keep your eyes on the lane: Real-time attention-guided lane detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 294–302. https://doi.org/10.1109/CVPR46437.2021.00036

Torres, L. T., Berriel, R. F., Paixão, T. M., Badue, C., De Souza, A. F., & Oliveira-Santos, T. (2020). PolyLaneNet: Lane estimation via deep polynomial regression. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 6150-6156). IEEE.

Wang, H., Chen, Y., Cai, Y., Chen, L., Li, Y., Sotelo, M. A., & Li, Z. (2022). SFNet-N: An improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes. IEEE Transactions on Intelligent Transportation Systems, 23(11), 21405–21417. https://doi.org/10.1109/TITS.2022.3177615

Wang, Y., Teoh, E. K., & Shen, D. (2004). Lane detection and tracking using B-Snake. Image and Vision Computing, 22(4), 269–280. https://doi.org/10.1016/j.imavis.2003.10.003

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., & Hu, H. (2022). Simmim: A simple framework for masked image modeling. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9653–9663.

Xu, H., Wang, S., Cai, X., Zhang, W., Liang, X., & Li, Z. (2020). CurveLane-NAS: Unifying lane-sensitive architecture search and adaptive point blending. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-030-58555-6_41

Zang, J., Zhou, W., Zhang, G., & Duan, Z. (2018). Traffic lane detection using fully convolutional neural network. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 305–311.

Zhang, J., Deng, T., Yan, F., & Liu, W. (2021). Lane detection model based on spatio-temporal network with double convolutional gated recurrent units. IEEE Transactions on Intelligent Transportation Systems, 23(7), 6666-6678. https://doi.org/10.1109/TITS.2021.3060258

Zhang, Y., Lu, Z., Zhang, X., Xue, J. H., & Liao, Q. (2022). Deep learning in lane marking detection: A survey. IEEE Transactions on Intelligent Transportation Systems, 23(7), 5976-5992. https://doi.org/10.1109/TITS.2021.3070111

Zheng, F., Luo, S., Song, K., Yan, C.-W., & Wang, M.-C. (2018). Improved lane line detection algorithm based on Hough transform. Pattern Recognition and Image Analysis, 28(2), 254–260.

Zou, Q., Jiang, H., Dai, Q., Yue, Y., Chen, L., & Wang, Q. (2020). Robust lane detection from continuous driving scenes using deep neural networks. IEEE Transactions on Vehicular Technology, 69(1), 41–54. https://doi.org/10.1109/TVT.2019.2949603.

5 Intelligent anomaly detection for lane rendering using Transformer with self-supervised pre-training and customised fine-tuning

Abstract

The burgeoning navigation services using digital maps provide great convenience to drivers. Nevertheless, the presence of anomalies in lane rendering map images occasionally introduces potential hazards, as such anomalies can mislead human drivers and consequently contribute to unsafe driving. In response to this concern, to accurately and effectively detect the anomalies, this study transforms lane rendering image anomaly detection into a classification problem and proposes a four-phase pipeline: data pre-processing, self-supervised pre-training with the masked image modelling (MiM) method, customised fine-tuning using cross-entropy based loss with label smoothing, and post-processing. Leveraging state-of-the-art deep learning techniques, especially those involving Transformer models, the pipeline demonstrates superior performance verified through various experiments. Notably, the self-supervised pre-training with MiM can greatly enhance the detection accuracy while significantly reducing the total training time. For instance, employing the Swin Transformer with Uniform Masking as self-supervised pretraining (Swin-Trans-UM) yielded a higher accuracy of 94.77% and an improved Area Under The Curve (AUC) score of 0.9743 compared with the pure Swin Transformer without pre-training (Swin-Trans) with an accuracy of 94.01% and an AUC of 0.9498. Furthermore, fine-tuning epochs were dramatically reduced to 41 from the original 280. Ablation study regarding techniques to alleviate the data imbalance between normal and abnormal instances further reinforces the model's overall performance. In conclusion, the proposed pipeline, with its incorporation of self-supervised pre-training using MiM and other advanced deep learning techniques, emerges as a robust solution for enhancing the accuracy and efficiency of lane rendering image anomaly detection in digital navigation systems.

This chapter is based on the journal publication:

Dong, Y., Lu, X., Li, R., Song, W., Van Arem, B., & Farah, H. (2025). Intelligent Anomaly Detection for Lane Rendering Using Transformer with Self-Supervised Pre-Training and Customized Fine-Tuning. Transportation Research Record: Journal of the Transportation Research Board. https://doi.org/10.1177/03611981251333341

5.1 Introduction

With the increase of private car ownership and the emergence of information and communication technology (ICT), navigation services have become popular, gaining increasing importance, forming a crucial component in driving, and providing convenience for drivers. Navigation services are always backed up by digital map applications (Vörös et al., 2022; L. Yang et al., 2021). A critical aspect of digital maps is the background, which is generated through data rendering. However, lane-level rendered map images may contain anomalies (errors and/or defects), such as irregular shapes and missing edges or corners. Examples of anomalies are illustrated in **Figure 5-1**. These anomalies can be confusing for human drivers, impairing their understanding and decision-making during navigation, which might result in critical unsafe situations.

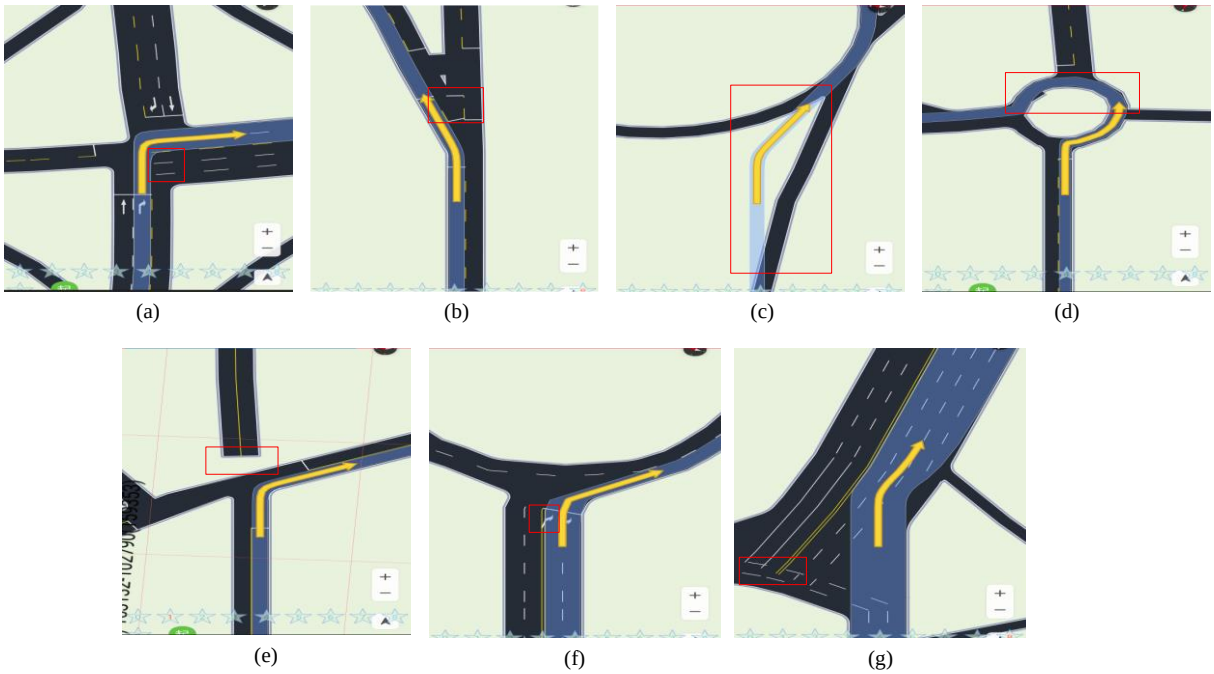


Figure 5-1. Illustration for examples of anomalous lane rendering images

Anomaly types notes: (a) Anomaly_1: The road centre line extends out of the junction; (b) Anomaly_2: The stop line is in the middle of a road; (c) Anomaly_3: The navigation route does not match actual roads; (d) Anomaly_4: The road shoulder is bumpy; (e) Anomaly_5: A part of the road is missing; (f) Anomaly_6: The road marking arrows overlap; (g) Anomaly_7: The lane lines overlap. The red boxes mark the specific regions where the anomalies are.

Similar anomalies can occur in high-definition (HD) maps used by automated vehicles (AVs) (Barsi & Barsi, 2022; Elghazaly et al., 2023). Accurate lane rendering in such maps is essential for various systems, including automated driving systems, Advanced Driver-Assistance Systems (ADAS), and smart traffic management systems, all of which rely heavily on precise and reliable mapping data to function effectively and safely. Anomalies in such maps can lead AVs into unsafe regions or induce dangerous driving behaviours.

Furthermore, this targeted problem is closely related to and can be easily transformed into relevant critical and practical real-world applications, such as road anomaly detection (Dib et al., 2020; Luo et al., 2020), road defect detection (Cao et al., 2020; Tong et al., 2020), as well as anomaly detection for lane and pavement marking on roads (Nguyen et al., 2009; Ruiz & Alzraiee, 2020; Sun et al., 2024). These issues are even more crucial for road safety. It is found that lane-related errors contribute to more than 10% of lane-change crashes (Isaksson-Hellman & Lindman, 2018), and misperception of lanes or lane boundaries is a leading factor in automated vehicle disengagements (Fu et al., 2024; Gershon et al., 2023). Thus, for example, the Federal Highway Administration (FHWA) in the USA has detailed guidelines on pavement markings essential for safe navigation and traffic management (NCUTCD, 2012). Similarly, China's Ministry of Transport emphasises the importance of accurate lane marking for reducing accidents and enhancing road safety (Ministry of Transport of the People's Republic of China, 2018).

Overall, it is vital to correctly detect these anomalies to prevent such unsafe situations. Fortunately, with the advancement of artificial intelligence algorithms, particularly in the domain of computer vision, it is now possible to carry out intelligent and automatic anomaly detection.

Conventional studies regarding anomaly detection in the relevant transportation domains principally focus on road surface anomalies (Bello-Salau et al., 2019; Dib et al., 2020), road traffic anomalies (Kumaran et al., 2020; Hengyuan Zhang et al., 2022), in-vehicle and vehicle-to-vehicle communication anomalies (Dong et al., 2022; Rajbahadur et al., 2018), abnormal driving behaviours (Hou et al., 2022; Hu et al., 2020; Dong et al., 2025), etc. Multi-modal and multi-source data have been utilised with various machine learning methods to do the detection. However, few studies have employed self-supervised methods to leverage unlabelled data. On the other hand, masked autoencoders and, to be general, masked image modelling (MiM) have become popular pre-training paradigms for self-supervised visual representation learning tasks. In MiM, a portion (usually a high ratio of 50% or above) of the input image is randomly masked using patches, and the model tries to reconstruct the masked pixels according to the target representations. The pre-trained model weights through MiM can be transferred to the downstream task for fine-tuning. Evidence in recent studies, e.g., (Bao et al., 2022; El-Nouby et al., 2021; He et al., 2022; R. Li & Dong, 2023; Xie et al., 2022), has demonstrated that self-supervised pre-training with MiM can boost the downstream tasks (e.g., classification, segmentation, and object detection) to achieve better desirable performance. Thus, it is worth exploring MiM-based pre-training for anomaly detection.

Furthermore, although various image datasets (e.g., animals, digital numbers, industrial inspection image MVTec AD datasets (Bergmann et al., 2019)) and vision-based anomaly detection methods have been developed (Bogdoll et al., 2022; Deecke et al., 2019; Kwon et al., 2020; Yan et al., 2021; J. Yang et al., 2021), to the best of the authors and after extensive review, there are no studies that tackle the abnormal lane rendering images in digital navigation maps.

To fill the aforementioned research gaps, this study develops a four-phase pipeline with self-supervised pre-training and customised fine-tuning and using state-of-the-art Transformer models (Bao et al., 2022; Dosovitskiy et al., 2021; Guo et al., 2022; X. Li et al., 2022; Liu et al., 2021; Parmar et al., 2018) to accurately and effectively detect lane rendering image anomalies. A large-scale lane rendering image dataset adjusted from the 2022 Global AI

Challenge[3] with both labelled and unlabelled data was adopted, and extensive experiments were carried out tackling the lane rendering image anomaly detection problem as a 2-class, 8-class, or 9-class classification task. Two MiM-based self-supervised pre-training methods, i.e., Uniform Masking (X. Li et al., 2022) and Bidirectional Encoder representation from Image Transformers (BEiT) (Bao et al., 2022), were customised and implemented. Extensive experiments, including ablation studies and comparative benchmarking, validate the pipeline's efficacy. To summarise, the main contributions of this study lie in:

1. Problem Reformulation: Transforming the lane rendering anomaly detection problem into a 2-class, 8-class, or 9-class classification problem.
2. Optimised Pipeline: Proposing a four-phase pipeline with especially self-supervised pre-training and customised fine-tuning to tackle the lane rendering image anomaly detection problem.
3. Utilisation and implementation of MiM Methods: Customising and implementing two MiM self-supervised pre-training methods within the proposed four-phase pipeline; extensive training, fine-tuning, and validating experiments demonstrated that with MiM the detection performance was greatly enhanced with improved AUC and reduced fine-tuning epochs.
4. State-of-the-art performance: Under the proposed pipeline, the best model delivered a performance at the accuracy of 94.82%, the Area Under the Curve (AUC) at 0.9756, and the F1-score at 0.7879, outperforming baseline models, e.g., Vision Transformer (ViT) (Dosovitskiy et al., 2021) and Swin Transformer (Liu et al., 2021).

Please note that the methods and models developed in this study can not only effectively detect lane rendering image anomalies but also can be readily adapted for related applications, such as detecting road surface anomalies and identifying abnormal lane markings.

The rest of this Chapter is arranged as follows: The next section, *Section 5.2 Methodology* describes the research methodology consisting of the proposed pipeline in detail, including the overall framework, data pre-processing, self-supervised pre-training, customised fine-tuning, and post-processing. Following this, *Section 5.3 Experiment and results* shows the experimental set-up and results comparing different models within the proposed pipeline. Then, *Section 5.4 Ablation study* introduces methods to alleviate data imbalance. Finally, *Section 5.5 Conclusion* draws the findings and proposes insights for further studies.

5.2 Methodology

In this section, the proposed method is introduced in detail. Firstly, the overall architecture of the proposed four-phase pipeline is illustrated and briefly explained. Then, each of the four phases, i.e., image pre-processing, self-supervised pre-training, fine-tuning classification, and post-processing, is depicted with comprehensive delineations sequentially.

[3] Global AI Challenge 2022:
https://developer.huawei.com/consumer/en/activity/digixActivity/digixdetail/201655283879815928

5.2.1 Overall pipeline description

This study proposes a pipeline of four phases to tackle the anomaly detection task for lane rendering images in digital navigation APPs. The overall pipeline of the four-phase method is illustrated in **Figure 5-2**.

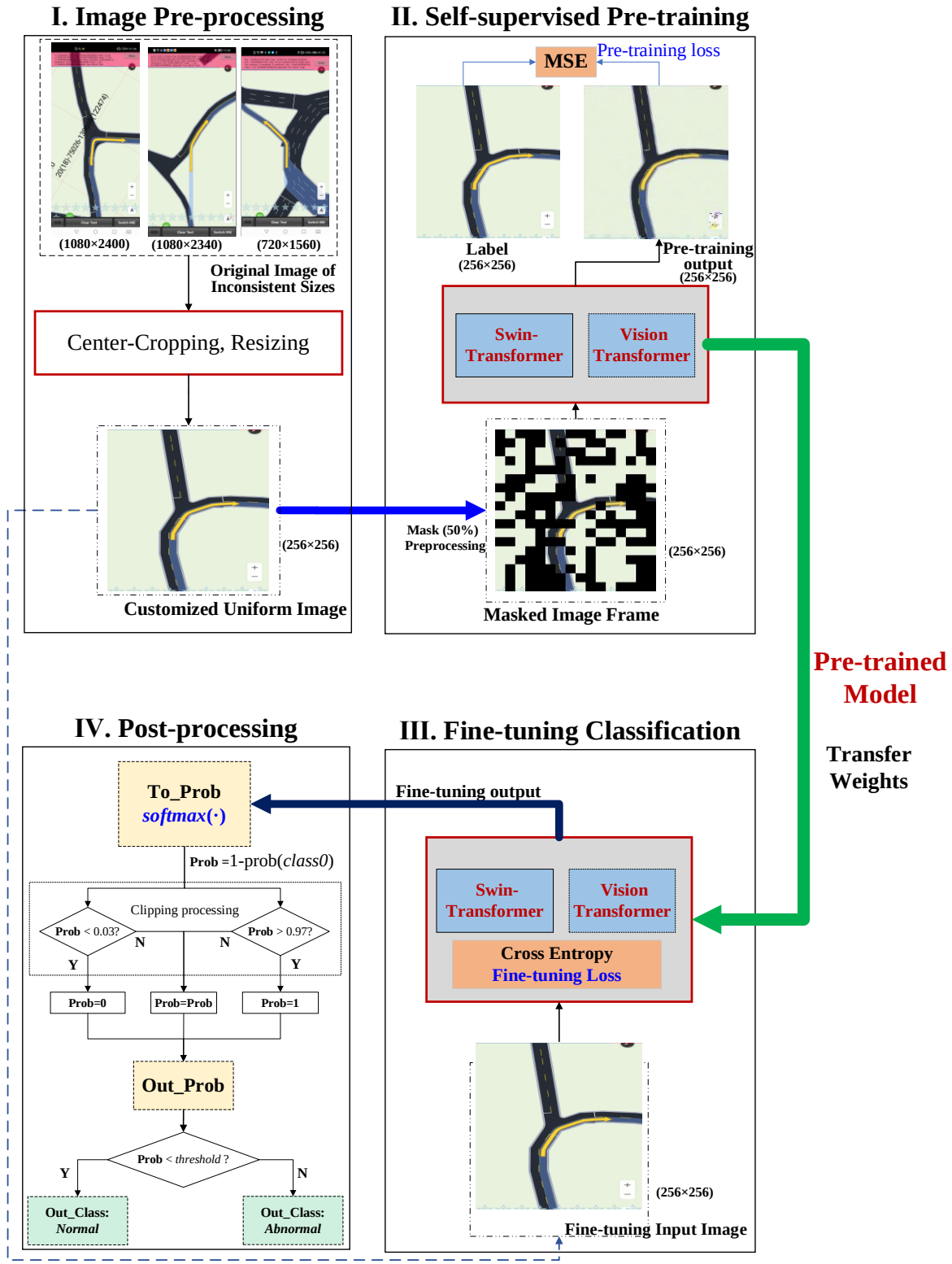


Figure 5-2. The architecture of the proposed four-phase pipeline

Note: class 0 is the normal class.

The designed 4 phases are 1) Image pre-processing, which normalises the inconsistent images into uniform size and format; 2) self-supervised pre-training, which is tackled by the masked image modelling (MiM) method using mean square error (MSE) loss and outputs the pre-trained model; 3) customised fine-tuning, which adopts the pre-trained model weights and further fine-tune the neural network model as a classification task using cross-entropy based loss (or its variants) with label smoothing; and 4) post-processing, which transforms the results of the last neural network layer (i.e., the output layer) into classification probabilities and outputs the final detection results with tuned probability threshold. The following subsections explain these four phases in more detail.

5.2.2 Image pre-processing

This study adopts the large-scale lane rendering image dataset adjusted and rearranged from the 2022 Global AI Challenge. The provided original images get different resolutions and sizes. The majority of them have a resolution of 1080 * 2400, while there are a few images with different resolutions, i.e., 1080 * 2340 and 720 * 1560. Furthermore, to focus on the relevant content of the images, the study identifies that the top and bottom portions contain non-map-related regions. Therefore, this study first carried out a centre-cropping operation by removing the 1080 * 300 pixels at the top and 1080 * 240 pixels at the bottom of the images, and then scaled the images to the same resolution of 256 * 256. Furthermore, since the images are only partly labelled with ground truth (i.e., class label of normal or anomaly type), while a large proportion of the images are unlabelled, this study constructs a pre-training dataset with both labelled images and unlabelled images, a fine-tuning dataset with partly random selected labelled image, and a testing dataset with a small proportion of the labelled images which is unseen in the fine-tuning dataset.

Similar image datasets can be created for other navigation maps by taking screenshots of the application software interface and applying the aforementioned pre-processing steps. The same process can be applied to real-world image datasets collected by cameras for anomaly detection of e.g., road lane line markings or pavement markings. It is important that after the image pre-processing phase, the images are in the uniform format, size, and resolution.

5.2.3 Self-supervised pre-training

For the lane rendering images in the navigation map APPs, lane lines account for only a small fraction of the whole image, as shown in **Figure 5-1**. There are 7 types of anomalies in the studied dataset, while the majority of the lane rendering images are normal ones. With these circumstances, it is assumed there is more spatial redundancy regarding image features for the abnormal lane rendering image detection task, and thus stronger feature extraction ability is required. Therefore, it is necessary to design a method to fully extract aggregated context information, as well as the critical features and correlations among pixels. Furthermore, as the examined dataset consists of massive unlabelled images (more than 80%), it is also vital to establish a pipeline to make full use of these unlabelled images.

Motivated by the aforementioned, this study proposes and customises the MiM method for self-supervised pre-training. In this phase, the total set of images serves as inputs for model pre-training regardless of whether labelled or unlabelled. The input image is randomly masked

using patches, and the pre-training model tries to reconstruct the masked pixels to match the target original images. Generally, the standard objective of self-supervised pre-training with MiM can be mathematically represented by equation (5-1):

$$\min \quad \frac{1}{\Omega(i_M)} ||\mathbf{r}_M - \mathbf{i}_M||_2 \tag{5-1}$$

where $\mathbf{i}, \mathbf{r} \in \mathbb{R}^{3\times H\times W}$ are the input original RGB values and the reconstructed RGB values, respectively (H is the height of the image, W is the width of the image, with $H \times W = 256 \times 256$ in this study); M represents the set of masked image pixels; $\Omega(\cdot)$ is the cardinality operator function to get the number of elements; $||\cdot||_2$ stands for ℓ_2-norm. Accordingly, the objective involves minimising the Root Mean Squared Error (RMSE), ℓ_2 loss, between the original and reconstructed pixel values for the masked regions. By focusing on accurately reconstructing the masked regions, the MiM approach encourages the model to learn rich and context-aware representations of the input image, which are crucial for downstream tasks.

Generally, there are two styles of implementing MiM: (1) raw pixel value regression, where the model directly reconstructs pixel values, and (2) converting the masked pixel signals into clusters or classes through methods such as vision tokenisation (Bao et al., 2022; Ramesh et al., 2021) or colour clustering (Chen et al., 2020), followed by performing a classification task for masked image prediction. Accordingly, this study customises and implements two distinct MiM methods, i.e., Uniform Masking (X. Li et al., 2022) and the method introduced in Bidirectional Encoder representation from Image Transformers (BEiT) (Bao et al., 2022). The Uniform Masking method was selected because it successfully enables efficient asymmetric structure, likewise in (He et al., 2022), of pixel-based Masked Autoencoder (MAE) style self-supervised pre-training, particularly for Pyramid-based Vision Transformers (ViTs). On the other hand, BEiT was selected because it serves as a typical and well-established representation of token-based methods. BEiT is the first to successfully adapt Masked Language Modelling (MLM) techniques from the Natural Language Processing (NLP) domain to the computer vision domain using ViT models. By introducing a discrete tokenisation mechanism for MiM, BEiT enables ViTs to process images in a manner analogous to how Transformers handle textual data, marking a significant milestone in bridging the gap between NLP and computer vision tasks.

Regarding the Uniform Masking method, two key operations play a central role in the self-supervised learning process:

1) ***Uniform Sampling***: This step ensures that one random patch is sampled from each 2 * 2 grid of patches within the image. As a result, 75% of the targeted region is dropped, which enforces a uniform yet sparse sampling pattern across the image.

2) ***Secondary Masking***: Since using only the uniform sampling can potentially make the self-supervisory task less challenging and largely hinders the representation quality (X. Li et al., 2022), after uniform sampling, an additional random masking operation (termed Secondary Masking) is applied to the sampled regions, further masking 25% of them (as used in this study) as shared learnable tokens.

Integrating uniform sampling and secondary masking together enables the pre-training method to support Pyramid-based ViTs, e.g., (Liu et al., 2021; Wang et al., 2021), while preserving better transferable visual representations. The Uniform Masking method pipeline for self-supervised learning is illustrated in **Figure 5-3**. The image is first divided into 16 * 16 patches

for Uniform Sampling, which drops up 75% of the original image, and the Secondary Masking is operated on the remaining patches. A compact 2D input, reduced to a quarter of the original image size, is constructed using the uniform-sampled patches combined with the secondary-masked tokens and is subsequently fed to the encoder. For the Pyramid-based ViT encoder, this study employs the Swin Transformer (Liu et al., 2021), which leverages a hierarchical architecture to effectively capture both local and global features, ensuring robust feature representation. For the decoder, the lightweight MAE Decoder, based on Vanilla ViT, is utilised, as adopted by (He et al., 2022). The MAE Decoder reconstructs the image using the encoder output features into the original size. These combinations ensure an efficient and effective architecture for self-supervised learning.

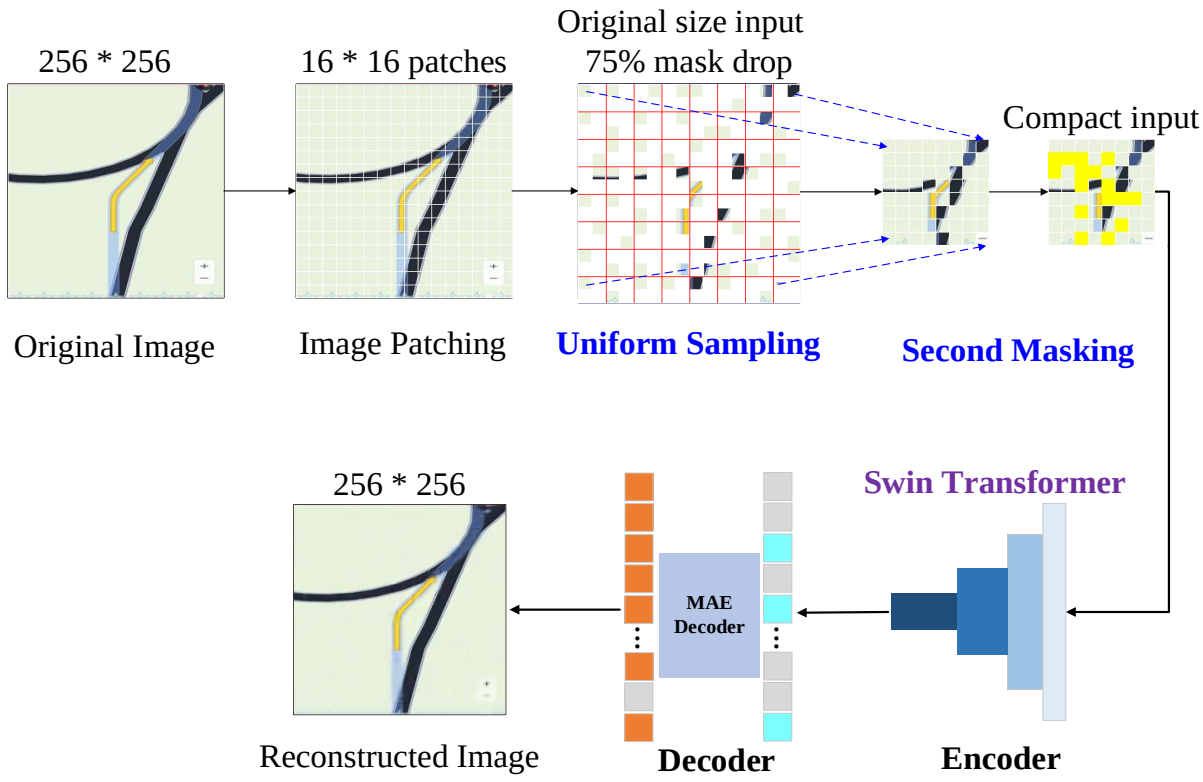


Figure 5-3. The illustration of the Uniform Masking method pipeline for MiM

The selection of the masked ratio at 75% in the uniform sampling process is based on the experiment results reported in (He et al., 2022; X. Li et al., 2022), while the selection of the secondary masking ratio of 25% is based on the ablation experiment results reported in (X. Li et al., 2022).

Regarding the BEiT self-supervised MiM method (Bao et al., 2022), each image is pre-trained with two complementary views, i.e., image patches (e.g., 16 * 16 pixels) and visual tokens (i.e., discrete tokens). **Figure 5-4** illustrates the method pipeline of BEiT for self-supervised MiM learning. The images are first "tokenised" into discrete visual tokens, which correspond to indices within a learned visual vocabulary. In this study, the visual vocabulary is generated using a discrete variational autoencoder (dVAE) tokeniser as in (Bao et al., 2022; Ramesh et al., 2021). Following tokenisation, some image patches are randomly masked and replaced with a special mask embedding before being fed into the ViT backboned encoder. Then, the objective

of the self-supervised MiM pretraining task involves predicting the visual tokens of the original image from the encoded representations of the corrupted image, which effectively enables the model to learn robust visual features. The prediction of the visual tokens is handled by the MiM head which consists of a single linear layer that converts the encoded features from the ViT encoder into a format compatible with the visual token space. Since the task involves finding the correct classes (i.e., the visual token indices), the Cross-Entropy loss function is employed for optimisation. To reconstruct the full image, the dVAE decoder takes the predicted discrete tokens as input and reconstructs their corresponding image patches. It is important to note that the MiM head is only used during the pre-training phase; during fine-tuning, task-specific decoders replace the MiM head. In this study, the original fine-tuned hyperparameters and network architecture from (Bao et al., 2022) are adopted.

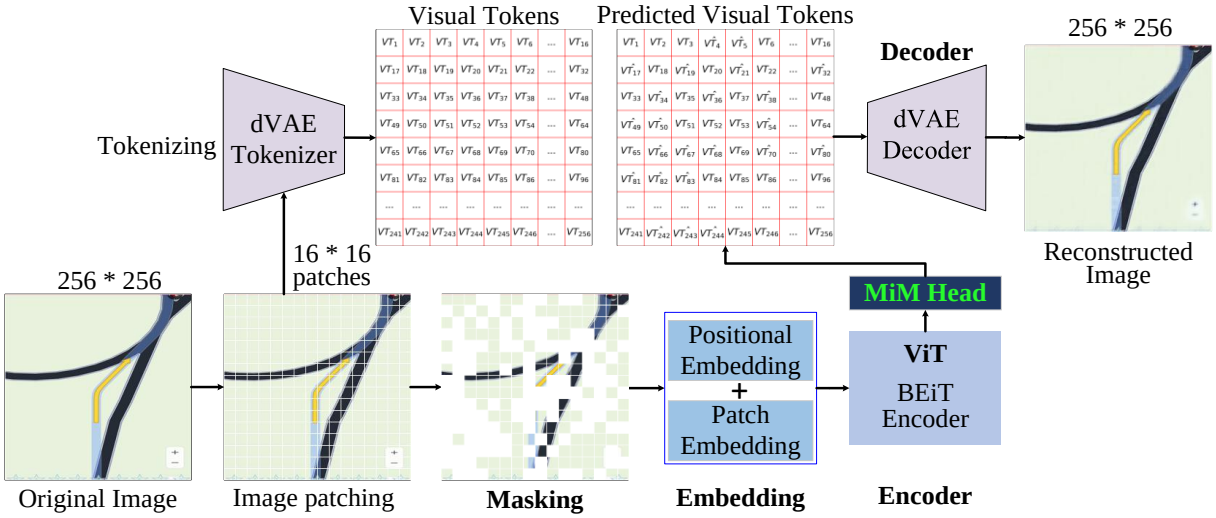


Figure 5-4. The illustration of the BEiT method pipeline for MiM

The described MiM task, implemented through either the Uniform Masking method or the BEiT method, forces the model to learn meaningful representations of images by understanding the context of the unmasked patches. For the Uniform Masking method, the Swin Transformer encoder is pre-trained using masked image regions, encouraging the model to effectively capture spatial relationships and hierarchical features. During the downstream classification task, the weights of the pre-trained Swin Transformer encoder are retained, and the MAE Decoder is replaced by a classification decoder. In contrast, for the BEiT method, the ViT encoder is pre-trained to predict discrete visual tokens corresponding to masked image regions. This approach emphasises token-based representations that align with concepts in the visual vocabulary. For the classification task, the pre-trained weights of the ViT encoder are preserved, and the MiM head is substituted with a task-specific classification decoder. Both methods leverage the robust features learned during the MiM task to enhance performance in the downstream tasks (i.e., the classification task of image types in this study), effectively transferring knowledge from the self-supervised pre-training phase to supervised fine-tuning.

This study also implemented and trained a Vision Transformer (ViT) model without the proposed self-supervised pretraining as a baseline.

5.2.4 Customised fine-tuning

In this study, the lane rendering images anomaly detection task is transferred into a 2-class, 8-class, or 9-class (multi-label) classification problem, with separating the 7 types of anomalies from the normal ones as the objective. The pre-training model weights in the self-supervised pre-training phase are transferred and further updated using the back-propagation mechanism with label smoothing Cross-Entropy as the loss function. To further boost the model performance, the MixUp technique (Hongyi Zhang et al., 2018) is adopted.

5.2.5 Post-processing

After customised fine-tuning, during the testing stage, the fine-tuned model will be applied to assign "new" testing images that are unseen in the training process into the normal class or the abnormal class. A post-processing phase is designed to aggregate the probability results and output the detection classification results.

In the post-processing, the neural network model outputs are first transformed into probabilities using $softmax(\cdot)$ function; and then the probability of each image being abnormal is calculated and truncated/clipped with up and down thresholds. After getting the truncated probability, the final detection result can be determined by fine-tuning a probability threshold to distinguish the anomalies and the normal image samples.

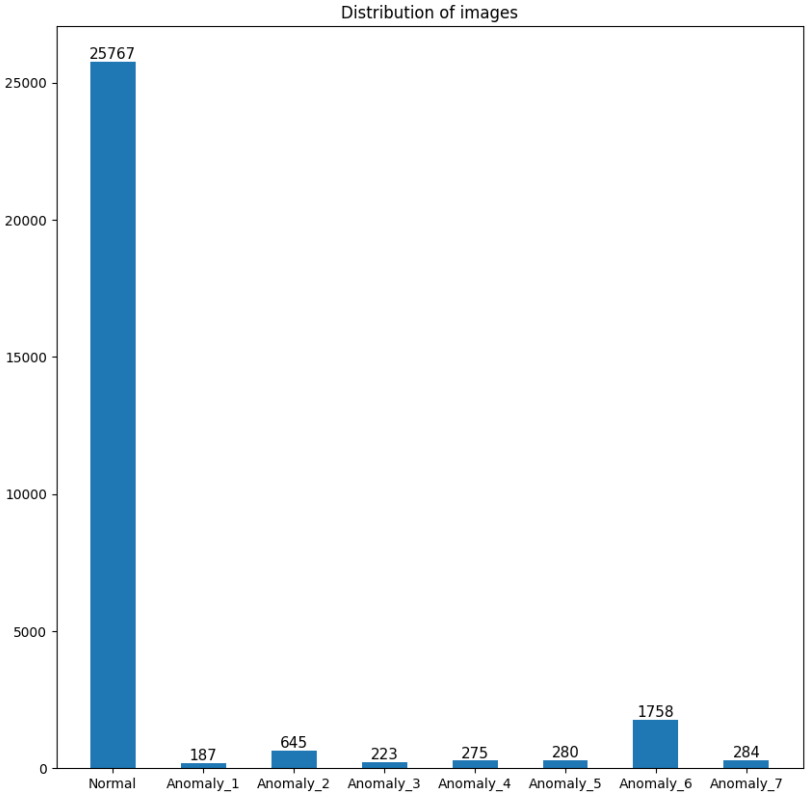
5.3 Experiments and results

To verify the effectiveness of the proposed pipeline, extensive experiments were carried out in various settings.

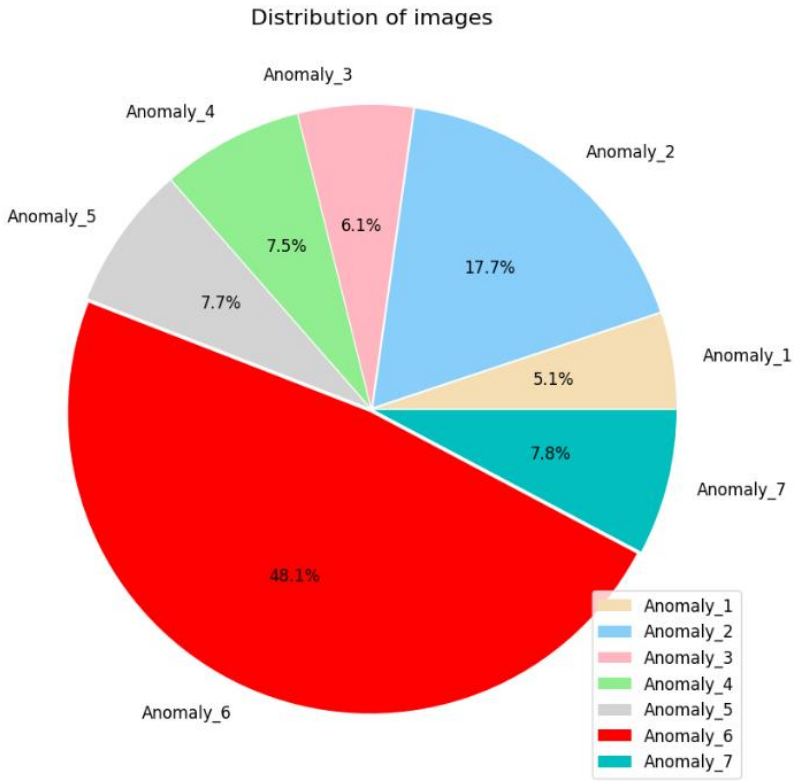
5.3.1 Dataset description

The lane-rendering digital map image data used in this study are adjusted and rearranged from the 2022 Global AI Challenge. As aforementioned, there are 7 types of anomalies, i.e., *Anomaly_1*: The road centre line extends out of the junction; *Anomaly_2*: The stop line is in the middle of a road; *Anomaly_3*: The navigation route does not match actual roads; *Anomaly_4*: The road shoulder is bumpy; *Anomaly_5*: A part of the road is missing; *Anomaly_6*: The road marking arrows overlap; and *Anomaly_7*: The lane lines overlap. Examples are shown in **Figure 5-1**.

In total, there are 161,772 images, with only 29,164 images labelled with the ground truth. Within the labelled ones, there are a total of 25,767 normal images and 3,397 images containing different kinds of anomalies (please note some images exhibit multiple different types of anomalies). **Figure 5-5 (a)** shows the histogram plot for the distribution of all labelled images, while **Figure 5-5 (b)** illustrates the pie chart for the distribution of each anomaly type within the labelled abnormal images. It is visible and clearly observed that within the 29,164 labelled images, the majority are normal ones. Furthermore, as illustrated in **Figure 5-5**, certain types of anomalies (e.g., Anomaly_6 and Anomaly_2) account for more samples than the other types of anomalies. Typically, Anomaly_6 takes up nearly half (48.1%) of the total quantity of abnormal images.



(a)



(b)

Figure 5-5. The distribution of labelled images: (a) histogram plot for the distribution of all labelled images and (b) pie chart for the distribution of each anomaly type within the labelled abnormal images

The labelled dataset is then randomly split into the training set, validation set, and test set, according to the ratio of 70%, 15%, and 15%, respectively. The images were classified according to error types, and images with multiple error types were put into multiple categories. Thus, it is a multi-class multi-label classification problem, and there are a few more training examples than the image quantity. To be specific, in practice, the number of instances in the training set is 20,764, the number of instances in the validation set is 4,310, and the number of instances in the test set is 4,346. However, all the available 161,772 images, regardless of whether labelled or not, are adopted in the self-supervised pre-training process.

5.3.2 Tested Transformer models

Two Transformer models, i.e., Vision Transformer (ViT) (Dosovitskiy et al., 2021) and Swin Transformer (Liu et al., 2021), are implemented and tested in this study. The two Transformer models are tested in modes of both with and without the self-supervised pre-training. Therefore, there are in total four model variants, i.e., 1) pure ViT without pretraining, 2) ViT variant, BEiT, with the pretraining method described in (Bao et al., 2022), 3) pure Swin Transformer (Swin-Trans for short), and 4) Swin Transformer with the Uniform Masking as self-supervised pre-training method (Swin-Trans-UM for short). The detailed model architectures, i.e., parameter settings for each layer of the tested models, are illustrated in Appendix **Table 5-A1**, **Table 5-A2**, **Table 5-A3**, and **Table 5-A4**.

5.3.3 Evaluation metrics

Various metrics are used to evaluate the overall performance of the selected models. Four basic terms, i.e., True-positive (TP) which represents the number of correctly detected lane rendering image anomalies, True-negative (TN) which represents the number of correctly detected normal lane rendering images, False-positive (FP) which represents the number of incorrectly detected anomalies, and False-negative (FN) which represents the number of incorrectly detected normal lane rendering images, are first obtained. Then, based on the four basic metrics, accuracy, precision, and recall were calculated.

Accuracy is the percentage of correctly predicted lane rendering image samples in regard to the total sample size, which can be defined as the following equation (5-2):

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{5-2}$$

Precision is the number of correctly predicted positive lane rendering image anomalies as a percentage of the total number of predicted positive anomaly observations, and it shows how close the measurements are to each other. The mathematical expression of precision is defined by equation (5-3):

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{5-3}$$

Recall ratio, illustrated in equation (5-4), is the percentage of positive anomaly observations correctly predicted in the actual category.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5-4}$$

Finally, the F1-score (F1 for short) provides an overall view of recall and precision (weighted average). F1 ranges from 0.0 to 1.0, with 1.0 indicating perfect precision and recall. And F1 can be obtained using the following equation (5-5):

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{5-5}$$

Another appropriate indicator for evaluating the two-class classification problem is the Receiver Operating Characteristic-Area Under the Curve (ROC AUC), commonly abbreviated as AUC. AUC assesses the model's ability to distinguish between normal and anomalous instances. It provides a single scalar value summarising the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different thresholds, offering insights into the model's classification performance regardless of the specific threshold applied. Given its threshold-independent nature and its ability to encapsulate the model's discriminative power, AUC is particularly suitable for imbalanced classification problems, such as the lane rendering image anomaly detection studied in this study. Accordingly, this study selects AUC as the primary evaluation metric for comparing and assessing the performance of the tested models.

To measure AUC, one needs the TPR, i.e., recall ratio, and the FPR. TPR and TNR can be obtained by the following two equations (5-6) and (5-7):

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5-6}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \tag{5-7}$$

5.3.4 Experiment set-up

Configuration details: In this study, to reduce the computational payload and save training time, the size of the images for both the training set and test set is set to a resolution of 256×256. In pre-training, the proportion of masked patches is set to 75%. Experiments were carried out on four NVIDIA Tesla V100 (32 GB memory) GPUs, using PyTorch version 1.9.0 with CUDA Deep Neural Network library (cuDNN) version 11.1. The batch size is set to be as large as possible, which is 60. The learning rate was initially set to 0.001 with decay applied after each epoch.

Data augmentation: A data augmentation technique, MixUp (Hongyi Zhang et al., 2018), where two samples (inputs and their labels) are linearly combined, is adopted to upgrade the model performance. The idea of MixUp is to create new synthetic samples to encourage the model to make predictions based on more diverse data.

The new synthetic training sample $(\tilde{x},\ \tilde{y})$ is given by equation (5-8):

$$\tilde{x} = \lambda x_a + (1-\lambda) x_b,\ \tilde{y} = \lambda y_a + (1-\lambda) y_b \tag{5-8}$$

where x_a, x_b are two raw input sample vectors, y_a, y_b are the corresponding one-hot encoded labels, λ is the MixUp parameter.

The MixUp technique helps the model generalise better by exposing it to more interpolated data points, leading to smoother decision boundaries.

Loss function details: As mentioned before, to make the proposed 4-phase pipeline work, different loss functions are adopted accordingly in the pre-training and fine-tuning phases. In the pre-training phase, the mean square error (MSE) is selected as the loss function for the Uniform Masking method since its objective is to reconstruct the masked patches directly at the pixel level. While the Cross-Entropy loss function is employed for the BEiT method since its MiM task involves identifying the correct visual token indices, framing the problem as a classification task over a visual vocabulary.

In the fine-tuning phase, the objective is to classify the lane rendering images into normal ones and anomalies, which can be regarded as a typical classification task. The Cross-Entropy loss with label smoothing is adopted for this imbalanced classification task, which is illustrated in equation (5-9):

$$\ell_{\mathrm{CE}} = \ell(y, \hat{y}) = -(1-\varepsilon)\, log\big(\hat{y}_y\big) - \frac{\varepsilon}{C-1}\sum_{c \neq y} log(\hat{y}_c) \tag{5-9}$$

where C is the number of classes; y is the one-hot encoded true label; $\hat{y}$ is the predicted probabilities output by the model, for example, $\hat{y}_y$ is the predicted probability for the true class, and $\hat{y}_c$ is the predicted probability for the true class c; ε is the smoothing factor controlling the amount of uncertainty applied, usually set between 0 and 1.

With label smoothing, the true labels are adjusted to distribute some of the target probability mass to other classes. The overall effect of this modification is to provide a softer target. The model is less confident solely on one class, promoting better learning from non-ideal scenarios, such as label noise or ambiguity, and potentially improving generalisation.

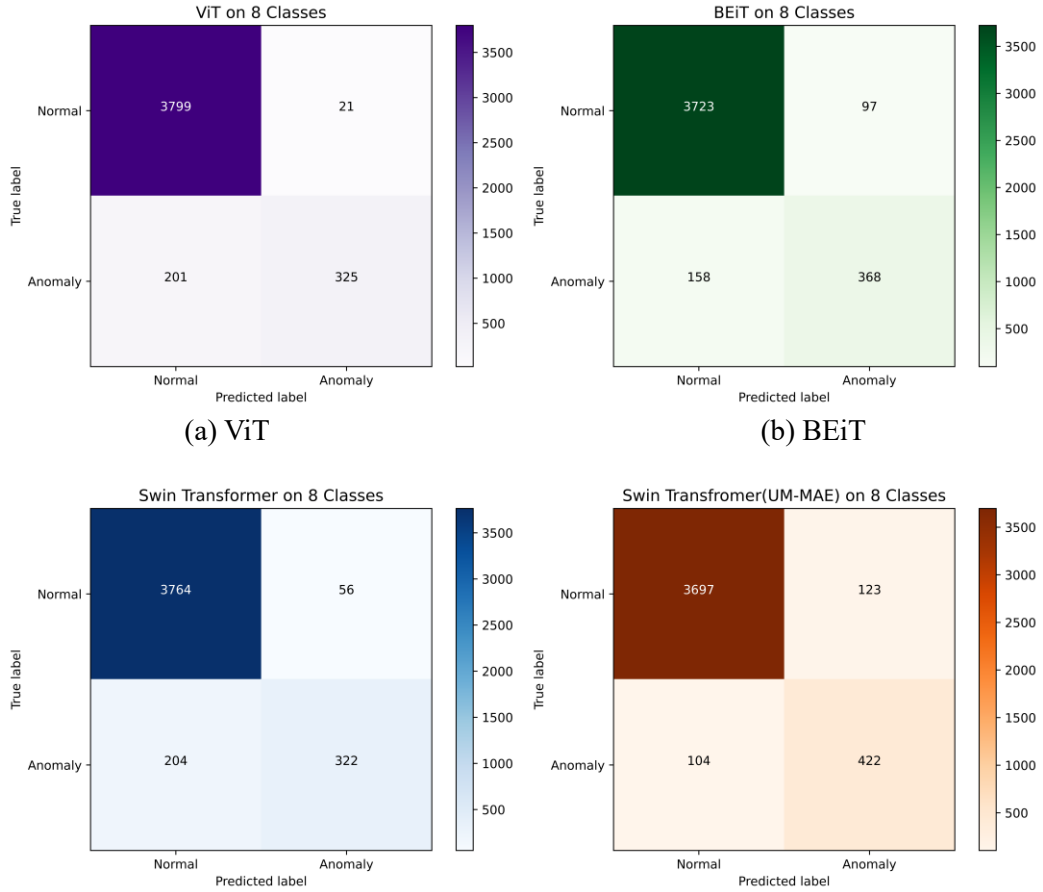
Optimiser details: To efficiently train and validate the proposed model pipeline, different optimisers were tested in different stages. Four optimisers, Stochastic Gradient Descent (SGD), Adaptive Moment Estimation (Adam), Rectified Adaptive Moment Estimation (RAdam), and Adam with decoupled weight decay (AdamW) (Loshchilov & Hutter, 2019), were tested in the pre-training and fine-tuning segmentation phases. Through the tests, AdamW performed the best in both the pre-training and the fine-tuning phases, therefore, it was finally chosen for both the two phases.

For other hyperparameters and experiment implementations, this study generally follows the fine-tuned settings reported in (Bao et al., 2022; He et al., 2022; X. Li et al., 2022).

5.3.5 Results

Various experiments were carried out to compare the model performance of the four tested Transformer models, i.e., pure ViT, pure Swin Transformer (Swin-Trans), ViT variant with self-supervised pretraining (BEiT), and Swin Transformer with Uniform Masking (Swin-Trans-UM). The obtained results of treating the problem as an 8-class classification task are illustrated in **Figure 5-6** and **Table 5-1**.

From **Table 5-1**, it is evident that the significant differences in the number of fine-tuning epochs stem from the influence of the adopted MiM pre-training. The stopping criterion utilised in this study is AUC convergence. Specifically, fine-tuning is terminated when the improvement in AUC between consecutive evaluation epochs falls below a predefined threshold, signalling that the model's performance has stabilised.



(a) ViT (b) BEiT

(c) Swin Transformer (d) Swin-Trans-UM

Figure 5-6. The testing results of the models visualised in confusion matrices

Table 5-1. The model performance regarding different metrics

Model	Accuracy	AUC	Precision	Recall	F1-score	Param (M)	Epoch Time (s)	Number of Fine-tuning Epochs
ViT	0.9489	0.9080	**0.9393**	0.6178	0.7454	632.20	4210	40
BEiT	0.9413	0.9481	0.7913	0.6996	0.7427	311.53	159	15
Swin-Trans	0.9401	0.9498	0.8518	0.6121	0.7123	86.90	120	280
Swin-Trans-UM	**0.9477**	**0.9743**	0.7743	**0.8022**	**0.7805**	194.95	223	41

With MiM pre-training, the Swin-Trans-UM and BEiT models converge in 15 epochs and 41 epochs, respectively. In contrast, without MiM pre-training, the original Vanilla ViT requires 40 epochs, and the original Vanilla Swin Transformer demands 280 epochs to converge.

The adoption of MiM pre-training considerably reduces the total number of fine-tuning epochs needed for convergence. This is achieved by equipping the model with rich, context-aware semantic features during pre-training, which provide a robust initialisation for the downstream classification task. As a result, models with MiM pre-training not only converge faster but also

maintain or improve their classification accuracy. This observed disparity underscores the efficiency and effectiveness of MiM pre-training in lowering computational requirements while delivering high performance.

Furthermore, regarding the primary and the most suitable overall model performance evaluation metric, AUC, both BEiT and Swin-Trans-UM outperform their variants without self-supervised pre-training, i.e., ViT and Swin-Trans. Especially, among the four models, Swin-Trans-UM obtains the best performance regarding Accuracy (94.77%), AUC (0.9743), Recall (0.8022), and F1-score (0.7805).

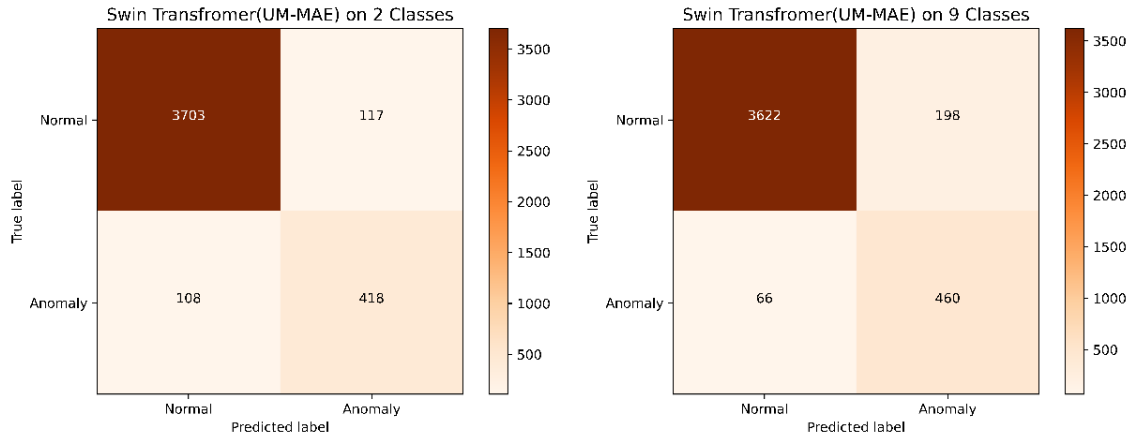
5.4 Ablation study

It is easy to identify that the quantity of abnormal and normal image samples is highly imbalanced. To alleviate this imbalance, two ablation studies are carried out using the Swin-Trans-UM model, regarding the abnormal lane rendering detection not as the original 8-class multi-label classification problem but as a 2-class classification problem (Swin-Trans-UM_2 as the corresponding model) or 9-class multi-label classification problem (Swin-Trans-UM_9 as the corresponding model) in the fine-tuning process.

5.4.1 Treated as a 2-class classification

When treated as a 2-class image classification problem, all abnormal images are grouped as one class, and together with the normal class, there are 2 classes in the fine-tuning process. In this way, the imbalance between the classes is alleviated since grouping abnormal classes together reduces the disparity between the number of normal instances and anomalies. By consolidating the abnormal classes into a single group, the number of anomaly-related instances is less sparse, making the distribution more balanced compared to treating each anomaly type separately.

The results of the tested Swin-Trans-UM_2 model performance under this setting are demonstrated in **Figure 5-7 (a)** and **Table 5-2**. It is evident that, except for Recall, all the other reported evaluation metrics (i.e., Accuracy, AUC, Precision, F1-score) for Swin-Trans-UM_2 are improved compared to the original approach which treats the problem as an 8-class classification (Swin-Trans-UM_8).



(a) Swin-Trans-UM_2 (b) Swin-Trans-UM_9

Figure 5-7. The confusion matrix of Swin-Trans-UM when treated as (a) a 2-class classification and (b) a 9-class multi-label classification

Table 5-2. The performance of the Swin-Trans-UM_2 and Swin-Trans-UM_9

Model	Accuracy	AUC	Precision	Recall	F1-score
Swin-Trans-UM_2	**0.9482**	**0.9756**	**0.7813**	0.7947	**0.7879**
Swin-Trans-UM_9	0.9392	0.9731	0.6990	**0.8745**	0.7770
Swin-Trans-UM_8	0.9477	0.9743	0.7743	0.8022	0.7805

5.4.2 Treated as a 9-class multi-label classification

When treated as a 9-class multi-label image classification problem, all abnormal images are grouped as one extra integrated class while still keeping each sub-abnormal class as in the dataset. Thus, 9 classes are obtained, and each abnormal instance will get at least two class labels. In this way, the imbalance between the classes is further alleviated. The results of the tested Swin-Trans-UM_9 model performance under this setting are demonstrated in **Figure 5-7 (b)** and **Table 5-2**. Except for Recall, all the other evaluation metrics of Swin-Trans-UM_9 are degraded compared with the original approach treated as an 8-class classification problem (Swin-Trans-UM_8). This might be due to the extra label for each abnormal instance confusing the model during the fine-tuning process when updating the model weights by backpropagation. Detailed reasons need further study.

5.5 Conclusions, limitations, and future research

Lane rendering is an important element in digital maps used for navigation services and other traffic-related applications. However, there might be anomalies in the lane rendering images. To accurately and effectively detect the anomalies, this study converts the problem of lane rendering image anomaly detection to a classification problem, which allows various state-of-the-art computer vision techniques to be applicable. Furthermore, this study proposes a four-phase pipeline consisting of data pre-processing, self-supervised pre-training with the masked image modelling (MiM) method, customised fine-tuning using cross-entropy loss with label smoothing, and post-processing. Various metrics are adopted to evaluate the model performance. Extensive experiments have demonstrated that the proposed pipeline effectively addresses the lane rendering image anomaly detection task, achieving outstanding performance in terms of high accuracy. And especially, the self-supervised pre-training with MiM can greatly improve the model accuracy, e.g., Swin Transformer with Uniform Masking as self-supervised pretraining (Swin-Trans-UM) obtained better accuracy at 94.77% and better AUC at 0.9743 compared with the pure Swin Transformer without pre-training (Swin-Trans) whose accuracy is 94.01%, AUC is 0.9498, while significantly reducing the model fine-tuning time, e.g., Swin-Trans-UM reduced the number of epochs of Swin-Trans at 280 to only 41. Ablation study regarding techniques to alleviate the data imbalance between normal and abnormal instances further enhances the model performance, with the 2-class classification variant of the Swin-Trans-UM model, i.e., Swin-Trans-UM_2 obtained the best performance on almost all the evaluation metrics, i.e., Accuracy (94.82%), AUC (0.9756), Precision (0.7813), and F1-score (0.7879). Lastly, regarding the societal benefits, the proposed method can improve the efficiency of lane rendering image data anomaly detection, reducing labour costs while keeping high accuracy.

As for limitations, due to the unavailability of other relevant datasets, this study only examined and evaluated the proposed method and results on a single dataset, which might potentially constrain the generalisability of the proposed method and corresponding results. Furthermore, limited by the properties of the data, the focus of this study is confined to discerning whether the lane rendering image is abnormal or normal. Further investigation into checking and diagnosing the specific anomaly types, as well as locating the anomalies within the images, could be intriguing directions for future studies. This would involve more detailed anomaly segmentation, which could provide valuable deeper insights into the nature and causes of detected anomalies. However, achieving such advancements would necessitate access to structured datasets equipped with labelled segmentation maps to facilitate robust anomaly localisation and classification tasks.

Moreover, certain anomaly images in the dataset have multiple labels, a complexity that this study did not address. Future studies should explore methods for handling multi-label classification to account for overlapping or co-occurring anomalies. Techniques such as multi-label learning algorithms (M. L. Zhang & Zhou, 2014), label correlation modelling (Yu et al., 2014; Zhu et al., 2018), or hierarchical classification approaches (Wehrmann et al., 2018) could be explored to tackle this issue. Addressing multi-label scenarios would enhance the robustness and applicability of anomaly detection systems in real-world contexts.

Lastly, the current study employs a supervised approach during the fine-tuning phase, necessitating high-quality ground truth labels. Future studies could explore the potential of semi-supervised or unsupervised machine learning approaches to distinguish anomalies from normal instances without relying on extensive labelled data. For example, Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) can perform zero-shot classification by learning from large-scale, unannotated data, aligning images with textual descriptions. Similarly, Bootstrapping Language-Image Pre-training (BLIP) (J. Li et al., 2022) can effectively perform image-text matching tasks in a self-supervised manner, which could help classify anomalies with minimal reliance on labelled data.

Acknowledgements

This work was supported by the Applied and Technical Sciences (TTW), a subdomain of the Dutch Institute for Scientific Research (NWO) through the Project *Safe and Efficient Operation of Automated and Human-Driven Vehicles in Mixed Traffic* (SAMEN) under Contract 17187. The authors thank the 2022 Global AI Challenge for providing the original data.

References

Bao, H., Dong, L., Piao, S., & Wei, F. (2022). BEIT: Bert pre-training of image transformers. ICLR 2022 - 10th International Conference on Learning Representations.

Barsi, M., & Barsi, A. (2022). Topological anomaly detection in automotive simulator maps. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives. https://doi.org/10.5194/isprs-archives-XLIII-B4-2022-343-2022

Bello-Salau, H., Onumanyi, A. J., Salawudeen, A. T., Mu'Azu, M. B., & Oyinbo, A. M. (2019). An examination of different vision based approaches for road anomaly detection. 2019

2nd International Conference of the IEEE Nigeria Computer Chapter, NigeriaComputConf 2019. https://doi.org/10.1109/NigeriaComputConf45974.2019.8949646

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD-A comprehensive real-world dataset for unsupervised anomaly detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2019.00982

Bogdoll, D., Nitsche, M., & Zollner, J. M. (2022). Anomaly detection in autonomous driving: A survey. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2022-June, 4487–4498. https://doi.org/10.1109/CVPRW56347.2022.00495

Cao, W., Liu, Q., & He, Z. (2020). Review of pavement defect detection methods. IEEE Access. https://doi.org/10.1109/aCCESS.2020.2966881

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. 37th International Conference on Machine Learning, ICML 2020.

Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., & Kloft, M. (2019). Image anomaly detection with generative adversarial networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-030-10925-7_1

Dib, J., Sirlantzis, K., & Howells, G. (2020). A Review on negative road anomaly detection methods. IEEE Access. https://doi.org/10.1109/ACCESS.2020.2982220

Dong, Y., Chen, K., Peng, Y., & Ma, Z. (2022). Comparative study on supervised versus semi-supervised machine learning for anomaly detection of in-vehicle CAN network. In 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC) (pp. 2914-2919). IEEE. https://doi.org/10.1109/ITSC55140.2022.9922235

Dong, Y., Zhang, L., Farah, H., Zgonnikov, A., & Van Arem, B. (2025). Data-driven semi-supervised machine learning with safety indicators for abnormal driving behavior detection. Transportation Research Record: Journal of the Transportation Research Board, 1-16. https://doi.org/10.1177/03611981241306752

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021 - 9th International Conference on Learning Representations.

El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., & Grave, E. (2021). Are large-scale datasets necessary for self-supervised pre-training? ArXiv Preprint ArXiv:2112.10740.

Elghazaly, G., Frank, R., Harvey, S., & Safko, S. (2023). High-definition maps: Comprehensive survey, challenges, and future perspectives. IEEE Open Journal of Intelligent Transportation Systems. https://doi.org/10.1109/OJITS.2023.3295502

Fu, Y., Seemann, J., Hanselaar, C., Beurskens, T., Terechko, A., Silvas, E., & Heemels, M. (2024). Characterization and mitigation of insufficiencies in automated driving systems. arXiv preprint arXiv:2404.09557.

Gershon, P., Mehler, B., & Reimer, B. (2023). Driver response and recovery following automation initiated disengagement in real-world hands-free driving. Traffic Injury Prevention. https://doi.org/10.1080/15389588.2023.2189990

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., Zhang, S. H., Martin, R. R., Cheng, M. M., & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. Computational Visual Media, 8(3), 331–368. https://doi.org/10.1007/s41095-022-0271-y

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR52688.2022.01553

Hou, M., Wang, M., Zhao, W., Ni, Q., Cai, Z., & Kong, X. (2022). A lightweight framework for abnormal driving behavior detection. Computer Communications, 184(May 2021), 128–136. https://doi.org/10.1016/j.comcom.2021.12.007

Hu, J., Zhang, X., & Maybank, S. (2020). Abnormal driving detection with ed driving behavior data: A deep learning approach. IEEE Transactions on Vehicular Technology, 69(7), 6943–6951. https://doi.org/10.1109/TVT.2020.2993247

Isaksson-Hellman, I., & Lindman, M. (2018). An evaluation of the real-world safety effect of a lane change driver support system and characteristics of lane change crashes based on insurance claims data. In Traffic Injury Prevention. https://doi.org/10.1080/15389588.2017.1396320

Kumaran, S. K., Dogra, D. P., & Roy, P. P. (2020). Anomaly detection in road traffic using visual surveillance: A survey. ACM Computing Surveys (CSUR), 53(6), 1-26.

Kwon, G., Prabhushankar, M., Temel, D., & AlRegib, G. (2020). Backpropagated gradient representations for anomaly detection. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16 (pp. 206-226). Springer International Publishing. https://doi.org/10.1007/978-3-030-58589-1_13

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning (pp. 12888-12900). PMLR.

Li, R., & Dong, Y. (2023). Robust lane detection through self pre-training with masked sequential autoencoders and fine-tuning with customized PolyLoss. IEEE Transactions on Intelligent Transportation Systems, 24(12), 14121–14132. https://doi.org/10.1109/TITS.2023.3305015

Li, X., Wang, W., Yang, L., & Yang, J. (2022). Uniform masking: Enabling MAE pre-training for pyramid-based vision Transformers with locality. 1–14. http://arxiv.org/abs/2205.10063

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision Transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992–10002. https://ieeexplore.ieee.org/document/9710580/

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. 7th International Conference on Learning Representations, ICLR 2019.

Luo, D., Lu, J., & Guo, G. (2020). Road anomaly detection through deep learning approaches. IEEE Access. https://doi.org/10.1109/ACCESS.2020.3004590

Ministry of Transport of the People's Republic of China. (2018). Specifications for highway geometric design (JTG D20—2017). Industry Standards of the People's Republic of China, 1–271.

NCUTCD. (2012). Manual on uniform traffic control devices for streets and highways - MUTCD - edition 2009. In FHWA.

Nguyen, T. S., Avila, M., & Begot, S. (2009). Automatic detection and classification of defect on road pavement using anisotropy measure. In 2009 17th European Signal Processing Conference (pp. 617-621). IEEE.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image Transformer. 35th International Conference on Machine Learning, ICML 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of Machine Learning Research.

Rajbahadur, G. K., Malton, A. J., Walenstein, A., & Hassan, A. E. (2018). A survey of anomaly detection for connected vehicle cybersecurity and safety. IEEE Intelligent Vehicles Symposium, Proceedings. https://doi.org/10.1109/IVS.2018.8500383

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In International Conference on Machine Learning, pp. 8821-8831. PMLR.

Ruiz, A. L., & Alzraiee, H. (2020). Automated pavement marking defects detection. Proceedings of the 37th International Symposium on Automation and Robotics in Construction, ISARC 2020: From Demonstration to Practical Use - To New Stage of Construction Robot. https://doi.org/10.22260/isarc2020/0011

Sun, Y., Tang, H., & Zhang, H. (2024). Automatic detection of pavement marking defects in road inspection images using deep learning. Journal of Performance of Constructed Facilities. https://doi.org/10.1061/jpcfev.cfeng-4619

Tong, Z., Yuan, D., Gao, J., & Wang, Z. (2020). Pavement defect detection with fully convolutional network and an uncertainty framework. Computer-Aided Civil and Infrastructure Engineering. https://doi.org/10.1111/mice.12533

Vörös, F., Gartner, G., Peterson, M. P., & Kovács, B. (2022). What does the ideal built-in car navigation system look like?—An investigation in the central European region. Applied Sciences (Switzerland), 12(8). https://doi.org/10.3390/app12083716

Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid vision Transformer: A versatile backbone for dense prediction without convolutions. Proceedings of the IEEE International Conference on Computer Vision. https://doi.org/10.1109/ICCV48922.2021.00061

Wehrmann, J., Cerri, R., & Barros, R. C. (2018). Hierarchical multi-label classification networks. 35th International Conference on Machine Learning, ICML 2018.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., & Hu, H. (2022). SimMIM: A simple framework for masked image modeling. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR52688.2022.00943

Yan, X., Zhang, H., Xu, X., Hu, X., & Heng, P. A. (2021). Learning semantic context from normal samples for unsupervised anomaly detection. 35th AAAI Conference on Artificial Intelligence, AAAI 2021, 4A, 3110–3118. https://doi.org/10.1609/aaai.v35i4.16420

Yang, J., Xu, R., Qi, Z., & Shi, Y. (2021). Visual anomaly detection for images: A systematic survey. Procedia Computer Science, 199(2021), 471–478. https://doi.org/10.1016/j.procs.2022.01.057

Yang, L., Bian, Y., Zhao, X., Liu, X., & Yao, X. (2021). Drivers' acceptance of mobile navigation applications: An extended technology acceptance model considering drivers' sense of direction, navigation application affinity and distraction perception. International Journal of Human Computer Studies. https://doi.org/10.1016/j.ijhcs.2020.102507

Yu, Y., Pedrycz, W., & Miao, D. (2014). Multi-label classification by exploiting label correlations. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2013.10.030

Zhang, Hengyuan, Zhao, S., Liu, R., Wang, W., Hong, Y., & Hu, R. (2022). Automatic traffic anomaly detection on the road network with spatial-temporal graph neural network representation learning. Wireless Communications and Mobile Computing. https://doi.org/10.1155/2022/4222827

Zhang, Hongyi, Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). MixUp: Beyond empirical risk minimization. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 1–13.

Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. In IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2013.39

Zhu, Y., Kwok, J. T., & Zhou, Z. H. (2018). Multi-label learning with global and local label correlation. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2017.2785795

Appendix

Note: *The following neural network structures are based upon 8-class classification in the fine-tuning phase. There are a few minor differences regarding the output layers for the models used in the self-supervised pretraining phase or for the 2-class and 9-class classifications.*

Multiply-Add, short for multiply-accumulate operation, which means computing the product of two numbers and adding that product to an accumulator. It is used as shorthand for the total number of operations in the model as popular layers such as convolution and linear layers multiply weights with inputs and then add the results of the multiplication (possibly with a bias).

Table 5-A1. Parameter settings for each layer of Vision Transformer

Layer	Kernel Shape	Input Shape	Output Shape	Param	Mult-Adds
VisionTransformer	--	[1, 3, 224, 224]	[1, 8]	253,440	--
PatchEmbed	--	[1, 3, 224, 224]	[1, 196, 1280]	--	--
Conv2d	[16, 16]	[1, 3, 224, 224]	[1, 1280, 14, 14]	984,320	192,926,720
Dropout	--	[1, 197, 1280]	[1, 197, 1280]	--	--
ModuleList (Consisting of 32 Blocks with the same structure as below)					
Block 1-32	--	[1, 197, 1280]	[1, 197, 1280]	--	--
LayerNorm	--	[1, 197, 1280]	[1, 197, 1280]	2,560	2,560
Attention	--	[1, 197, 1280]	[1, 197, 1280]	6,554,880	6,554,880
Identity	--	[1, 197, 1280]	[1, 197, 1280]	--	--
LayerNorm	--	[1, 197, 1280]	[1, 197, 1280]	2,560	2,560
Mlp	--	[1, 197, 1280]	[1, 197, 1280]	13,113,600	13,113,600
Identity	--	[1, 197, 1280]	[1, 197, 1280]	--	--
LayerNorm	--	[1, 197, 1280]	[1, 197, 1280]	2,560	2,560
Linear	--	[1, 1280]	[1, 8]	10,248	10,248

Table 5-A2. Parameter settings for each layer of BEiT

Layer	Kernel Shape	Input Shape	Output Shape	Param	Mult-Adds
BEiT	--	[1, 3, 224, 224]	[1, 8]	768	--
PatchEmbed	--	[1, 3, 224, 224]	[1, 196, 768]	--	--
Conv2d	[16, 16]	[1, 3, 224, 224]	[1, 768, 14, 14]	590,592	115,756,032
Dropout	--	[1, 197, 768]	[1, 197, 768]	--	--
ModuleList (Consisting of 12 Blocks with the same structure as below)					
Block 1-12	--	[1, 197, 768]	[1, 197, 768]	1,536	--
LayerNorm	--	[1, 197, 768]	[1, 197, 768]	1,536	1,536
Attention	--	[1, 197, 768]	[1, 197, 768]	2,370,384	590,592
Identity	--	[1, 197, 768]	[1, 197, 768]	--	--
LayerNorm	--	[1, 197, 768]	[1, 197, 768]	1,536	1,536
Mlp	--	[1, 197, 768]	[1, 197, 768]	4,722,432	4,722,432
Identity	--	[1, 197, 768]	[1, 197, 768]	--	--
LayerNorm	--	[1, 768]	[1, 768]	1,536	1,536
Linear	--	[1, 768]	[1, 8]	6,152	6,152

Table 5-A3. Parameter settings for each layer of Swin Transformer

Layer (type:depth-idx)	Kernel Shape	Input Shape	Output Shape	Param	Mult-Adds
SwinTransformerV2	--	[1, 3, 256, 256]	[1, 8]	--	--
PatchEmbed	--	[1, 3, 256, 256]	[1, 4096, 96]	--	--
Conv2d	[4, 4]	[1, 3, 256, 256]	[1, 96, 64, 64]	4,704	19,267,584
LayerNorm	--	[1, 4096, 96]	[1, 4096, 96]	192	192
Dropout	--	[1, 4096, 96]	[1, 4096, 96]	--	--
ModuleList					
BasicLayer	--	[1, 4096, 96]	[1, 1024, 192]	--	--
ModuleList	--	--	--	--	--
SwinTransformerBlock	--	[1, 4096, 96]	[1, 4096, 96]	114,819	673,632
SwinTransformerBlock	--	[1, 4096, 96]	[1, 4096, 96]	114,819	673,632
PatchMerging	--	[1, 4096, 96]	[1, 1024, 192]	--	--
Linear	--	[1, 1024, 384]	[1, 1024, 192]	73,728	73,728
LayerNorm	--	[1, 1024, 192]	[1, 1024, 192]	384	384
BasicLayer	--	[1, 1024, 192]	[1, 256, 384]	--	--
ModuleList	--	--	--	--	--
SwinTransformerBlock	--	[1, 1024, 192]	[1, 1024, 192]	449,286	894,144
SwinTransformerBlock	--	[1, 1024, 192]	[1, 1024, 192]	449,286	894,144
PatchMerging	--	[1, 1024, 192]	[1, 256, 384]	--	--
Linear	--	[1, 256, 768]	[1, 256, 384]	294,912	294,912
LayerNorm	--	[1, 256, 384]	[1, 256, 384]	768	768
BasicLayer	--	[1, 256, 384]	[1, 64, 768]	--	--
ModuleList	--	--	--	--	--
SwinTransformerBlock	--	[1, 256, 384]	[1, 256, 384]	1,781,772	1,782,144
SwinTransformerBlock	--	[1, 256, 384]	[1, 256, 384]	1,781,772	1,782,144
SwinTransformerBlock	--	[1, 256, 384]	[1, 256, 384]	1,781,772	1,782,144
SwinTransformerBlock	--	[1, 256, 384]	[1, 256, 384]	1,781,772	1,782,144
SwinTransformerBlock	--	[1, 256, 384]	[1, 256, 384]	1,781,772	1,782,144
SwinTransformerBlock	--	[1, 256, 384]	[1, 256, 384]	1,781,772	1,782,144
PatchMerging	--	[1, 256, 384]	[1, 64, 768]	--	--
Linear	--	[1, 64, 1536]	[1, 64, 768]	1,179,648	1,179,648
LayerNorm	--	[1, 64, 768]	[1, 64, 768]	1,536	1,536
BasicLayer	--	[1, 64, 768]	[1, 64, 768]	--	--
ModuleList	--	--	--	--	--
SwinTransformerBlock	--	[1, 64, 768]	[1, 64, 768]	7,100,952	5,329,920
SwinTransformerBlock	--	[1, 64, 768]	[1, 64, 768]	7,100,952	5,329,920
LayerNorm	--	[1, 64, 768]	[1, 64, 768]	1,536	1,536
AdaptiveAvgPool1d	--	[1, 768, 64]	[1, 768, 1]	--	--
Linear	--	[1, 768]	[1, 8]	6,152	6,152

Table 5-A4. Parameter settings for each layer of Swin Transformer with Uniform Masking

Layer (type: depth-idx)	Kernel Shape	Input Shape	Output Shape	Param	Mult-Adds
Swin (Swin)	--	[1, 3, 256, 256]	[1, 8]	--	--
PatchEmbed (patch embed): 1-1	--	[1, 3, 256, 256]	[1, 4096, 192]	--	--
Conv2d (proj): 2-1	[4, 4]	[1, 3, 256, 256]	[1, 192, 64, 64]	9,408	38,535,168
LayerNorm (norm): 2-2	--	[1, 4096, 192]	[1, 4096, 192]	384	384
ModuleList (blocks): 1-2	--	--	--	--	--
SwinBlock (0): 2-3	--	[1, 4096, 192]	[1, 4096, 192]	--	--
LayerNorm (norm1): 3-1	--	[1, 4096, 192]	[1, 4096, 192]	384	384
WindowAttention (attn): 3-2	--	[16, 256, 192]	[16, 256, 192]	148,806	612,642,816
Identity (drop path): 3-3	--	[1, 4096, 192]	[1, 4096, 192]	--	--
LayerNorm (norm2): 3-4	--	[1, 4096, 192]	[1, 4096, 192]	384	384
Mlp (mlp): 3-5	--	[1, 4096, 192]	[1, 4096, 192]	295,872	295,872
Identity (drop path): 3-6	--	[1, 4096, 192]	[1, 4096, 192]	--	--
SwinBlock (1): 2-4	--	[1, 4096, 192]	[1, 4096, 192]	--	--
LayerNorm (norm1): 3-7	--	[1, 4096, 192]	[1, 4096, 192]	384	384
WindowAttention (attn): 3-8	--	[16, 256, 192]	[16, 256, 192]	148,806	612,642,816
DropPath (drop path): 3-9	--	[1, 4096, 192]	[1, 4096, 192]	--	--
LayerNorm (norm2): 3-10	--	[1, 4096, 192]	[1, 4096, 192]	384	384
Mlp (mlp): 3-11	--	[1, 4096, 192]	[1, 4096, 192]	295,872	295,872
DropPath (drop path): 3-12	--	[1, 4096, 192]	[1, 4096, 192]	--	--
SwinBlock (2): 2-5	--	[1, 4096, 192]	[1, 1024, 384]	--	--
PatchMerge (downsample): 3-13	--	[1, 4096, 192]	[1, 1024, 384]	295,680	302,383,488
LayerNorm (norm1): 3-14	--	[1, 1024, 384]	[1, 1024, 384]	768	768
WindowAttention (attn): 3-15	--	[4, 256, 384]	[4, 256, 384]	592,332	257,169,408
DropPath (drop path): 3-16	--	[1, 1024, 384]	[1, 1024, 384]	--	--
LayerNorm (norm2): 3-17	--	[1, 1024, 384]	[1, 1024, 384]	768	768
Mlp (mlp): 3-18	--	[1, 1024, 384]	[1, 1024, 384]	1,181,568	1,181,568
DropPath (drop path): 3-19	--	[1, 1024, 384]	[1, 1024, 384]	--	--
SwinBlock (3): 2-6	--	[1, 1024, 384]	[1, 1024, 384]	--	--
LayerNorm (norm1): 3-20	--	[1, 1024, 384]	[1, 1024, 384]	768	768
WindowAttention (attn): 3-21	--	[4, 256, 384]	[4, 256, 384]	592,332	257,169,408
DropPath (drop path): 3-22	--	[1, 1024, 384]	[1, 1024, 384]	--	--
LayerNorm (norm2): 3-23	--	[1, 1024, 384]	[1, 1024, 384]	768	768
Mlp (mlp): 3-24	--	[1, 1024, 384]	[1, 1024, 384]	1,181,568	1,181,568
DropPath (drop path): 3-25	--	[1, 1024, 384]	[1, 1024, 384]	--	--
SwinBlock (4): 2-7	--	[1, 1024, 384]	[1, 256, 768]	--	--
PatchMerge (downsample): 3-26	--	[1, 1024, 384]	[1, 256, 768]	1,181,184	302,187,264
LayerNorm (norm1): 3-27	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-28	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-29	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-30	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-31	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-32	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (5): 2-8	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-33	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-34	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-35	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-36	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-37	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432

DropPath (drop path): 3-38	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (6): 2-9	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-39	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-40	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-41	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-42	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-43	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-44	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (7): 2-10	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-45	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-46	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-47	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-48	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-49	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-50	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (8): 2-11	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-51	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-52	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-53	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-54	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-55	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-56	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (9): 2-12	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-57	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-58	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-59	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-60	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-61	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-62	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (10): 2-13	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-63	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-64	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-65	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-66	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-67	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-68	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (11): 2-14	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-69	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-70	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-71	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-72	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-73	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-74	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (12): 2-15	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-75	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-76	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-77	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-78	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-79	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-80	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (13): 2-16	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-81	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536

WindowAttention (attn): 3-82	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-83	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-84	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-85	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-86	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (14): 2-17	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-87	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-88	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-89	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-90	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-91	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-92	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (15): 2-18	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-93	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-94	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-95	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-96	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-97	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-98	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (16): 2-19	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-99	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-100	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-101	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-102	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-103	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-104	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (17): 2-20	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-105	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-106	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-107	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-108	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-109	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-110	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (18): 2-21	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-111	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-112	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-113	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-114	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-115	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-116	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (19): 2-22	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-117	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-118	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-119	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-120	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-121	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-122	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (20): 2-23	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-123	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-124	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-125	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-126	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536

Mlp (mlp): 3-127	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-128	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (21): 2-24	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm1): 3-129	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
WindowAttention (attn): 3-130	--	[1, 256, 768]	[1, 256, 768]	2,364,120	117,181,440
DropPath (drop path): 3-131	--	[1, 256, 768]	[1, 256, 768]	--	--
LayerNorm (norm2): 3-132	--	[1, 256, 768]	[1, 256, 768]	1,536	1,536
Mlp (mlp): 3-133	--	[1, 256, 768]	[1, 256, 768]	4,722,432	4,722,432
DropPath (drop path): 3-134	--	[1, 256, 768]	[1, 256, 768]	--	--
SwinBlock (22): 2-25	--	[1, 256, 768]	[1, 64, 1536]	--	--
PatchMerge (downsample): 3-135	--	[1, 256, 768]	[1, 64, 1536]	4,721,664	302,089,728
LayerNorm (norm1): 3-136	--	[1, 64, 1536]	[1, 64, 1536]	3,072	3,072
WindowAttention (attn): 3-137	--	[1, 64, 1536]	[1, 64, 1536]	9,446,640	23,009,280
DropPath (drop path): 3-138	--	[1, 64, 1536]	[1, 64, 1536]	--	--
LayerNorm (norm2): 3-139	--	[1, 64, 1536]	[1, 64, 1536]	3,072	3,072
Mlp (mlp): 3-140	--	[1, 64, 1536]	[1, 64, 1536]	18,882,048	18,882,048
DropPath (drop path): 3-141	--	[1, 64, 1536]	[1, 64, 1536]	--	--
SwinBlock (23): 2-26	--	[1, 64, 1536]	[1, 64, 1536]	--	--
LayerNorm (norm1): 3-142	--	[1, 64, 1536]	[1, 64, 1536]	3,072	3,072
WindowAttention (attn): 3-143	--	[1, 64, 1536]	[1, 64, 1536]	9,446,640	23,009,280
DropPath (drop path): 3-144	--	[1, 64, 1536]	[1, 64, 1536]	--	--
LayerNorm (norm2): 3-145	--	[1, 64, 1536]	[1, 64, 1536]	3,072	3,072
Mlp (mlp): 3-146	--	[1, 64, 1536]	[1, 64, 1536]	18,882,048	18,882,048
DropPath (drop path): 3-147	--	[1, 64, 1536]	[1, 64, 1536]	--	--
LayerNorm (fc norm): 1-3	--	[1, 1536]	[1, 1536]	3,072	3,072
Linear (head): 1-4	--	[1, 1536]	[1, 8]	2,296	12,296

6 Data-driven semi-supervised machine learning with safety indicators for abnormal driving behaviour detection

Abstract

Detecting abnormal driving behaviour is critical for road traffic safety and the evaluation of drivers' behaviour. With the advancement of machine learning (ML) algorithms and the accumulation of naturalistic driving data, many ML models have been adopted for abnormal driving behaviour detection (also referred to as anomalies). Most existing ML-based detectors rely on supervised methods, which require substantial labelled data. However, ground truth labels are not always available in the real world, and labelling large amounts of data is tedious. Thus, there is a need to explore unsupervised or semi-supervised methods to make the anomaly detection process more feasible and efficient. To fill this research gap, this study analyses large-scale real-world data revealing several abnormal driving behaviours (e.g., sudden acceleration, rapid lane-changing) and develops a Hierarchical Extreme Learning Machine (HELM)-based semi-supervised ML method using partly labelled data to accurately detect the identified abnormal driving behaviours. Moreover, previous ML-based approaches predominantly utilised basic vehicle motion features (e.g., velocity and acceleration) to label and detect abnormal driving behaviours, while this study seeks to introduce event-level safety indicators as input features for ML models to improve detection performance. Results from extensive experiments demonstrate the effectiveness of the proposed semi-supervised ML model with the introduced safety indicators serving as important features. The proposed semi-supervised ML method outperforms other baseline semi-supervised or unsupervised methods regarding various metrics, e.g., delivering the best accuracy (99.58%) and the best F1-score (0.9913). The ablation study further highlights the significance of safety indicators for advancing the detection performance.

This chapter is based on the journal publication:

Dong, Y., Zhang, L., Farah, H., Zgonnikov, A., & Van Arem, B. (2025). Data-driven Semi-supervised Machine Learning with Safety Indicators for Abnormal Driving Behavior Detection. Transportation Research Record: Journal of the Transportation Research Board, 1-16. https://doi.org/10.1177/03611981241306752

6.1 Introduction

Road traffic safety has become a growing concern worldwide. The World Health Organization (2023) reported that approximately 1.19 million people die each year in road traffic crashes, with over 30 million suffering non-fatal injuries. These crashes not only resulted in disabilities but also caused significant economic loss, reaching as high as 3% of the gross domestic product in some countries. It is alarming that in most crashes, human factors were identified as contributing factors (Bucsuházy et al., 2020; Elvik et al., 2009; Saiprasert & Pattara-Atikom, 2013). This highlights the urgent need to identify abnormal driving behaviours and find ways to prevent or mitigate crashes caused by such abnormal human driving behaviours.

Driving behaviour encompasses various variables and factors, including driving performance, environmental awareness, risk-taking propensity, and reasoning abilities (Mohammadnazar et al., 2021). Abnormal driving behaviour refers to reckless actions that deviate from safe and normal driving, posing risks to oneself, passengers, and other road users, and typically occurs within a short period of time (Ma et al., 2023). Examples of such behaviour include excessive speeding, tailgating, and erratic lane changes (Matousek et al., 2019). These abnormal driving behaviours frequently engender severe traffic altercations, including collisions, crashes, and other minor incidents, thereby underscoring the necessity of addressing and precluding these actions (Ma et al., 2023; Matousek et al., 2019). Effective monitoring of abnormal driving behaviours is integral to augmenting driving safety, enhancing driver awareness of driving patterns, and reducing the chances of prospective road crashes.

Machine learning (ML)-based approaches have shown great promise in detecting abnormal driving behaviours. They can learn complex patterns, adapt to changing scenarios, handle large and diverse datasets, and detect unusual behaviours with optimised processes (Sarker, 2021). However, most of the available studies adopted fully-supervised ML models to do the detection, and few of them explored unsupervised or semi-supervised ML methods. While in the real world, ground truth labels are sometimes missing or inaccurate, plus labelling large amounts of data is tedious and even dangerous under certain critical situations. Therefore, examining and developing unsupervised or semi-supervised methods is imperative to achieve more feasible and efficient abnormal driving behaviour detection.

On the other hand, safety indicators and particularly Surrogate Measures of Safety (SMoS) offer a proactive approach to safety evaluation by using proximity measures. Since SMoS do not rely directly on crash data, employing them allows road safety assessment without the need to collect crash data (Nikolaou et al., 2023). As Tarko (2018) notes, SMoS facilitates detecting excessive crash risk, better understanding crash-precipitating conditions, and estimating countermeasure efficacy. By providing insights into potential safety issues, the safety indicators help prioritise improvement efforts. C. Wang et al. (2021) categorise safety indicators into three classes: time-based (e.g., time-to-collision (TTC) and post-encroachment time (PET)), deceleration-based (e.g., deceleration rate to avoid a crash (DRAC)), and energy-based (e.g., DeltaV). Commonly, these safety indicators are applied in road safety research in combination with thresholds to identify traffic conflicts (Bonela & Kadali, 2022; C. Lu et al., 2021; Nikolaou et al., 2023). There is no doubt that safety indicators can serve as important features in various tasks, e.g., in traffic safety assessment and in detecting traffic conflicts, however, for data-driven-based abnormal driving behaviour detection, previous studies predominantly employed

basic vehicle motion (e.g., speed, acceleration) as features to label and detect abnormal behaviours, and seldom explored the benefits of safety indicators.

To fill the aforementioned research gaps, this study aims to develop a data-driven approach for abnormal driving behaviour detection using real-world naturalistic driving data and leveraging semi-supervised ML with self-supervised training to enhance the performance and effectiveness of the detection method. Specifically, this study first analyses a large-scale dataset, i.e., the CitySim dataset (Zheng et al., 2022), with vivid visualisations, and extracts various abnormal driving behaviours. Then, the study develops a Hierarchical Extreme Learning Machine (HELM)-based semi-supervised ML model using unlabelled data to carry out self-supervised pre-training and leveraging only partly labelled data to fine-tune the model for accurately detecting the identified abnormal driving behaviours. Furthermore, this study conducts a significative ablation study introducing event-level safety indicators as input features for the developed semi-supervised ML model to further improve the detection performance. Extensive experiments verified the proposed method. The proposed semi-supervised HELM model using safety indicators as input features outperforms other baseline models, delivering the best accuracy at 99.58% and the best F1-score at 0.9913.

In short, by filling the research gap and addressing the limitations of existing methods in the literature, this research endeavours to improve road safety and reduce accidents caused by abnormal driving behaviours. It addresses the limitations of traditional supervised approaches and overcomes the scarcity of labelled abnormal driving data. The study analyses publicly available vehicle trajectory datasets and provides meaningful insights into the identification of abnormal human driving behaviour. The conclusions and limitations of this study, as well as future research directions, are discussed at the end of this study.

6.2 Related work

Several studies have investigated abnormal driving behaviours, with typical examples of Chen et al. (2015) and Kim et al. (2016) putting forth definitions reflecting different conceptualisations of driving, as shown in **Table 6-1**.

Table 6-1. Different classifications of abnormal driving behaviour

Chen et al. (2015)	Kim et al. (2016)
Weaving	Sudden start
Swerving	Speeding
Sideslipping	Long-standing speeding
Fast U-turn	Sudden braking
Turning with a wide radius	Sudden overtaking
Sudden braking.	Sudden changing lanes
---	Sudden turning

Chen et al. (2015) emphasised whether the vehicle's location complies with regulations, while Kim et al. (2016) prioritised speed modulation. In combination, despite these different emphases, both studies suggest that sudden changes in speed or location are key indicators of abnormal driving, regardless of the country where the driving occurs. Building on this, the current study delineates abnormal driving based on changes in position and velocity,

concentrating on behaviours of abrupt starts, emergency braking, as well as rapid and close lane changes. This definition is supported by a comprehensive review of the existing literature, indicating a focus on both the spatial and temporal aspects of driving behaviour.

ML-based approaches for detecting abnormal driving behaviours have gained substantial research attention and exhibited robust performance. Both supervised and unsupervised methodologies have been commonly utilised in prior investigations of abnormal driving behaviour. Supervised ML techniques necessitate labelled data during model training, whereby the system ascertains the mapping between inputs and outputs to categorise and predict new data points. For example, Jia et al. (2020) devised a model integrating long short-term memory (LSTM) neural network and convolutional neural network (CNN) architectures to pinpoint instances of extreme acceleration and deceleration. Shahverdy et al. (2021) proposed a lightweight 1-dimensional CNN (1D-CNN) exhibiting high efficiency and low computational overhead for classifying drivers' behaviour into safe, distracted, aggressive, drunk, and drowsy driving. Ryan et al. (2021) simulated an end-to-end model leveraging CNN to compare human and autonomous vehicle driving patterns and adopted a Gaussian Processes-based method to detect driving anomalies.

Conversely, unsupervised ML techniques entail training models using raw, unlabelled data. This approach is frequently utilised during exploratory phases to derive insights from the dataset. As an illustration, Mohammadnazar et al. (2021) developed an architecture leveraging unsupervised ML to quantify driving performance and categorise driving styles across diverse spatial contexts. Feng et al. (2019) proposed a Support Vector Clustering methodology to classify driving styles (e.g., aggressive, normal, defensive) robustly. Existing literature denotes substantial challenges in accurately identifying anomalies through solely unsupervised ML. As Chandola et al. (2009) concluded from their review, unsupervised anomaly detection approaches often demonstrate inferior detection rates and heightened false positive rates on real-world problems. Correspondingly, Pimentel et al. (2014) found via benchmark assessments that complete dependency on unsupervised anomaly detection is not recommended, as these techniques fail to detect all anomalies. Erfani et al. (2016) further emphasised that purely unsupervised methodologies lack the learning guidance to precisely differentiate normal from abnormal patterns. Synthesising these conclusions, utilising only unsupervised ML without any labelled data to achieve accurate anomaly detection is hardly possible. Even if viable, the detection performance based on pure unsupervised ML is highly possible to be further enhanced by labelled data. Therefore, there is a research consensus regarding the necessity for making use of at least partially labelled data to supervise and augment anomaly detection capabilities with semi-supervised ML approaches.

Concerning the features utilised as input for ML models, traditional indicators such as velocity, acceleration, and steering angle have been extensively employed (Dai et al., 2010; Dhar et al., 2014; Jia et al., 2020; Li et al., 2015; Lim & Yang, 2016). For example, Lim & Yang (2016) considered vehicular data comprising velocity, acceleration, steering angle, and gas pedal position and leveraged a CNN model to estimate driver drowsiness, workload, and distraction levels. Li et al. (2015) collected lateral vehicle position, steering angle, and speed-related information and implemented a Support Vector Machine (SVM) model to differentiate between normal and intoxicated driving states. Incorporating safety indicators (e.g., TTC) into ML-based methods is supposed to be promising for abnormal driving detection but has seldom been investigated yet. To the best of the authors' knowledge and after extensive review, only one

study was found to be relevant, i.e., J. Lu et al. (2022) integrated the representation of TTC together with the driver manoeuvre profiles into a deep unsupervised learning and clustering method with their proposed Transformer encoder based model to identify traffic conflicts and non-conflicts. However, they only investigated situations of one intersection and one roundabout in the United States, neglecting other various types of driving anomalies, especially those related to highway driving.

Investigating the potential of semi-supervised approaches, which utilise both labelled and unlabelled data, is imperative to enhance abnormal driving behaviour detection, yet limited research has explored this direction. By harnessing the additional information from unlabelled data, semi-supervised learning might be able to uncover subtle patterns and behaviours that conventional supervised or unsupervised techniques may overlook. This study endeavours to address this research gap. Moreover, input features are fundamental for ML-based approaches. To enhance detection performance, it is advisable to explore more effective features. In this line, this study seeks to investigate the benefits of event-level safety indicators as input variables and conducts ablation analyses to verify their efficacy in upgrading the detection accuracy.

6.3 Dataset and data analysis

6.3.1 Description of the data

To conduct data-driven research, a high-quality dataset is imperative. After extensive exploration, the study utilises the CitySim dataset (Zheng et al., 2022), comprising video-based trajectory data concentrating on traffic safety in the United States. The CitySim dataset encompasses vehicle trajectory information extracted from videos at 30 frames per second (FPS) captured by 12 drones, spanning six road geometry typologies, including freeway segments, signalised intersections, and stop-controlled junctions. The dataset provides precise positional details with measurements accurate to approximately 10 centimetres in various formats, including pixels, feet, and GPS coordinates, alongside data on velocity, heading angle, and vehicle lane numbers. **Table 6-2** provides the fields of the raw data record and provides one example accessible within the dataset.

Table 6-2. Data sample of the CitySim dataset

Features	Value
frameNum	0
carId	582
carCenterX (ft)	462.4
carCenterY (ft)	184.8
headX (ft)	469.6
headY (ft)	184.8
tailX (ft)	455.3
tailY (ft)	184.8
Speed (mph)	39.5
Heading (°)	180.7
laneId	10

Table notation: ft---feet; mph---miles per hour; ° --- degree

Following the research objectives, supplementary features were derived from the CitySim dataset, encompassing, for example, *longitudinal acceleration*, *lateral acceleration*, and *inter-vehicle distances*, which facilitate the calculation of event-level safety indicators. By integrating these computed variables with the original dataset, this study endeavours to strengthen the data foundations necessary for the model. However, the dataset initially still contains noisy and inconsistent data. Rigorous pre-processing techniques were employed to enhance the quality and reliability, ensuring robustness in subsequent analysis and model training. Firstly, entries with missing values and *NULL* were identified and treated using the *dropna* function in the Python *pandas* library, eliminating instances with incomplete information. Then, entries with extreme values such as distance or speed beyond the normal range were cleared. For example, negative values in either distance or speed and speed values beyond 100 m/s (360km/h) are considered extreme values.

Furthermore, a data smoothing technique with exponential smoothing was applied to attenuate high-frequency noise while preserving the underlying trends and patterns of the data.

Table 6-3 exhibits examples of the data used after the pre-processing. As illustrated, the data after pre-processing includes features of coordinates, i.e., carCenterX and carCenterY, speed, heading angle, and distance. Since carCenterX, carCenterY, speed, and heading angle are provided in the original data, they were the fields used when smoothing the data. Whereas, the data fields of distance together with the later introduced longitudinal and lateral acceleration were calculated after the pre-processing using the relevant fields. For example, distance was calculated using carCenterX and carCenterY of the adjacent two vehicles.

The upcoming *Section 6.4 Methodology* delineates the precise calculations done to derive the additional features from the raw CitySim dataset, including, as well, the selected event-level safety indicators.

Table 6-3. Data examples after data pre-processing

frameNum	carCenterX (m)	carCenterY (m)	Speed (m/*s*)	Heading (°)	2DTTC (*s*)	Distance (m)	Abnormal=1/ Normal=0
10	53.258	32.155	14.985	359.632	1.110	0.482	1
1737	251.998	27.466	11.095	359.742	104.794	131.453	0
1739	248.537	31.095	12.300	359.707	6001.553	128.168	0
1760	251.607	27.355	11.064	359.656	110.943	131.392	0
11940	128.567	31.653	16.368	359.220	0.415	0.593	1
11966	127.897	31.653	16.217	359.082	0.376	0.482	1
11981	127.115	31.653	16.218	358.865	0.295	0.457	1
12000	126.836	31.542	16.277	358.864	0.311	0.387	1

Table notation: the original distance measure "feet" is converted to "meters".

6.3.2 Abnormal driving behaviours identified in the dataset

Based on the classification and definition of abnormal driving behaviour in the reviewed literature (check *Section 6.2 Related work*), this section illustrates the specific abnormal driving behaviours observed in the examined CitySim dataset. Each abnormal behaviour is associated with one or two indicators, measured or calculated at various locations.

Rapid acceleration and emergency brake behaviour

The acceleration data corresponding to each velocity datum in the vehicle trajectory dataset is exhibited in **Figure 6-1**. Extreme acceleration and deceleration observations can be derived, denoting abnormal manoeuvres such as sudden braking or accelerating. Identifying these extreme observations enables the segmentation of abnormal driving behaviours versus normal ones. A specific proportion of extreme acceleration can be pinpointed by statistically scrutinising all acceleration observations at identical speeds across all journeys. Determining an appropriate ratio to differentiate extreme/abnormal points from normal ones is imperative. A 15% threshold appears sensible based on reiterative experimentation and associated existing research (Jia et al., 2020; X. Wang et al., 2015).

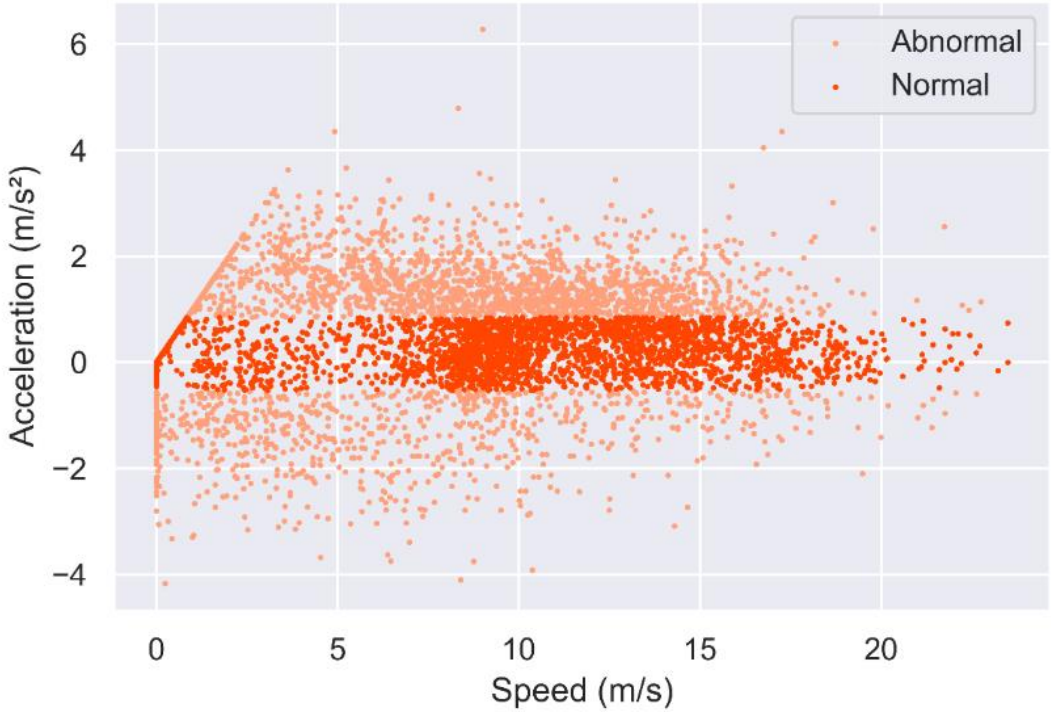


Figure 6-1. Longitudinal acceleration and deceleration scatterplot at different speeds: normal observations (dark orange dots in the middle), and abnormal observations (light orange dots)

Rapid lane-changing behaviour

Rapid lane-changing behaviour is characterised by sudden and instantaneous abnormal lateral accelerations that occur for a short duration. In normal driving patterns, vehicles exhibit relatively stable lateral acceleration around zero (as shown in **Figure 6-2**). However, abnormal lane-changing behaviour manifests an abrupt variation in the vehicle's lateral acceleration.

The majority of vehicles exhibiting lane divergence comportment demonstrate a lateral acceleration bounded by ±1 m/s^2, whereby they execute lane diversions seamlessly at a fixed velocity. However, the accelerations of some vehicles appear as outliers in **Figure 6-3**. A normal distribution with a mean of 0 and a standard deviation of 1.3 was examined. According to the characteristics of a normal distribution, approximately 68% of the data falls within ±1 standard deviation from the mean. These outliers beyond ±1 standard deviation from the mean accounted for approximately 32% of the total data points. A ratio of approximately 15% is considered reasonable based on repeated experiments and related research (Jia et al., 2020; X. Wang et al.,

2015). This satisfies the heuristic definition of outliers as observations that differ significantly from most data. Examining outliers based on standard deviation thresholds aligns with statistically grounded techniques for anomaly detection using the sigma principle for normal distributions (Jia et al., 2020; X. Wang et al., 2015). According to the normal distribution, values greater than 1.3 m/s² and less than -1.3m/s² were used as the filter condition for abnormal instances.

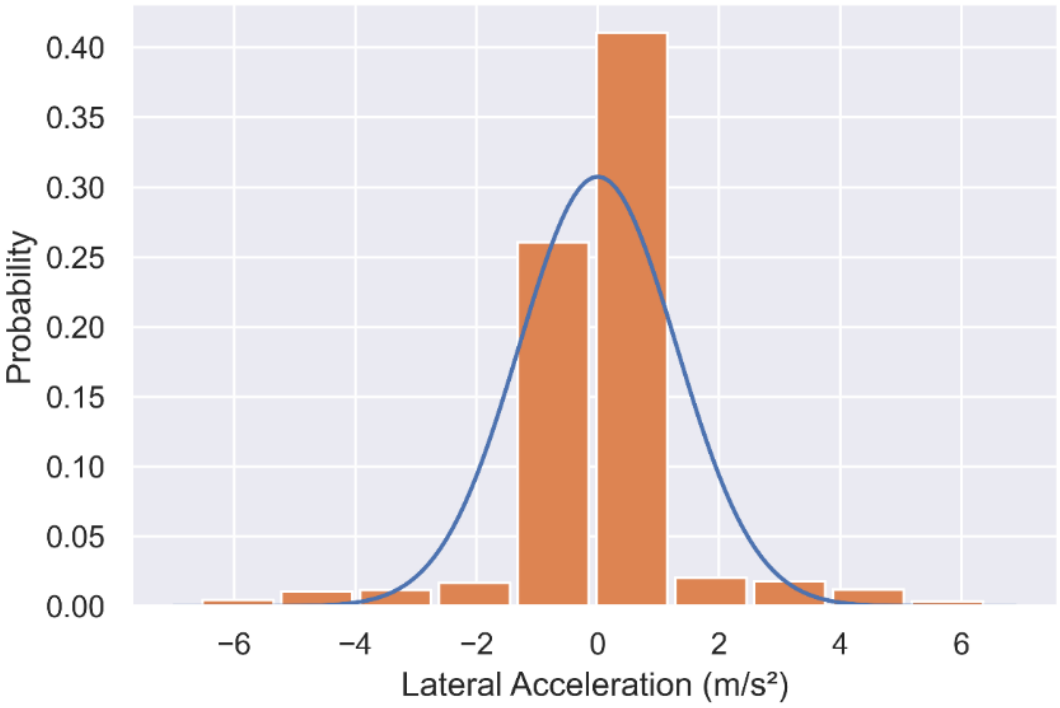


Figure 6-2. Illustration of the distribution of lateral acceleration

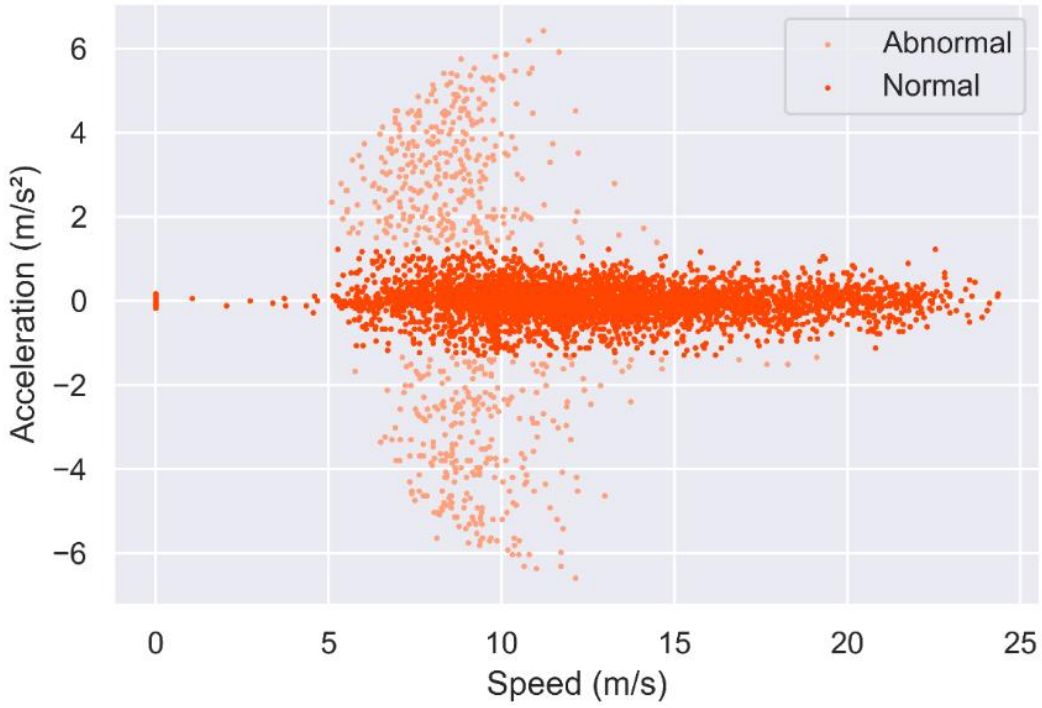


Figure 6-3. Lateral acceleration scatterplot at different speeds: normal observations (dark orange dots in the middle), and abnormal observations (light orange dots)

Close lane-changing behaviour

Close lane-changing behaviour is characterised by sudden and instantaneous abnormal lane-changing actions with very short distances from adjacent vehicles that occur for a short duration. Vehicles in normal driving patterns maintain a certain distance between themselves and adjacent lanes. However, during abnormal close lane-changing behaviour, there is a significant decrease in the distance between the vehicle and vehicles in the adjacent lanes, indicating a close lane change. In this study, when the distance between the car performing the lane-changing manoeuvre and its surrounding vehicles is less than 0.5 meters, it is considered severe abnormal driving behaviour. In contrast, when the distance is less than 1.0 meters but greater than 0.5 meters, it is considered weak abnormal driving behaviour, as seen in **Figure 6-4**.

Based on the aforementioned criteria, the labels of the driving data samples were further examined by human experts to remove inaccurate labelling, improving the quality of the finalised labels. Referring to the method adopted by Jia et al. (2020), firstly, the data samples during the periods with large longitudinal accelerations and decelerations, large lateral accelerations, and extreme close distances were selected to be checked and verified, i.e., grey area data. By observing the changes in the distribution of the extreme longitudinal acceleration and deceleration data points, lateral accelerations, as well as distance-changing dynamics, the human expert combined these observations with their knowledge and experience to verify the labels. If the human expert was not certain with high confidence about the labelling for the data sample, that specific data sample would be removed. It should be noted that this human expert examination-based verification method may only correct the labels of false alarms, to the degree it is possible through examining kinematic variables, and will not correct missed abnormal instances.

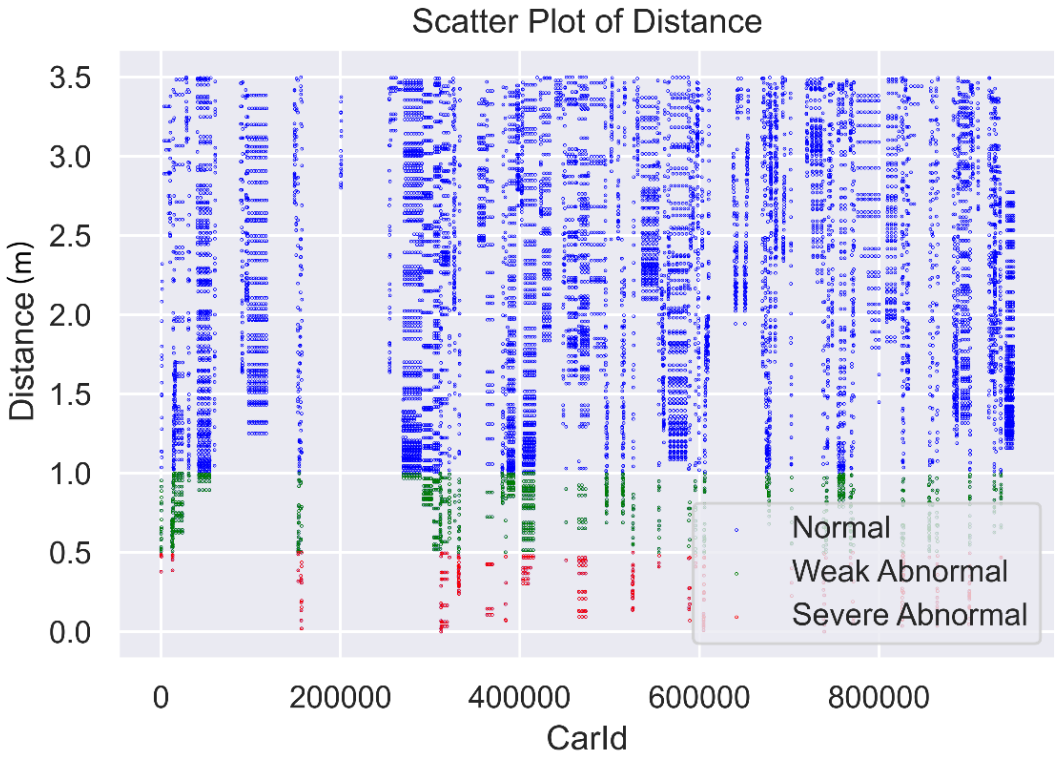


Figure 6-4. Scatterplot of distance during lane-changing for different *carId*

6.4 Methodology

This section first introduces event-level safety indicators, especially the adopted Two-Dimensional Time-To-Collision (2D-TTC) (Jiao, 2024; Ward et al., 2015). Two ML models, i.e., Isolation Forest and Robust Covariance, are then presented as baseline methods for comparison. Finally, a customised semi-supervised model named Hierarchical Extreme Learning Machine (HELM) is proposed and explained in detail.

6.4.1 Safety indicators

In the literature, several safety indicators were developed and introduced, and a comprehensive overview can be found in (Arun et al., 2021; Nikolaou et al., 2023). One of the most popular and commonly used safety indicators is the time-to-collision (TTC) which is a time-based proximity measure. TTC is defined as the time required for two road users, on a collision course, to collide if no evasive action is taken, which can be and is generally computed continuously (Svensson, 1998). Its simplistic form is when road users' speed and path are assumed to remain unchanged (Hayward, 1971). For example, the TTC value, for a car-following situation, assuming motion prediction with constant speed, is calculated as:

$$\text{TTC} = \frac{D}{v_1 - v_2} \tag{6-1}$$

where D is the distance between the following vehicle and the leading vehicle, while v_1 and v_2 are the speeds for the two vehicles, respectively.

Over the years, several studies have further extended the TTC safety indicator. For example, Time Exposed Time-to-collision (TET) and Time Integrated Time-to-collision (TIT) were introduced by (Minderhoud & Bovy, 2001) to measure the risk associated with the duration of dangerous driving conditions. The Modified Time-to-Collision (MTTC), proposed by Ozbay et al. (2008), provides an alternative way to calculate TTC at each instant, e.g., in a car-following traffic scenario, by considering the accelerations of both the lead and following vehicles. Other approaches involve incorporating site-specific motion patterns of road users and calculating TTC with respect to the distribution of possible trajectories (Saunier et al., 2007; St-Aubin et al., 2015). In this study, a TTC-based safety indicator, entitled the two-dimensional TTC (2D-TTC) (Jiao, 2024; Ward et al., 2015), was implemented as an input feature, which can capture proximities of vehicles' movements and interactions in a plane in various traffic scenarios besides a car following scenario. The illustration of 2D-TTC is demonstrated in **Figure 6-5**. 2D-TTC is calculated as follows:

$$\text{2D-TTC} = \begin{cases} \frac{|\overrightarrow{DTC}|}{|v_i - v_j|}, & if\ the\ direction\ of\ \overrightarrow{\boldsymbol{DTC}}\ is\ the\ same\ as\ \boldsymbol{v}_{ij} = (\boldsymbol{v}_i - \boldsymbol{v}_j) \\ inf,\ if\ \left|\overrightarrow{\boldsymbol{DTC}}\right| = inf\ OR\ if\ the\ direction\ of\ \overrightarrow{\boldsymbol{DTC}}\ is\ opposite\ to\ \boldsymbol{v}_{ij} = (\boldsymbol{v}_i - \boldsymbol{v}_j) \end{cases} \tag{6-2}$$

where $\left|\overrightarrow{\boldsymbol{DTC}}\right|$ is the distance-to-collision, which refers to the minimum distance between the bounding boxes of target vehicle i and another interacting vehicle j along their relative speed $\boldsymbol{v}_{ij} = (\boldsymbol{v}_i - \boldsymbol{v}_j)$ direction, while $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are the speeds for the two vehicles respectively. If their relative movement decreases the distance-to-collision, they are approaching each other, and a potential collision exists. Otherwise, the vehicles are moving away from each other and

no potential collision exists. For more detailed information about the demonstration and calculation of the adopted 2D-TTC, the reader is advised to refer to (Jiao, 2024).

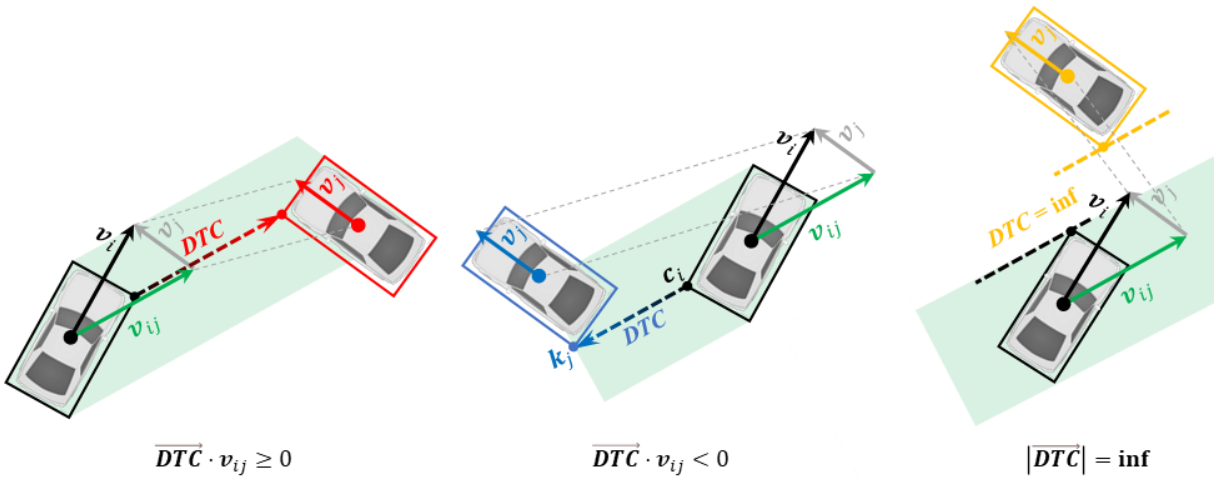


Figure 6-5. Illustration of 2D-TTC (adjusted from (Jiao, 2024))

In general, according to the literature, only encounters with a minimum TTC below 1.5 seconds are deemed critical, with trained observers consistently applying this threshold in practice (van der Horst & Hogema, 1994). This study explores the effects of the input feature 2D-TTC on the detection performance of abnormal driving behaviour. Specifically, the vehicle angle in the dataset decomposes each vehicle's velocity into *x*-*y* coordinate components, yielding velocity vectors based on the dataset parameters. 2D-TTC is then calculated per these velocity vectors and the corresponding distance along the same direction. This approach highlights how 2D-TTC can be computed from the raw dataset by leveraging the vehicle angle data to obtain velocity vectors in coordinate space. The derived 2D-TTC is analysed and integrated with input features such as position, speed, and acceleration to evaluate abnormal driving behaviour detection performance using the given dataset.

6.4.2 Baseline models

Isolation Forest and Robust Covariance are selected as two baseline methods considering their interpretability, effectiveness, and broad utilisation in various domains.

The *Isolation Forest*, initially developed by (Liu et al., 2008), constitutes an effective algorithm typically utilised for data anomaly detection. The Isolation Forest algorithm is based on the principle that anomalous data points are more readily separable from the majority of normal samples. To isolate an abnormal data point, the algorithm iteratively generates partitions of the sample by randomly selecting a feature attribute and subsequently randomly choosing a split value within the permissible minimum and maximum values for the selected feature attribute. Through recursive binary partitioning, data points that require fewer splits to become isolated are deemed more anomalous.

The Isolation Forest algorithm capitalises on the premise that anomalies are few and different from the rest of the data, and thereby manifest topological shorter path lengths from the root to the external node (leaf), (which is elucidated by averaging this value across the trees) when random partitioning is employed. Therefore, it leverages an ensemble of isolation trees

generated through such recursive random partitioning to identify anomalies, with shorter average path lengths corresponding to greater anomaly scores.

In practice, the Isolation Forest anomaly detection algorithm involves two primary phases. Firstly, a collection of isolation trees (*iTrees*) is constructed utilising recursive partitioning on a training dataset. During recursive partitioning, splits are performed by randomly selecting an attribute and random split value to isolate a data point. Secondly, each instance in the test set is propagated through the ensemble of *iTrees* and assigned an *anomaly score* based on the average path length for that instance across the *iTrees*. Shorter average path lengths correspond to fewer partitions required to isolate the instance, indicating more anomalous behaviour and higher *anomaly scores*. After computing *anomaly scores* for all test instances, those data points with a score exceeding a predefined threshold specific to the domain can be classified as anomalies.

The *Robust Covariance* estimation algorithm presupposes that normal data points exhibit a Gaussian distribution, and accordingly approximates the morphology of the joint distribution (namely, estimates the mean and covariance of the multivariate Gaussian distribution) (Nikita Butakov, 2020).

In statistical analysis, the deviation can be measured by the Z-score. The generalisation of the Z-score for a point x_i in the case of a p-dimensional multi-variate probability distribution with some mean μ and covariance matrix Σ is known as Mahalanobis distance d_i, which is given by:

$$d_i = \sqrt{(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)} \tag{6-3}$$

It is based on the premise that outliers increase the values (entries) in Σ, thereby making the data dispersion appear more extensive. Consequently, |Σ| (the determinant) will also be larger, which could theoretically decrease if extreme samples are removed. Rousseeuw and Van Driessen (1999) devised a computationally efficient algorithm capable of furnishing robust covariance approximations. The approach assumes that at minimum h of the n samples are "normal" (h denoting a hyperparameter). The algorithm begins with k arbitrary samples containing (p+1) points. For each k sample, μ, Σ, and |Σ| are estimated, the distances are computed and sorted in ascending order, and the smallest h distances are employed to update the estimates. In their original publication, the process of computing distances and revising the estimations of μ, Σ, and |Σ| is entitled a "C-step" whereby two such increments are typically sufficient to identify effective candidates (for μ and Σ) among the k arbitrary samples. In the succeeding step, a subset of magnitude m with the lowest |Σ| (the optimal candidates) is contemplated for computation until convergence, and the sole estimate whose |Σ| is minimal is furnished as output.

Please note that, although Isolation Forest and Robust Covariance are usually considered unsupervised ML approaches, in this study, only normal data samples are input to train them; thus, in this study, they can be regarded as semi-supervised approaches and are comparable to the proposed semi-supervised machine learning method.

6.4.3 Hierarchical extreme learning machine-based semi-supervised machine learning

The Hierarchical Extreme Learning Machine algorithm, originally proposed by Tang et al. (2016), constitutes an advanced extension of the Extreme Learning Machines (ELM) algorithm that can enhance performance in both training speed and generalisation capability. This

approach integrates a feed-forward neural network structure with multiple latent layers, and it operates through two primary steps: unsupervised feature representation and supervised feature classification. In the initial step, the HELM is intended to ascertain a sparse encoder in an unsupervised manner, which transforms the raw input into superior-level representation. The encoder is structured with multiple latent layers which are processed sequentially, with each layer building upon the previous one to capture increasingly abstract features of the data. The second step involves using these learned features for supervised classification or approximation tasks. By leveraging the rich, hierarchical features extracted in the first step, HELM aims to achieve effective and accurate predictions. This two-step process enables HELM to combine the advantages of both unsupervised and supervised learning, resulting in improved overall performance.

Given a training set with N samples, indicated by (X_i, Y_i) $(X_i \in R^n, Y_i \in R^t, i = 1,2,3, \dots, N)$, where X_i and Y_i denote the feature representation and the targeted output of the ith sample, respectively. Suppose the encoder consists of K hidden layers, each with $L_i (1 \leq i \leq K)$ neurons. The output $O = [o_1, o_2, \dots, o_N]^T$ can be expressed as:

$$\sum_{i=1}^{K} \beta_i g(W_i \cdot x_j + b_j) = o_j, \ j = 1,2, \dots, N \tag{6-4}$$

where $g(\cdot)$ is the activation function, β_i is the output weight, W_i is the input weight, and b_j is the bias. Ideally, there should be:

$$\sum_{j=1}^{N} \|o_j - Y_j\| = 0 \tag{6-5}$$

This implies that there exist weights β_i, W_i, and biases b_i such that

$$\sum_{i=1}^{K} \beta_i g(W_i \cdot x_j + b_j) = Y_j, \text{ for } j = 1,2, \dots, N \tag{6-6}$$

In matrix form, this can be represented by

$$H\beta = Y \tag{6-7}$$

where H is the output of the hidden layer node, β is the output weight, and Y is the desired output.

$$H(W_1, W_2, \dots, W_K, b_1, b_2, \dots, b_K, x_1, x_2, \dots, x_N) = \begin{bmatrix} g_1(X_1) & \cdots & g_{K_1}(X_1) \\ \vdots & \ddots & \vdots \\ g_1(X_N) & \cdots & g_{K_1}(X_N) \end{bmatrix} \tag{6-8}$$

To train a single hidden layer ELM neural network is equivalent to obtaining $\hat{\beta}$ such that

$$\|H\hat{\beta} - Y\| = \min_{\beta} \|H\beta - Y\| \tag{6-9}$$

When choosing the mean square error (MSE) as the measure, this formula is equivalent to minimising the following loss function:

$$Loss = \sum_{j=1}^{N} (\sum_{i=1}^{K} \beta_i g(W_i \cdot x_j + b_i) - Y_j)^2 \tag{6-10}$$

Traditional ELMs allow the weights β and the deviations between the latent layers and the inputs to be set arbitrarily, drawn from any distribution. This flexibility means that the learning process primarily adjusts these weights to find the optimal connections between the latent layers

and the output. However, standard ELMs can be limited in their ability to effectively process complex data, even with a large number of hidden nodes.

In this study, the customised HELM was introduced to address this limitation, which stacks multiple layers of ELM to create a deeper and more profound structure. This hierarchical approach enhances the model's ability to capture intricate data patterns. The proposed HELM-based semi-supervised learning consists of two phases, i.e., 1) self-supervised training for feature learning, where the model extracts and learns useful features from the data in an unsupervised manner, and 2) supervised fine-tuning, where the model is further optimised using samples of labelled data to improve its performance on the abnormal driving behaviour detection task, as visualised in **Figure 6-6**.

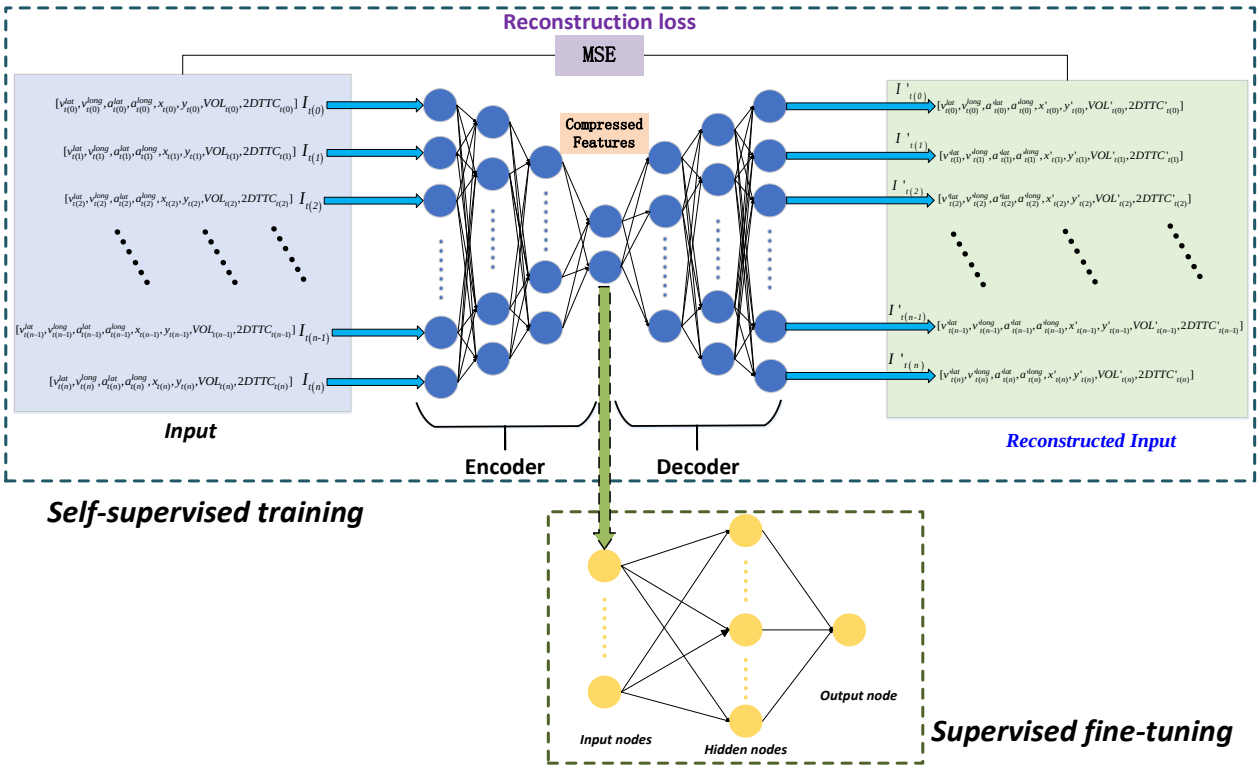


Figure 6-6. The framework of HELM-based semi-supervised machine learning method

The HELM model is initially trained purely self-supervised on normal data samples exclusively, with all anomalous examples excluded from this training set. During this phase, by minimising a reconstruction error loss function, the stacked ELM autoencoder layers learn to capture the most salient features of the input data that represent its intrinsic normal characteristics. These extracted feature representations can encapsulate the essential properties of standard normal behaviour. Subsequently, the learned feature embeddings are transferred to a one-class classifier, which undergoes further supervised training to obtain a decision threshold τ. This threshold calibration phase notably utilises an unseen validation dataset containing only normal data samples. Withholding this validation set during ELM feature learning prevents overfitting the threshold to any potential anomalies in the original training data. Overall, this staged approach enables robust unsupervised feature extraction from normal data, followed by supervised

threshold tuning to facilitate effective anomaly detection. Usually, a good threshold τ can be expressed by

$$\tau = \gamma \cdot percentile_p(|1 - \boldsymbol{Y}^{\text{valid}}|) \tag{6-11}$$

where $\boldsymbol{Y}^{\text{valid}}$ is the output of the one-class classifier, $percentile_p$ is a function of the $p-$th percentile with hyperparameters p and $\gamma \geq 0$.

Finally, in the deployment phase, newly observed data samples are propagated through the trained HELM model to obtain the corresponding outputs from the one-class classifier. These outputs, denoted by $\boldsymbol{Y}^{\text{test}}$, are compared against the decision threshold τ established during the training process. Recall that this threshold was calibrated on the separate validation dataset to avoid overfitting. The label assignment for each new test sample is then determined by thresholding its one-class output as follows:

$$\text{Label}_{\boldsymbol{Y}^{\text{test}}} = \text{sgn}\ (\tau - |1 - \boldsymbol{Y}^{\text{test}}|) \tag{6-12}$$

In summary, the trained HELM model generates layered feature representations of newly observed test data in a purely data-driven manner. Anomalies can be effectively detected by propagating these examples through the model and comparing the resulting one-class classifier decisions to the calibrated threshold τ. This approach benefits from the model's unsupervised learning of salient features from normal training data, and the deep HELM architecture captures robust intrinsic representations of standard normal behaviour. By thresholding the one-class outputs relative to τ, deviations from the learned normality are identified during deployment. Overall, this framework provides a self-supervised feature learning mechanism to represent normal data and a thresholding technique for effective anomaly detection in practice. The model framework of the HELM-based semi-supervised machine learning method is delineated in **Figure 6-6**.

6.5 Experiment and results

6.5.1 Dataset arrangement

This study carries out comprehensive experiments to assess the performance of various models and the impact of different input feature conditions on the detection of abnormal driving behaviour. Initially, the built training dataset contained 290,690 instances, which included noisy and inconsistent data. In this study, several techniques were employed to address these issues, such as utilising the *dropna* function in the *pandas* library to eliminate instances with *NULL*, missing, and blank values, as well as refining the original data by employing smoothing techniques to attenuate noise.

The dataset itself includes the following features: *frameNum*, *carId*, *carCenterX (ft)*, *carCenterX (m)*, *carCenterY (ft)*, *carCenterY (m)*, *headX (ft)*, *headY (ft)*, *tailX (ft)*, *tailY (ft)*, *speed (mph)*, *speed (m/s)*, *heading (°)*, and *laneId*, as shown in **Table 6-2**. Next, the time interval was determined by calculating the difference in timestamp values using *frameNum* between adjacent later samples and their corresponding former ones. Based on this, the speed and acceleration (both longitudinal and lateral) for each vehicle were computed. Subsequently, using the *frameNum* as the index, the distances and 2D-TTCs between all relevant vehicles at

the same timestamp were calculated. As the quantity of normal driving data samples is far beyond the abnormal ones, to balance the quantity of abnormal and normal data samples, this study sampled the normal driving samples. In the end, the examined dataset comprised a total of 23,605 samples, consisting of 12,125 normal instances and 11,480 anomaly instances. All anomaly instances are utilised for testing, and 3,638 normal instances are adopted for testing. As anomaly instances are more critical, this study examined more anomaly instances in the estimation of model performance.

6.5.2 Evaluation metrics

Various metrics are adopted to evaluate the overall performance of the selected model, and the discrimination evaluation of the optimal model can be defined based on the confusion matrix (Hossin & Sulaiman, 2015), as shown in **Table 6-4**.

Table 6-4. Confusion matrix and the corresponding array representation

	Actual Positive Class	**Actual Negative Class**
Predicted Positive Class	True-positive (TP)	False-negative (FN)
Predicted Negative Class	False positive (FP)	True-negative (TN)

In binary classification, one class constitutes the positive class, whereas the other delineates the negative class. The positive class epitomises the events the model endeavours to detect, i.e., abnormal driving in this study, while the negative class constitutes other contingencies, i.e., normal driving in this study. True Positive (TP) and True Negative (TN) denote the quantity of accurately classified positive and negative exemplars. In this study, TP represents the correctly detected abnormal driving behaviour data sample, and TN constitutes the accurately detected normal driving samples. On the other hand, False Positive (FP) and False Negative (FN) represent the number of misclassified positive and negative instances, meaning incorrect detection of abnormal driving behaviour/normal driving behaviour instances. Accuracy, Precision, and Recall were computed based on these four terms.

Accuracy refers to the proportion of true results among the total number of cases examined.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{6-13}$$

Precision is utilised to gauge the accurate prediction of positive patterns among the total predicted patterns in a positive class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{6-14}$$

Another widely utilised measure is Recall, which accounts for the proportion of actual Positives that are correctly classified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{6-15}$$

The F1-score is a measure combining and balancing Precision and Recall, and it is defined as the harmonic mean of Precision and Recall:

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision+recall}} \tag{6-16}$$

Finally, the True Positive Rate (TPR) and False Positive Rate (FPR) are also examined as evaluation metrics. As indicated in their names, TPR and FPR are calculated as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{6-17}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \tag{6-18}$$

6.5.3 Ablation study regarding features

Three experimental settings with distinct feature representations are designed to evaluate the impact of input information on model performance. As illustrated in **Table 6-5**, *Setting 1* utilises only the raw coordinates, velocity, and vehicle angle features inherently present in the dataset. *Setting 2* augments *Setting 1* by incorporating two additional engineered features of lateral acceleration and inter-vehicle distance. Finally, *Setting 3* further supplements *Setting 2* by including the 2D-TTC feature capturing temporal proximity. By comparing results between these controlled settings, the incremental value of providing basic motion features (*Setting 2*) and safety indicators, i.e., 2D-TTC (*Setting 3*), over the raw dataset (*Setting 1*) can be quantified. The proposed three experimental settings serve to illustrate the effect of step-wise enriching the feature space on the learning capabilities of the model under controlled conditions.

Table 6-5. Input features in different settings

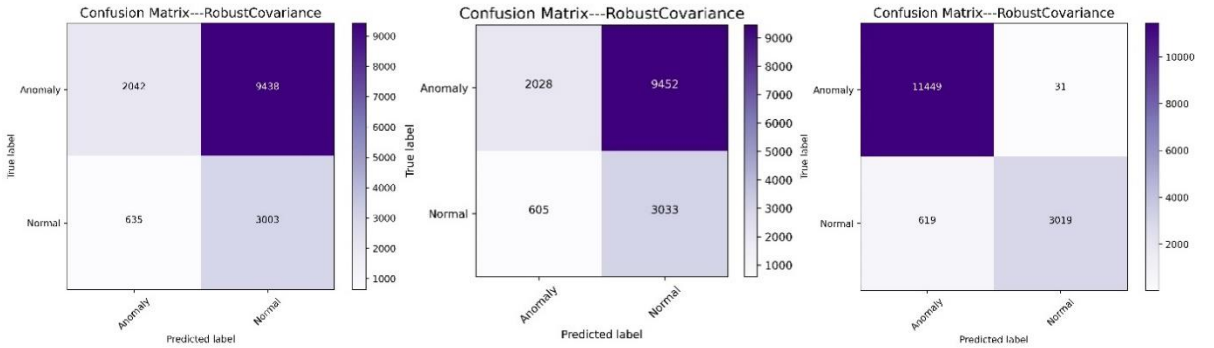
Experimental Setting	Input Features
1	coordinates/velocity/angle
2	coordinates/velocity/angle/acceleration/distance
3	coordinates/velocity/angle/2D time-to-collision

6.5.4 Results and comparison

The testing results of the proposed HELM model, together with the two baselines, are illustrated in **Table 6-6**, as well as **Figures 6-7**, **6-8**, and **6-9**. In general, the HELM model outperforms Robust Covariance and Isolation Forest, with the best variant delivering the best accuracy at 99.58% and the best F1-score at 0.9913.

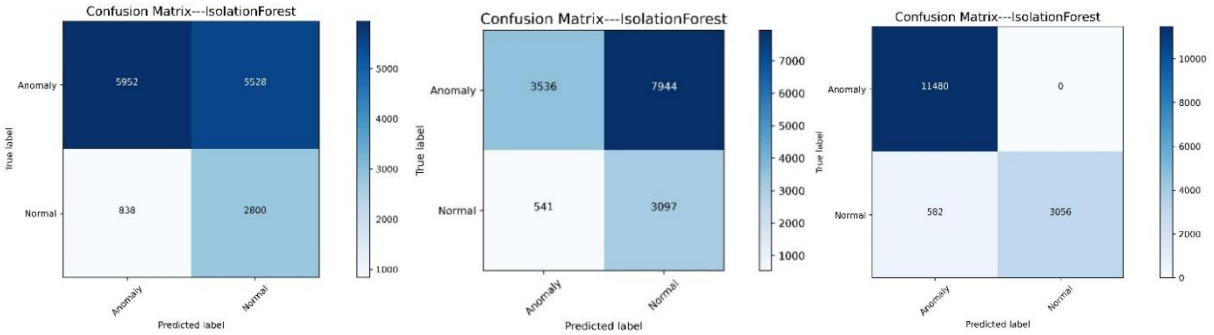
Table 6-6. Results comparison under different settings

Model	Setting	Accuracy	Precision	Recall	F1-score	FPR	TPR
Robust Covariance	1	0.3337	0.7628	0.1779	0.3735	0.1745	0.1779
	2	0.3348	0.7702	0.1767	0.3762	0.1663	0.1767
	3	0.9570	0.9487	0.9973	0.9028	0.1701	0.9973
Isolation Forest	1	0.5789	0.8766	0.5185	0.4680	0.2303	0.5185
	2	0.4387	0.8673	0.3080	0.4219	0.1487	0.3080
	3	0.9615	0.9517	**1.0000**	0.9131	0.1600	**1.0000**
HELM	1	0.9471	0.9349	**1.0000**	0.8766	0.2196	**1.0000**
	2	0.9614	0.9561	0.9949	0.9144	0.1440	0.9949
	3	**0.9958**	**0.9963**	0.9983	**0.9913**	**0.0118**	0.9983



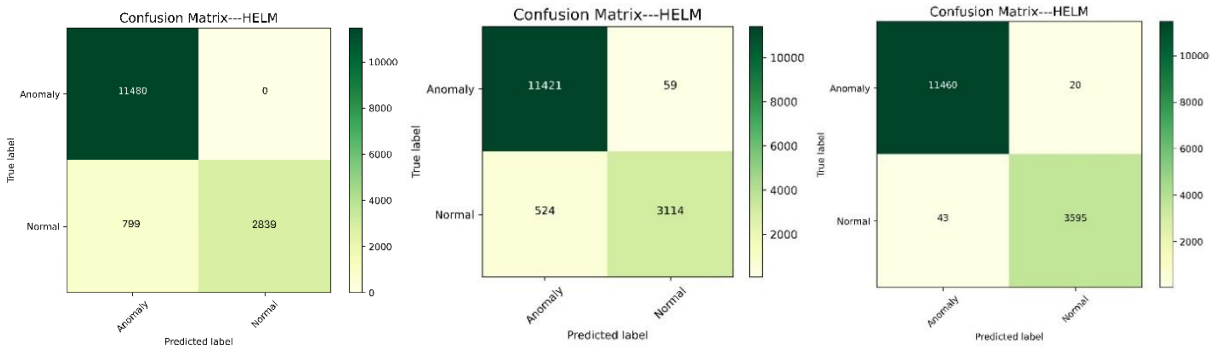
(a) *Setting 1* (b) *Setting 2* (c) *Setting 3*

Figure 6-7. Robust Covariance performance under *Setting 1*, *Setting 2*, and *Setting 3*



(a) *Setting 1* (b) *Setting 2* (c) *Setting 3*

Figure 6-8. Isolation Forest performance under *Setting 1*, *Setting 2*, and *Setting 3*



(a) *Setting 1* (b) *Setting 2* (c) *Setting 3*

Figure 6-9. HELM performance under *Setting 1*, *Setting 2*, and *Setting 3*

Experiments across three experimental settings demonstrate enhanced abnormal driving behaviour identification capabilities by incorporating the safety indicator of 2D-TTC. Furthermore, the proposed semi-supervised HELM model achieves consistently superior performance compared to the alternative baseline models in all three experimental settings.

In the baseline *Setting 1*, with only raw coordinates, velocity, and angle serving as the input features, the HELM model attains an accuracy of 0.9471. Then, augmenting with acceleration and inter-vehicle distance features, in *Setting 2*, the accuracy of HELM is improved to 0.9614. Notably, further inclusion of the adopted 2D-TTC safety indicator in *Setting 3*, the accuracy of HELM is dramatically enhanced to 0.9958, alongside near-perfect scores for Precision (0.9963),

Recall (0.9983), F1-score (0.9913), and FPR (0.0118). This underscores the outstanding value of 2D-TTC as an important spatial-temporal feature for this task.

Similarly, unsupervised models (which work in a semi-supervised way in this study) exhibit substantial gains when endowed with 2D-TTC. For instance, the Precision and Recall of Robust Covariance are improved by over 20%, while the Accuracy and F1-score of Isolation Forest are increased by 5% and 10%, respectively. Nevertheless, the semi-supervised HELM approach outperforms these two baseline models across all metrics except for TPR and Recall.

Finally, scatter visualisation of the result obtained by the proposed semi-supervised method using HELM is provided in **Figure 6-10**. From the visualisation, it is further demonstrated that the HELM can distinguish between normal and abnormal driving behaviours. However, it can not tell the severe abnormal apart from the weak abnormal instances, as the values of their $|1 - \boldsymbol{Y}^{\text{test}}|/\tau$ are similar. How to distinguish the severity of abnormal driving behaviour using semi-supervised machine learning can be an interesting future research direction.

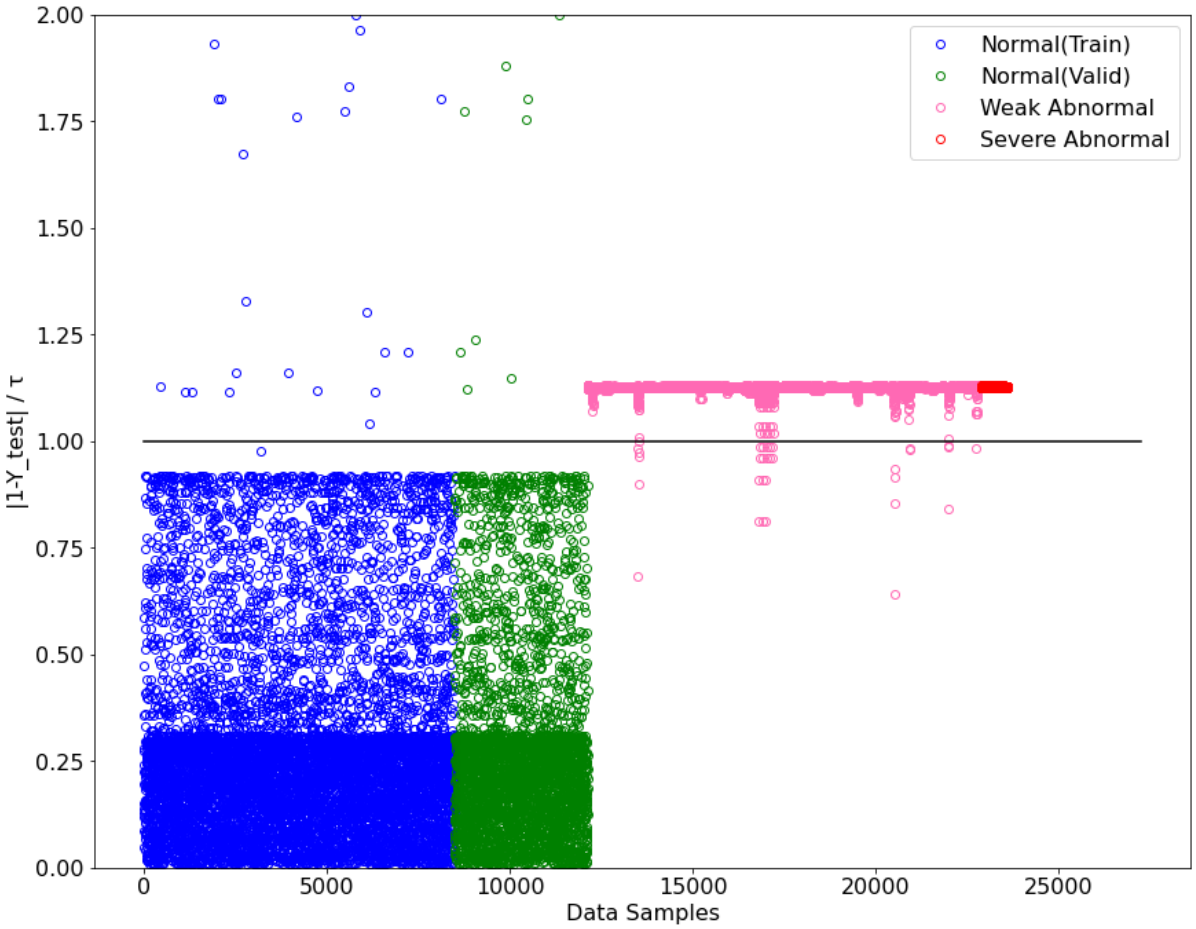


Figure 6-10. Scatter visualisation of the result obtained by semi-supervised HELM

In summary, augmenting the feature space with the adopted safety indicator, i.e., 2D-TTC, consistently improves the anomaly detection capabilities across models. The HELM framework integrating 2D-TTC markedly surpasses other baseline models, demonstrating the advantages and superiority of the proposed semi-supervised learning method together with the spatial-temporal feature engineering for anomalous driving behaviour identification.

6.6 Conclusion and future work

This study presented a semi-supervised machine learning framework leveraging event-level safety indicators to enhance abnormal driving behaviour detection. A large-scale real-world naturalistic driving dataset was analysed, and various abnormal driving behaviours were revealed and categorised in this study. A Hierarchical Extreme Learning Machine (HELM) model was proposed, which harnesses unlabelled data for self-supervised pre-training and partially labelled data for fine-tuning. The 2D-TTC safety indicator was introduced as an important feature, with experiments demonstrating that integrating 2D-TTC significantly improves the detection accuracy by over 5% for all the tested models compared to baseline experimental feature settings.

By training on unlabelled data and employing only a small sample of labelled data for fine-tuning, the proposed semi-supervised approach achieved competitive performance while reducing dependency on fully labelled datasets, making it well-suited for real-world applications with limited labelled data. Notably, the incorporation of event-level safety indicators, in this case, 2D-TTC, greatly enhanced the model performance. These compelling results underscore the critical value of safety indicators in effectively detecting abnormal driving behaviours across diverse ML algorithms. This fusion of semi-supervised ML and utilisation of safety indicators as input features showcase the potential for advancing abnormal driving behaviour detection capabilities, with significant implications for safety-oriented research and evaluations. To further upgrade the detection performance, future studies could explore other and more advanced safety indicators.

Furthermore, the current study focused on detection; future research should explore predictive capabilities to enable earlier identification of impending abnormal behaviours before manifestation. This involves inputting multi-step time-series driving data and computing the features (e.g., TTC, 2D-TTC, MTTC) over a continuous duration period based on observed historical driving behaviour data to predict the status of the next time step or the next few time steps. Additionally, incorporating motion prediction (e.g., for more accurate TTC calculation) and adoption of driving risk field related metrics such as the human perceived Driver's Risk Field (DRF) (Kolekar et al., 2020) and the Probabilistic Driving Risk Field (PDRF) (Mullakkal-Babu et al., 2020), together with developing techniques to extract robust spatial-temporal patterns as model inputs, could further enhance the detection and prediction performance.

Lastly, regarding other limitations, the adopted dataset encompassed only three abnormal driving behaviour types in this study. Future research should incorporate an expanded diversity of abnormal driving behaviours and more advanced safety indicators to enrich the understanding and identification of anomalies. Additionally, ground truth labels are the prerequisite for evaluating the model performance. The current human expert examination-based verification method adopted in this study can not detect missed abnormal driving behaviour instances, but it may correct possible false alarms to upgrade the label quality. It is suggested to adopt more advanced approaches to obtain and verify high-quality ground truth labels, e.g., employing online crowd-sourcing with multiple experts, and using more comprehensive datasets with corresponding video recordings, as well as incorporating fine-labelled accident data from road authorities.

Acknowledgements

This work was supported by the Applied and Technical Sciences (TTW), a subdomain of the Dutch Institute for Scientific Research (NWO) through the Project Safe and Efficient Operation of Automated and Human-Driven Vehicles in Mixed Traffic (SAMEN) under Contract 17187. Additionally, partial support was provided by the Transport & Mobility Institute at Delft University of Technology (TU Delft).

References

Arun, A., Haque, M. M., Bhaskar, A., Washington, S., & Sayed, T. (2021). A systematic mapping review of surrogate safety assessment using traffic conflict techniques. Accident Analysis and Prevention. https://doi.org/10.1016/j.aap.2021.106016

Bonela, S. R., & Kadali, B. R. (2022). Review of traffic safety evaluation at T-intersections using surrogate safety measures in developing countries context. In IATSS Research (Vol. 46, Issue 3, pp. 307–321). Elsevier B.V. https://doi.org/10.1016/j.iatssr.2022.03.001

Bucsuházy, K., Matuchová, E., Zůvala, R., Moravcová, P., Kostíková, M., & Mikulec, R. (2020). Human factors contributing to the road traffic accident occurrence. Transportation Research Procedia. https://doi.org/10.1016/j.trpro.2020.03.057

Butakov, N. (2020). How to build robust anomaly detectors with machine learning. https://www.ericsson.com/en/blog/2020/4/anomaly-detection-with-machine-learning. [Online] Accessed March 9, 2024.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. In ACM Computing Surveys. https://doi.org/10.1145/1541880.1541882

Chen, Z., Yu, J., Zhu, Y., Chen, Y., & Li, M. (2015). D3: Abnormal driving behaviors detection and identification using smartphone sensors. 2015 12th Annual IEEE International Conference on Sensing, Communication, and Networking, SECON 2015, 524–532. https://doi.org/10.1109/SAHCN.2015.7338354

Dai, J., Teng, J., Bai, X., Shen, Z., & Xuan, D. (2010). Mobile phone based drunk driving detection. 2010 4th International Conference on Pervasive Computing Technologies for Healthcare, Pervasive Health 2010. https://doi.org/10.4108/ICST.PERVASIVEHEALTH2010.8901

Dhar, P., Shinde, S., & Bhaduri, A. (2014). Unsafe driving detection system using smartphone as sensor platform. In International Journal of Enhanced Research in Management & Computer Applications (Vol. 3).

Elvik, Rune., Høye, A., Vaa, T., & Sørensen, M. (2009). Factors contributing to road accidents. In The handbook of road safety measures (pp. 36–80). Emerald.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition, 58, 121–134. https://doi.org/10.1016/j.patcog.2016.03.028

Feng, Y., Pickering, S., Chappell, E., Iravani, P., & Brace, C. (2019). A support vector clustering based approach for driving style classification. International Journal of Machine Learning and Computing, 9(3), 344–350. https://doi.org/10.18178/ijmlc.2019.9.3.808

Jia, S., Hui, F., Li, S., Zhao, X., & Khattak, A. J. (2020). Long short-term memory and convolutional neural network for abnormal driving behaviour recognition. IET Intelligent Transport Systems, 14(5), 306–312. https://doi.org/10.1049/iet-its.2019.0200

Jiao, Y. (2024). A fast calculation of two-dimensional Time-To-Collision. https://Github.Com/Yiru-Jiao/Two-Dimensional-Time-To-Collision. [Online] Accessed March 9, 2024.

Hayward, J. C. (1971). Near misses as a measure of safety at urban intersections [Master Thesis]. The Pennsylvania State University.

Kim, D. G., Lee, C., & Park, B. J. (2016). Use of digital tachograph data to provide traffic safety education and evaluate effects on bus driver behavior. Transportation Research Record, 2585, 77–84. https://doi.org/10.3141/2585-09

Kolekar, S., de Winter, J., & Abbink, D. (2020). Human-like driving behaviour emerges from a risk-based driver model. Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-18353-4

Li, Z., Jin, X., & Zhao, X. (2015). Drunk driving detection based on classification of multivariate time series. Journal of Safety Research, 54(June), 61.e29-64. https://doi.org/10.1016/j.jsr.2015.06.007

Lim, S., & Yang, J. H. (2016). Driver state estimation by convolutional neural network using multimodal sensor data. Electronics Letters, 52(17), 1495–1497. https://doi.org/10.1049/el.2016.1393

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. Proceedings - IEEE International Conference on Data Mining, ICDM, 413–422. https://doi.org/10.1109/ICDM.2008.17

Lu, C., He, X., van Lint, H., Tu, H., Happee, R., & Wang, M. (2021). Performance evaluation of surrogate measures of safety with naturalistic driving data. Accident Analysis and Prevention, 162. https://doi.org/10.1016/j.aap.2021.106403

Lu, J., Grembek, O., & Hansen, M. (2022). Learning the representation of surrogate safety measures to identify traffic conflict. Accident Analysis and Prevention, 174. https://doi.org/10.1016/j.aap.2022.106755

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process. https://doi.org/10.5121/ijdkp.2015.5201

Ma, Y., Xie, Z., Chen, S., Qiao, F., & Li, Z. (2023). Real-time detection of abnormal driving behavior based on long short-term memory network and regression residuals. Transportation Research Part C: Emerging Technologies. https://doi.org/10.1016/j.trc.2022.103983

Matousek, M., El-Zohairy, M., Al-Momani, A., Kargl, F., & Bosch, C. (2019). Detecting anomalous driving behavior using neural networks. IEEE Intelligent Vehicles Symposium, Proceedings, 2019-June(Iv), 2229–2235. https://doi.org/10.1109/IVS.2019.8814246

Minderhoud, M. M., & Bovy, P. H. L. (2001). Extended time-to-collision measures for road traffic safety assessment. Accident Analysis and Prevention. https://doi.org/10.1016/S0001-4575(00)00019-1

Mohammadnazar, A., Arvin, R., & Khattak, A. J. (2021). Classifying travelers' driving style using basic safety messages generated by connected vehicles: Application of unsupervised

machine learning. Transportation Research Part C: Emerging Technologies, 122. https://doi.org/10.1016/j.trc.2020.102917

Mullakkal-Babu, F. A., Wang, M., He, X., van Arem, B., & Happee, R. (2020). Probabilistic field approach for motorway driving risk assessment. Transportation Research Part C: Emerging Technologies, 118. https://doi.org/10.1016/j.trc.2020.102716

Nikolaou, D., Ziakopoulos, A., & Yannis, G. (2023). A review of surrogate safety measures uses in historical crash investigations. In Sustainability (Switzerland) (Vol. 15, Issue 9). MDPI. https://doi.org/10.3390/su15097580

Ozbay, K., Yang, H., Bartin, B., & Mudigonda, S. (2008). Derivation and validation of new simulation-based surrogate safety measure. Transportation Research Record, 2083, 105–113. https://doi.org/10.3141/2083-12

Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. In Signal Processing (Vol. 99, pp. 215–249). https://doi.org/10.1016/j.sigpro.2013.12.026

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics. https://doi.org/10.1080/00401706.1999.10485670

Ryan, C., Murphy, F., & Mullins, M. (2021). End-to-end autonomous driving risk analysis: A behavioural anomaly detection approach. IEEE Transactions on Intelligent Transportation Systems, 22(3), 1650–1662. https://doi.org/10.1109/TITS.2020.2975043

Saiprasert, C., & Pattara-Atikom, W. (2013). Smartphone enabled dangerous driving report system. Proceedings of the Annual Hawaii International Conference on System Sciences, 1231–1237. https://doi.org/10.1109/HICSS.2013.484

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. In SN Computer Science (Vol. 2, Issue 3). Springer. https://doi.org/10.1007/s42979-021-00592-x

Saunier, N., Sayed, T., & Lim, C. (2007). Probabilistic collision prediction for vision-based automated road safety analysis. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC. https://doi.org/10.1109/ITSC.2007.4357793

Shahverdy, M., Fathy, M., Berangi, R., & Sabokrou, M. (2021). Driver behaviour detection using 1D convolutional neural networks. Electronics Letters, 57(3), 119–122. https://doi.org/10.1049/ell2.12076

St-Aubin, P., Saunier, N., & Miranda-Moreno, L. (2015). Large-scale automated proactive road safety analysis using video data. Transportation Research Part C: Emerging Technologies. https://doi.org/10.1016/j.trc.2015.04.007

Svensson, Å. (1998). A method for analysing the traffic process in a safety perspective. [Master Thesis]. In Dept. of traffic planning and engineering. The Pennsylvania State University.

Tang, J., Deng, C., & Huang, G. Bin. (2016). Extreme learning machine for multilayer perceptron. IEEE Transactions on Neural Networks and Learning Systems, 27(4), 809–821. https://doi.org/10.1109/TNNLS.2015.2424995

Tarko, A. P. (2018). Surrogate measures of safety. In Safe mobility: Challenges, methodology and solutions (Vol. 11, pp. 383–405). Emerald Publishing Limited. https://doi.org/10.1108/S2044-994120180000011019

van der Horst, R., & Hogema, J. (1994). Time-to-collision and collision avoidance systems. Proceedings of the 6th International Cooperation on Theories and Concepts in Traffic Safety (ICTCT) Workshop. https://www.researchgate.net/publication/237807114

Wang, C., Xie, Y., Huang, H., & Liu, P. (2021). A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. Accident Analysis and Prevention, 157. https://doi.org/10.1016/j.aap.2021.106157

Wang, X., Khattak, A. J., Liu, J., Masghati-Amoli, G., & Son, S. (2015). What is the level of volatility in instantaneous driving decisions? Transportation Research Part C: Emerging Technologies, 58, 413–427. https://doi.org/10.1016/j.trc.2014.12.014

Ward, J. R., Agamennoni, G., Worrall, S., Bender, A., & Nebot, E. (2015). Extending time to collision for probabilistic reasoning in general traffic scenarios. Transportation Research Part C: Emerging Technologies, 51, 66–82. https://doi.org/10.1016/j.trc.2014.11.002

World Health Organization. (2023). Global status report on road safety 2023. https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023

Zheng, O., Abdel-Aty, M., Yue, L., Abdelraouf, A., Wang, Z., & Mahmoud, N. (2023). CitySim: A drone-based vehicle trajectory dataset for safety oriented research and digital twins. Transportation Research Record: Journal of the Transportation Research Board. https://doi.org/10.1177/03611981231185768. https://github.com/ozheng1993/UCF-SST-CitySim-Dataset.

7 Towards developing socially compliant automated vehicles: Advances, expert insights, and a conceptual framework

Abstract

Automated Vehicles (AVs) hold promise for revolutionising transportation by improving road safety, traffic efficiency, and overall mobility. Despite the steady advancement in high-level AVs in recent years, the transition to full automation entails a period of mixed traffic, where AVs of varying automation levels coexist with human-driven vehicles (HDVs). Making AVs socially compliant and understood by human drivers is expected to improve the safety and efficiency of mixed traffic. Thus, ensuring AVs compatibility with HDVs and social acceptance is crucial for their successful and seamless integration into mixed traffic. However, research in this critical area of developing socially compliant AVs (SCAVs) remains sparse. This study carries out the first comprehensive scoping review to assess the current state of the art in developing SCAVs, identifying key concepts, methodological approaches, and research gaps. An expert interview was also conducted to identify critical research gaps and expectations towards SCAVs. Based on the scoping review and expert interview input, a conceptual framework is proposed for the development of SCAVs. The conceptual framework is evaluated using an online survey targeting researchers, technicians, policymakers, and other relevant professionals worldwide. The survey results provide valuable validation and insights, affirming the significance of the proposed conceptual framework in tackling the challenges of integrating AVs into mixed-traffic environments. Additionally, future research perspectives and suggestions are discussed, contributing to the research and development agenda of SCAVs.

This chapter is accepted for journal publication by *Communications in Transportation Research* (currently under publication process), and it has been pre-printed on arXiv.

Dong, Y., Van Arem, B., & Farah, H. (2025). Towards Developing Socially Compliant Automated Vehicles: Advances, Expert Insights, and A Conceptual Framework. arXiv preprint arXiv:2501.06089. https://doi.org/10.48550/arXiv.2501.06089

7.1 Introduction

Automated vehicles (AVs) are expected to benefit traffic safety and efficiency (Greenblatt & Shaheen, 2015; Jamson et al., 2011; Talebpour & Mahmassani, 2016; Yaqoob et al., 2020). Although steady development of higher levels of AVs is gradually witnessed, their deployment will not happen overnight. Instead, a transition period is inevitable, during which AVs with various automation levels will share the same road environment with human drivers, leading to mixed-traffic conditions.

The Society of Automotive Engineers (SAE) defines six levels of driving automation (SAE International, 2021), ranging from No Driving Automation (Level 0) to Full Driving Automation (Level 5). Level 0 has no automation, and the driver is fully responsible for all aspects of driving. Levels 1 and 2 introduce partial automation, where the driver remains responsible for driving, even with the assistance of automated features, and must supervise these features continuously. The difference between Levels 1 and 2 lies in the scope of control supported: Level 1 supports either steering or brake/acceleration, while Level 2 supports both simultaneously, encompassing longitudinal and lateral control. At levels 3, 4, and 5, the automated system monitors the environment with full automation capabilities when the automated driving (AD) features are engaged. However, distinctions exist among these levels. At Level 3, known as conditional automation, drivers must be prepared to intervene and resume control when prompted by the AD features. While at Levels 4 and Level 5, the AD features will never make such requests. For Level 4, the AD features can operate the vehicle only under specific conditions defined by the Operational Design Domain (ODD). In contrast, Level 5 allows the AD features to operate the vehicle under all conditions.

The deployment of AVs with varying levels of automation in mixed traffic introduces new challenges and novel interactions which may potentially create uncertainties and issues that affect both road safety and efficiency (Fagnant & Kockelman, 2015; Farah et al., 2022; Fraedrich et al., 2015; Raju et al., 2022). Moreover, there is a pressing need to ensure the acceptance of AVs by human drivers to seamlessly integrate them into existing traffic systems (Łach & Svyetlichnyy, 2024; Orieno et al., 2024).

Regarding the development of AVs' driving behaviours, previous studies have traditionally prioritised aspects such as safety, efficiency, comfort, and energy consumption (Du et al., 2022; ElSamadisy et al., 2024; Vasile et al., 2023; M. Zhu et al., 2020). While these elements are essential, the growing complexity of mixed-traffic environments, where AVs must coexist with human-driven vehicles (HDVs), highlights the importance of ensuring that AVs' driving behaviours are socially compliant. Referring and upgrading upon the definition provided in (Schwarting et al., 2019), socially compliant driving of AVs can be defined as behaving predictably and complying with the social expectations of human drivers and other surrounding road users (including other AVs) when encountering social dilemmas during driving with intensive interactions (e.g., driving through unsignalised intersections, roundabouts, on-ramp/off-ramp merging, or unprotected left turning). This encompasses compliance with different local driving cultures, norms, cues, formal and informal traffic rules, and behaviours expected in specific contexts. The capability of AVs to drive in a predictable and socially compliant way is critical not only for enhancing safety and efficiency but also for fostering understanding and acceptance of AVs by human drivers. Consequently, there is a growing interest in designing and developing socially compliant automated driving systems. AVs with

socially compliant driving capabilities, i.e., socially compliant AVs (SCAVs), generally correspond to Level 3 to Level 5 automation. While infrequent, certain aspects of socially compliant driving might also be observed at Level 2 or Level 1 automation, where partial driver assistance needs to be provided when requested. Nevertheless, the full potential of socially compliant AVs is most relevant and impactful at higher levels of automation, where AVs are expected to make independent decisions in complex traffic scenarios.

Some preliminary efforts have been made in the domain of socially compliant driving, e.g., (Hang et al., 2021; Kolekar et al., 2020; Schwarting et al., 2019; W. Wang et al., 2022). These studies have laid the important groundwork by exploring various aspects of social compliance of AVs, including modelling social interactions, understanding the dynamics between HDVs and AVs, and developing models for socially aware perception, decision-making, or trajectory planning. However, despite these advancements, research on this emerging topic remains relatively limited, particularly in areas such as the modelling of different driving norms and implicit communication in different cultural backgrounds. The current studies lack a comprehensive, integrated approach that fully addresses the complexities, multidisciplinary, and multifaceted nature of socially compliant driving. Therefore, there is a clear and pressing need for the development of an integrated conceptual framework that can guide future research, providing a holistic understanding of socially compliant driving and helping to design a research agenda to bridge the gaps in the current literature.

To advance research in the domain of SCAVs, this study embarks on a comprehensive approach employing an integrated research method. It begins with a scoping review of the current state of the art, aimed at identifying key concepts, methodological approaches, and research gaps. Additionally, an informal expert interview was conducted to gather insights into critical issues and research expectations towards SCAVs. Subsequently, leveraging the findings from the scoping review and expert interview, a conceptual framework is proposed. This framework incorporates all aspects deemed necessary, based on the scoping review and expert interviews, for the development of SCAVs. To validate and refine the proposed conceptual framework as well as gain further insights, an online survey was developed, and responses from experts worldwide were collected. The survey results provide valuable validation and insights, affirming the significance of the framework for developing SCAVs to safely and efficiently integrate them into mixed-traffic environments. Additionally, suggestions for future enhancements are elicited, contributing to the continuous development of AV technology and guiding potential directions for further research and development.

7.2 Scoping literature review

In this study, a scoping review is adopted to synthesise the current research evidence and state of the practice in scientific peer-reviewed publications, as well as identify the key concepts, predominant research approaches, and research gaps related to SCAVs.

A scoping review was selected over a systematic review due to the exploratory nature of the research objective. Compared to systematic reviews, which aim to provide a synthesis and critical appraisal of the published evidence (Munn et al., 2018), scoping reviews are more suitable to summarise and report the research evidence on emerging and burgeoning topics, where evidence is limited and not yet systematically consolidated. As outlined in (Arksey & O'Malley, 2005; Tafidis et al., 2022), scoping reviews aim to provide a broad overview of

available research, identifying relevant key concepts, methodologies, and gaps that require further investigation. Considering that AVs, especially SCAVs, are still in the early stages of development, with a relatively small body of research, a scoping review approach is more appropriate for mapping the current state of the field.

The scope of the review specifically targets methodologies and technical developments (i.e., the methods, algorithms, platforms, tools, and datasets that have been employed), as well as the substantive content of reviewed studies (e.g., what has been done, what scenarios/manoeuvres have been covered), that are relevant to SCAVs. This focus aligns with the study's goal of proposing a conceptual framework to guide future research and development. The descriptive nature of the scoping review allows for an expansive exploration of the research landscape, offering a foundation for conceptualising SCAVS in the context of mixed-traffic environments. It is important to note that detailed analyses and discussions of the findings and conclusions from the reviewed studies are beyond the scope of this study, as the primary focus is on synthesising key methodological insights to inform the proposed framework.

7.2.1 Five-step approach

In this study, a five-step scoping review was utilised to identify and report related existing literature and map the results. The five steps of the methodological approach are:

- Step 1: Setting up eligibility criteria and information sources
- Step 2: Developing search strategy and process
- Step 3: Screening and selecting studies
- Step 4: Charting and visualising the studies
- Step 5: Summarising, synthesising, and reporting the results

This five-step approach is a condensed version of the well-designed PRISMA Extension for Scoping Reviews (PRISMA-ScR) (Tricco et al., 2018), developed in consultation with an international panel of experts to enhance research and scientific publications.

Step 1: Setting up eligibility criteria and information sources

In this step, eligibility criteria and information sources are established to guide the selection of studies for the scoping review. In principle, only peer-reviewed research papers published in journals and conference proceedings in English up till May 21, 2024, were considered eligible for the scoping review. It is essential that the pertinent studies involve the social interactions between AVs and HDVs, or between AVs and other road users (e.g., cyclists and pedestrians). Publications solely discussing and modelling the social interactions and behaviours among humans (e.g., drivers, cyclists, and pedestrians) without insights into SCAVs are deemed ineligible and thus excluded from the review process. There have been a few review papers including such publications, e.g., (Benrachou et al., 2022; Crosato, Tian, et al., 2023; W. Wang et al., 2022; T. Zhang et al., 2023). Therefore, the main difference and key contribution of the literature review part in this study lie in its dedicated focus on socially compliant driving, specifically emphasising interactions involving AVs or insights toward this goal as a core criterion.

Various academic databases and repositories were used, including Scopus, Web of Science (Web of Science All Databases not the Web of Science Core Collection), IEEE Xplore, and Transport Research International Documentation (TRID) The four databases provide access to a wide range of peer-reviewed research papers published in journals and conference proceedings, offering comprehensive coverage of scholarly literature in the field of transportation and automated driving research.

Step 2: Developing search strategy and process

In this step, a systematic search strategy is developed to identify relevant studies for inclusion in the scoping review. The search strategy encompasses a combination of keywords and controlled vocabulary terms related to socially compliant automated driving, social-aware automated driving, social interaction, automated driving, and other associated concepts. Recognising the varied terminologies used in the domain of automated driving, the search includes different spellings, synonyms, and variants of related concepts to ensure inclusivity. The keywords for each associated term are illustrated in **Table 7-1**, facilitating a nuanced and exhaustive search process.

Table 7-1. Keywords used for each associated term.

Term	Relevant Keywords
Automated vehicle	(Autonomous OR automated OR driverless OR driver-less OR self-driving OR selfdriving) AND (car OR vehicle); (Autonomous OR automated) driving
Socially compliant driving	(Social OR social-aware OR socially compliant OR human-like) AND (driving OR interaction OR behaviour OR behavior OR navigation OR decision-making OR trajectory planning OR planning and control); driving AND (social compliance OR social acceptance)

Boolean operators and truncation were utilised to enhance the precision and comprehensiveness of the search. Furthermore, the employed search strings were tailored to meet the specific requirements (e.g., in length) and functionalities of each selected database. The time range was set to 2000-2024. The language of publications was limited to English. Furthermore, only the publications within the subject areas of Mathematics, Psychology, Physics, Neurosciences, Computer Science, Behavioural Sciences, Social Sciences, Operations Research and Management Science, Engineering (including Transportation, Robotics, Telecommunications, Automation Control Systems, etc.), as well as Science Technology, were considered valid. Publications falling into the other domains, e.g., Art, Architecture, Demography, International Relations, Public Administration, Social Issues, etc., were excluded.

It is also important to mention that the literature search was carried out in two phases. One phase before the conceptual design and online questionnaire survey, and the second phase afterwards to capture new publications that had emerged during that time period.

Step 3: Screening and selecting studies

In this phase, the screening process commences with an initial evaluation of the titles, abstracts, and keywords of the search results to determine their alignment with the research objectives and relevance to the study topic. This preliminary assessment serves to identify potentially

eligible studies for further consideration. Subsequently, the full-text articles of the identified studies undergo a thorough review to assess their eligibility. Only studies that are deemed truly pertinent to the research objectives are selected for inclusion in the scoping review.

Furthermore, to ensure the comprehensiveness of the literature coverage, a backward and forward snowballing technique was employed. This technique involves examining the reference lists of the selected papers and the papers that cite the selected papers to identify additional relevant studies that may have been missed in the initial search.

Step 4: Charting and visualising the studies

In this step, the selected studies undergo abstraction and charting to capture their general characteristics, including authorship details, year of publication, source of publication, the disciplinary focus of the journal or conference, keywords, abstract content, number of citations, etc. This process enables a comprehensive overview of the literature landscape and facilitates the identification of trends, patterns, and relationships among the selected studies.

Furthermore, keyword network analysis using VOSviewer (van Eck & Waltman, 2010) and Sankey diagram visualisation techniques were employed to visually represent the relationships among key terms of methodologies adopted and targeted use cases in the identified studies. Keyword network analysis provides insights into the interconnectedness of key terms and concepts within the literature, highlighting prominent themes and areas of focus. By analysing the co-occurrence and relationships between keywords, researchers can identify clusters of related concepts and uncover overarching themes. The Sankey diagram visualisation offers a graphical representation of the flow of information between different categories or variables, illustrating the distribution and relationships between various elements in the selected studies and providing a holistic view of the research landscape. By visualising the flow of information, researchers can identify patterns, trends, and relationships that may not be immediately apparent from textual analysis alone.

By leveraging these visualisation techniques, the findings of the scoping review are presented in a clear and concise manner, enabling stakeholders to easily interpret and understand the key findings and insights derived from the selected studies. Additionally, visualising the data enhances the accessibility and communicability of the research findings and facilitates knowledge dissemination. So that researchers can gain deeper insights into the structure and content of the literature, ultimately contributing to a more comprehensive understanding of the research field.

Step 5: Summarising, synthesising, and reporting the results

In this final step, the results of the scoping review are synthesised and mapped based on the extracted and charted data, as well as the findings from keyword network analysis and Sankey diagram visualisation. The synthesised results are organised into clusters, highlighting key themes, methodological approaches, application cases, study designs, models, metrics used, and broad findings identified in the selected studies. This allows for the identification of commonalities and differences among studies and provides a comprehensive overview of the literature landscape.

Furthermore, relevant research gaps were identified based on the synthesised results, highlighting areas where further investigation is needed and providing valuable insights for the

development of an integrated conceptual framework that addresses key challenges and opportunities in the development of SCAVs.

7.2.2 Scoping literature review results

(1) Selection of pertinent studies

The literature search through the four selected academic databases and under the aforementioned search process originally returned 1,542 records, i.e., there were 432 records (361 published documents and 71 preprints) by Scopus, 258 records by Web of Science (publications and preprints together), 634 records by IEEE Xplore (including early access articles), and 218 records by TRID. Additionally, 11 studies that were identified during the screening process through snowballing were also added, so that in total, 1,553 studies were qualified for the screening process.

These records were exported as comma-separated values (CSV) files and processed using the Pandas Python Data Analysis Library to merge and group the records and remove the duplicates. Together with the manual examination of the titles, a total of 1,327 valid unique records proceeded to the preliminary checking process. Then based on the title and abstract, 209 studies were identified to be either directly relevant to, or capable of, providing valuable insights into automated driving interactions with HDVs in mixed traffic, among which four are review or survey papers (Benrachou et al., 2022; Crosato, Tian, et al., 2023; W. Wang et al., 2022; T. Zhang et al., 2023), and one is about cognitive architecture design and perspectives (Xie et al., 2020). Following a detailed examination regarding their full text, 68 were finally screened out due to their potential to contribute significantly to the understanding and development of socially compliant automated driving in mixed traffic. Thus, the 68 studies were ultimately selected for in-depth review. **Figure 7-1** illustrates the selection process of pertinent studies under the PRISMA pipeline. A full list of the 209 studies is provided in Supplementary Attachment 1 at: https://lnkd.in/gpceU6gQ.

(2) Charting, visualising, summarising, synthesising, and reporting the results

Firstly, to visualise the key terms, methods, and concepts related to socially compliant driving and the development of SCAVs, the relevant publications identified by the Web of Science search engine were visualised using the keyword network plot by VOSviewer, shown in **Figure 7-2**. Please note this study selected Web of Science as the sole database for visualisation due to VOSviewer's limitations and the practical challenges associated with integrating multiple databases. Using the Web of Science database effectively captured the primary information and relationships between key terms and concepts, making it a suitable choice for constructing the keyword network visualisation. The size of the nodes and thickness of the links depict the scale of the publications in the corresponding areas of the keyword, and the different colour depicts the clusters.

The analysis shows that *decision making* appears to be the most frequent keyword, followed by terms like *agent*, *policy*, *dataset*, *robotics*, *human driver*, *robots*, *safety*, and *efficiency*, among others. From the visualisations, one can also identify the commonly adopted methods and terms, such as *deep learning*, *neural networks*, *game theory*, *model predictive control*, and *optimization*. These results provide valuable global insights for understanding the target domain of socially compliant driving.

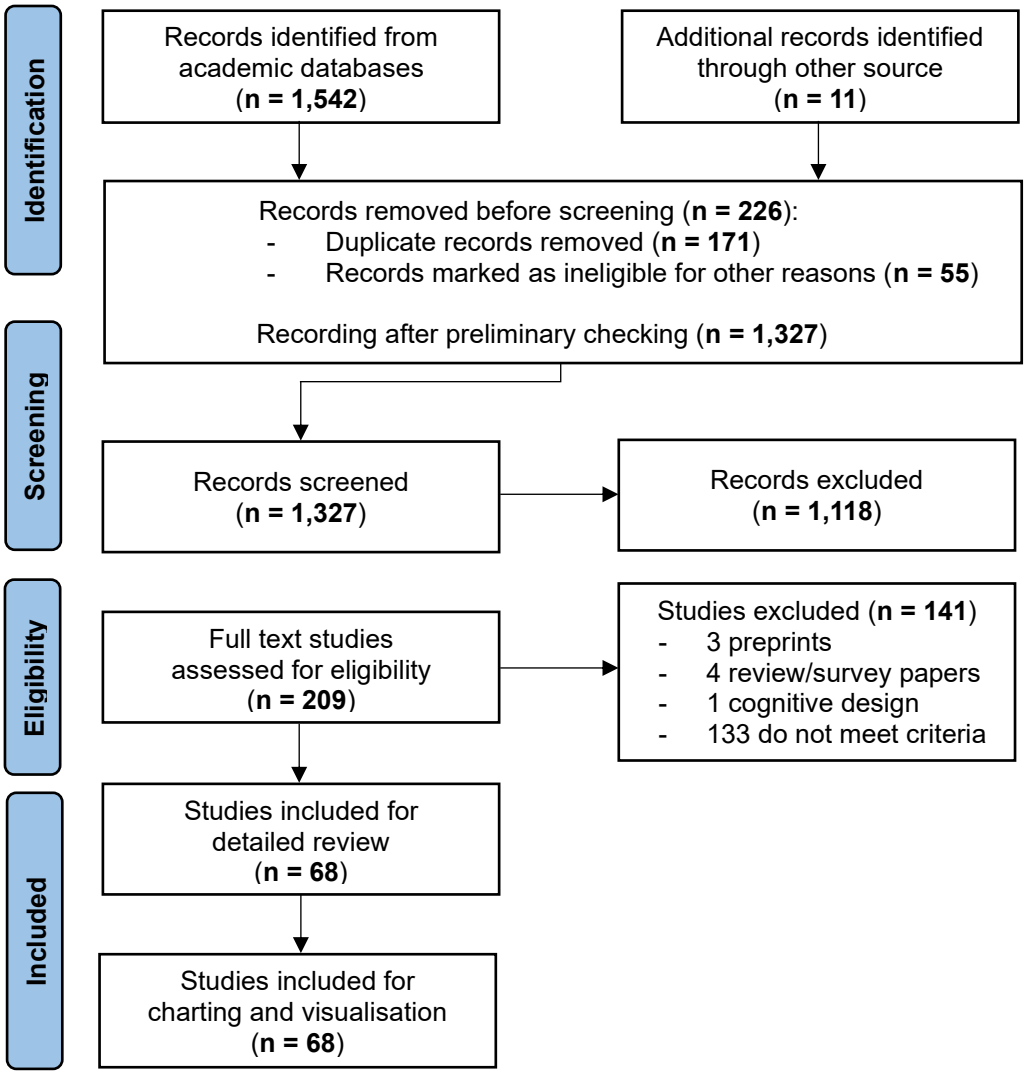


Figure 7-1. Pertinent studies selection process flow diagram

To design SCAVs, methodologies identified in the reviewed literature can be broadly grouped into learning-based and model/utility-based approaches. In practice, learning-based and model-based approaches often complement each other to achieve more robust and adaptable performance. Specifically, the detailed methodologies can be roughly classified into five key sub-categories:

1) Imitation Learning of Social Driving Behaviours from Human Drivers

This approach focuses on replicating the social driving behaviours of human drivers through imitation learning techniques, such as behaviour cloning (Lingguang Wang et al., 2023a, 2023b; Z. Zhu & Zhao, 2023), inverse reinforcement learning (IRL) (Geng et al., 2023; Sun et al., 2019), and generative adversarial imitation learning, e.g., in (Da & Hua, 2023). The AV learns to mimic human-like decision-making and driving patterns by observing and imitating from either expert demonstrations or processed empirical real-world driving data. This method can work in an end-to-end pipeline, but it is not necessary to do so. Representative works in this direction include (Da & Hua, 2023; Sun et al., 2019; Z. Wang et al., 2021; C. Xu et al., 2023).

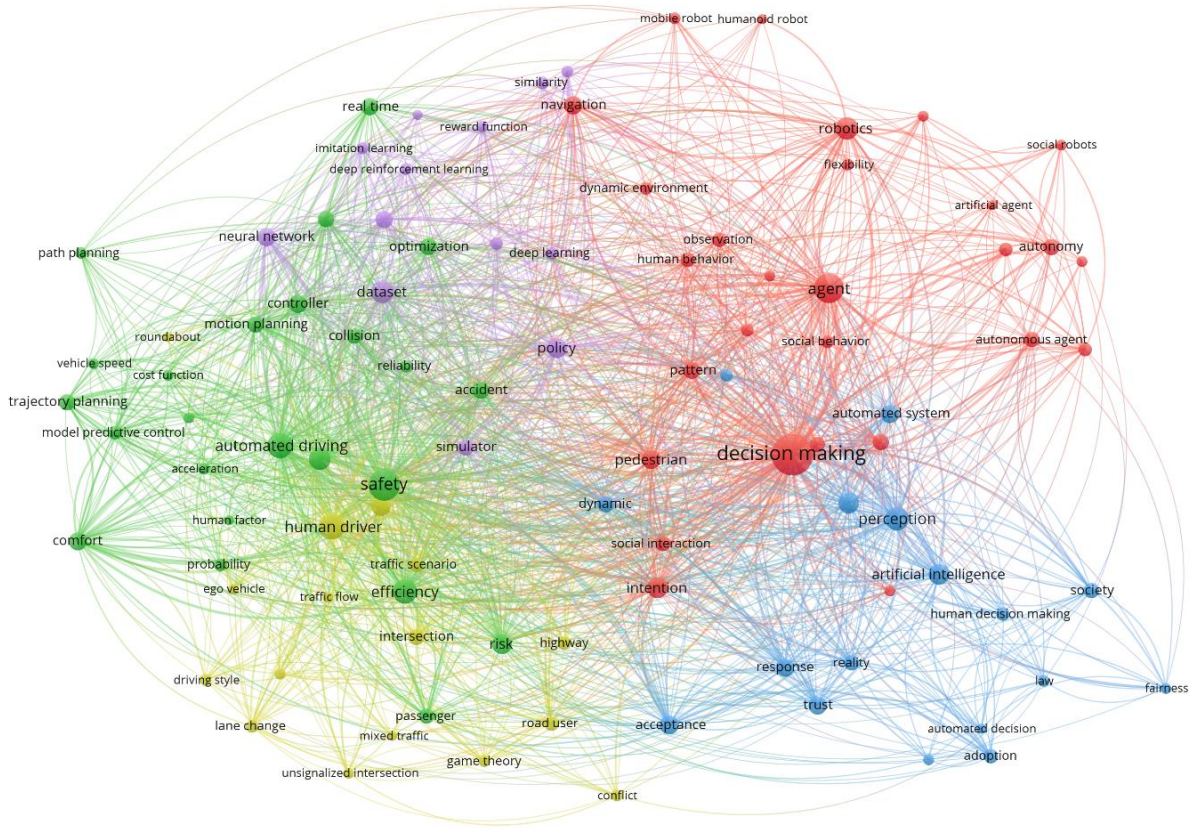


Figure 7-2. Keyword network visualisation by VOSviewer

2) Reinforcement Learning Combined with Utility-based Models

In this approach, reinforcement learning (RL) is employed to infer the underlying utility (also referred to as reward in many studies) functions that govern social driving behaviours from observed human (expert) demonstrations or empirical driving data. The utility functions quantify social factors, such as deterministic courtesy (Sun et al., 2018), and the magnitude of the concern people have for others relative to themselves, e.g., through Social Value Orientation (Liebrand & McClintock, 1988; Murphy & Ackermann, 2014; Schwarting et al., 2019). This method enables AVs to learn and adapt the relevant social factors influencing human decision-making to achieve socially compliant behaviour (Buckman et al., 2019; Larsson et al., 2021; Nan et al., 2024; Schwarting et al., 2019; Letian Wang et al., 2021; Xue et al., 2023; Yoon & Ayalew, 2019).

3) Model-Based Generation of Human-Like Behaviours

This category encompasses approaches that leverage mathematical models to replicate human driving behaviours and/or inform socially aware decision-making. Techniques within this category, such as game theory, social force models, driving risk field models, and potential field models, simulate the complex interaction dynamics between AVs and other road users, including HDVs, pedestrians, and cyclists. Game theory, in particular, provides a framework for strategic decision-making by modelling interactions as a series of cooperative or competitive scenarios where AVs make decisions based on anticipated responses from surrounding agents (Hang et al., 2021; Hang, Lv, et al., 2022; Shu et al., 2023). Other models, like the social force model, e.g., in (Chen et al., 2024; Reddy et al., 2021; Yoon & Ayalew, 2019), driving risk field

model, e.g., in (Geng et al., 2023; Kolekar et al., 2020; J. Wang et al., 2023), and potential field model,e.g., in (Bhatt et al., 2022; Yan et al., 2022; Zhao et al., 2024), capture the forces, risks, and potential outcomes of interactions in mixed-traffic environments, allowing for a more nuanced emulation of human-like behaviours. These model-based approaches are valuable for predicting and generating socially compliant driving behaviours by considering both explicit rules and inferred human tendencies. Notable contributions in this area include, e.g., (Bhatt et al., 2022; Ferrer & Sanfeliu, 2014; Hang et al., 2021; Hang, Lv, Huang, Xing, et al., 2020; Kolekar et al., 2020; L. Zhang et al., 2023; J. Liu, Qi, et al., 2024; Shu et al., 2023; J. Wang et al., 2023).

It can be noted that, usually, these models can be integrated with learning-based approaches (especially RL) to enhance their adaptability and responsiveness in real-time applications, as seen in works like (J. Liu, Qi, et al., 2024; Xiao Wang et al., 2024).

4) Trajectory Prediction through Integration of Social Factors with Machine Learning for Encouraging Socially Compliant Behaviours

This sub-category focuses on the use of machine learning (ML) models, integrated with social factors, to predict trajectories that reflect socially compliant behaviour. Unlike categories (1) and (2), which generally deliver driving control actions, the approaches here rely on deep learning (DL) using deep neural networks (DNNs) or IRL aided by social factor models to analyse and learn from large datasets and forecast the socially compliant trajectories of surrounding HDVs, pedestrians, and/or other road users. By accurately predicting these trajectories, the ego AV can then adjust its actions to achieve corresponding socially compliant driving behaviour, thus ensuring smoother and safer interactions in mixed-traffic scenarios (Geng et al., 2023; Vemula et al., 2018; Yoon & Ayalew, 2019). The prediction can then be used for RL control (Valiente et al., 2024) to leverage prediction and social awareness in RL decision-making, to improve safety and efficiency.

5) Optimisation-Based Tuning of Social Driving Parameters

This approach leverages optimisation techniques to fine-tune the parameters of driving models to achieve desired social objectives, such as individualistic, altruistic, or pro-social driving behaviour. By adjusting and optimising these parameters, the models aim to balance trade-offs between safety, efficiency, and comfort while considering the benefits of the ego AV versus surrounding vehicles or other road participants in mixed-traffic environments. Representative studies in this category include, e.g., (Larsson et al., 2021).

To provide a holistic view of the five identified methodological categories, **Table 7-2** presents a comparative overview of their key characteristics, including advantages, disadvantages, and typical applications. This comparison elucidates the trade-offs and appropriate contexts for each approach, facilitating a deeper understanding of their roles in SCAV development.

As depicted in **Table 7-2**, each category offers distinct strengths and faces specific challenges. Imitation learning excels at replicating human behaviour but is constrained by the diversity and quality of available training data. Reinforcement learning offers adaptability to complex, dynamic settings, yet its effectiveness hinges on well-designed reward functions. Model-based approaches provide interpretable and theoretically sound frameworks for understanding interactions, such as through game theory, though they often demand significant computational

resources and may struggle with adaptability and real-time application. Trajectory prediction, enriched by social factors, improves the anticipation of other road users' movements but depends on robust social data. Lastly, optimisation-based tuning allows precise adjustments to driving parameters, though it may miss nuanced, dynamic social cues. Notably, these methodologies are complementary; combining them can harness their respective strengths to develop more robust and effective SCAV systems.

Table 7-2. Comparison of the five identified methodological categories

Methodological Category	Advantages	Disadvantages	Typical Application
Imitation learning of social driving behaviours from human drivers	Effectively replicates human behaviour	Limited by diversity and quality of training data	Learning social driving norms from expert demonstrations
Reinforcement learning combined with utility-based models	Highly adaptable to dynamic environments	Sensitive to reward function design	Optimising long-term socially compliant behaviour
Model-based generation of human-like behaviours	Structured, interpretable, and theoretically grounded	Computationally intensive; may lack adaptability and real-time feasibility	Modelling strategic multi-agent interactions (e.g., using game theory)
Trajectory prediction through integration of social factors with machine learning	Enhances prediction accuracy with social context	Reliant on the quality and availability of social data	Anticipating movements of HDVs or pedestrians
Optimisation-based tuning of social driving parameters	Offers precise control over driving parameters	May overlook dynamic or implicit social cues	Fine-tuning AV responses for specific social scenarios

These aforementioned methodologies collectively represent the current state of research in socially compliant driving behaviour for AVs. They highlight the multidisciplinary nature of the field, which combines elements of artificial intelligence (AI) (e.g., ML, DL, RL), physics, human factors, control theory, social psychology, and transportation engineering. The integration of multidisciplinary knowledge is crucial for developing AVs capable of safely and efficiently interacting with HDVs and other road users in complex traffic environments. It is important to note that the different approaches categorised are not mutually exclusive: in practice, they can be utilised in combination to enhance the robustness and reliability of AV behaviour. A detailed illustration of the models, terms, and methods adopted by the studies is provided in **Table 7-3** and **Figure 7-3**.

From **Table 7-3**, it is noticed that the majority of studies adopt machine learning approaches, and more specifically, deep learning (e.g., Convolutional Neural Network (CNN), Generative Adversarial Networks (GAN), Long Short-Term Memory (LSTM) neural networks, Multi-Layer Perceptron (MLP), Gated Recurrent Unit (GRU), Transformer), and deep reinforcement learning (e.g., IRL, Deep Q-learning, and Actor-Critic methods). Typically, driving decision-making is modeled as the Markov Decision Process (MDP), e.g., in (Crosato et al., 2021; Da & Hua, 2023; Ding et al., 2022; Hang et al., 2021; Z. Huang, Wu, et al., 2023; J. Liu, Zhou, et al., 2024; Zong et al., 2023), or as the partially observable Markov decision process (POMDP), e.g., in (Ding et al., 2022; J. Liu, Zhou, et al., 2024; Peng et al., 2021) to account for uncertainties. Additionally, a substantial number of studies employ game theory (e.g., Stackelberg game,

Table 7-3. Clustering of methods identified in the papers reviewed

(A) Machine Learning based methods

Methods and Terms Adopted			Related Publications
Machine Learning (DL, RL)	Deep Learning	CNN	(Ding et al., 2022; Hirose et al., 2024; Pérez-Dattari et al., 2022; Qin et al., 2021; Valiente et al., 2024)
		GAN	(Da & Hua, 2023; Gupta et al., 2018; Kothari & Alahi, 2023; Sadeghian et al., 2019; Z. Wang et al., 2021)
		LSTM	(Alahi et al., 2016; W. J. Chang et al., 2023; Da & Hua, 2023; Ding et al., 2022; Gupta et al., 2018; Z. Huang, Liu, et al., 2023; Kothari et al., 2021; Kothari & Alahi, 2023; Pérez-Dattari et al., 2022; Sadeghian et al., 2019; Vemula et al., 2018; Xueyang Wang et al., 2024; Z. Wang et al., 2021)
		MLP	(W. J. Chang et al., 2023; Da & Hua, 2023; Z. Huang, Liu, et al., 2023; Kothari & Alahi, 2023; Xue et al., 2023; Z. Zhu & Zhao, 2023)
		Transformer	(Geng et al., 2023; B. Huang & Sun, 2023; Z. Huang, Liu, et al., 2023; Xiao Wang et al., 2024)
		Attention Module	(Kothari & Alahi, 2023; J. Liu, Zhou, et al., 2024; Qin et al., 2021; Sadeghian et al., 2019; Vemula et al., 2018; Z. Wang et al., 2021; Xue et al., 2023)
		Graph Attention Network	(Xueyang Wang et al., 2024)
		Autoencoder	(J. Liu, Zhou, et al., 2024; Valiente et al., 2024; Zong et al., 2023)
		GRU	(J. Liu, Zhou, et al., 2024; Zong et al., 2023)
		Social Pooling Layer	(Alahi et al., 2016; Gupta et al., 2018)
	Reinforcement Learning	Actor-Critic	(Crosato et al., 2021; Crosato, Shum, et al., 2023; Z. Huang, Wu, et al., 2023; J. Liu, Zhou, et al., 2024; L. Liu et al., 2020; Toghi et al., 2021a; Tong et al., 2024; Xue et al., 2023; Zong et al., 2023)
		Deep Q-learning	(Z. Huang, Wu, et al., 2023; Lu et al., 2022; Nan et al., 2024; Taghavifar & Mohammadzadeh, 2024; Toghi et al., 2021b, 2022; Valiente et al., 2024)
		IRL	(Geng et al., 2023; Z. Huang, Liu, et al., 2023; Nan et al., 2024; Schwarting et al., 2019; Sun et al., 2018, 2019; C. Xu et al., 2023; Zhao et al., 2024)
		PPO	(Crosato, Shum, et al., 2023; J. Liu, Zhou, et al., 2024)
		Coordinated Policy Optimisation	(Peng et al., 2021)

(B) Game theory, field-based models, and social psychological factor related methods

Methods and Terms Adopted		Related Publications
Game Theory	Stackelberg Game	(Hang et al., 2021; Hang, Huang, et al., 2022b; Hang, Lv, Huang, Cai, et al., 2020; C. Li et al., 2022; Schwarting et al., 2019; Letian Wang et al., 2021; Zhao et al., 2024)
	Nash-Equilibrium based Game	(Galati et al., 2022; Hang et al., 2021; Hang, Huang, et al., 2022b; J. Liu, Qi, et al., 2024; M. Liu et al., 2024; Shu et al., 2023; J. Wang et al., 2023)
	POSG	(Toghi et al., 2021b, 2022; Valiente et al., 2024; Xue et al., 2023)
	Coalitional Game	(Hang, Huang, et al., 2022a; Hang, Lv, et al., 2022)
	Potential Game	(M. Liu et al., 2024)
Social Psychological Factor	SVO	(Buckman et al., 2019; Crosato et al., 2021; Crosato, Shum, et al., 2023; Peng et al., 2021; Schwarting et al., 2019; Taghavifar & Mohammadzadeh, 2024; Toghi et al., 2021a, 2021b, 2022; Tong et al., 2024; Valiente et al., 2024; Xue et al., 2023; L. Zhang et al., 2023; Zhao et al., 2024)
	Courtesy	(W. J. Chang et al., 2023; C. Li et al., 2022; Sun et al., 2018; Letian Wang et al., 2021)
	Coordination Tendency	(J. Liu, Zhou, et al., 2024)
	Social Preference	(Lu et al., 2022)
	Social Cohesion	(Landolfi & Dragan, 2018)
	Social Anchor	(Kothari et al., 2021)
Field-based Models	Potential Field	(Bhatt et al., 2022; Hang et al., 2021; Hang, Huang, et al., 2022a, 2022b; Hang, Lv, Huang, Cai, et al., 2020; Reddy et al., 2021; Yan et al., 2022; Zhao et al., 2024)
	Risk Field	(Geng et al., 2023; Kolekar et al., 2020; J. Wang et al., 2023; Xiao Wang et al., 2024; L. Zhang et al., 2023)

(C) Other models and methods

Methods and Terms Adopted	Related Publications
Model Predictive Control	(Bhatt et al., 2022; Hang et al., 2021; Hang, Huang, et al., 2022a; Hang, Lv, Huang, Cai, et al., 2020; Landolfi & Dragan, 2018; Larsson et al., 2021; Pérez-Dattari et al., 2022; Sun et al., 2018, 2019; J. Wang et al., 2023; Letian Wang et al., 2021; Yan et al., 2022; Yoon & Ayalew, 2019; L. Zhang et al., 2023)
Markov Decision Process	(Crosato et al., 2021; Crosato, Shum, et al., 2023; Da & Hua, 2023; Ding et al., 2022; Z. Huang, Wu, et al., 2023; J. Liu, Zhou, et al., 2024; Peng et al., 2021; Song et al., 2016; Zong et al., 2023)
Expert Demonstration	(Da & Hua, 2023; Z. Huang, Liu, et al., 2023; Z. Huang, Wu, et al., 2023; J. Liu, Qi, et al., 2024; Nan et al., 2024; Qin et al., 2021; Z. Zhu & Zhao, 2023)
Social Force Model	(Chen et al., 2024; Crosato, Shum, et al., 2023; Ferrer & Sanfeliu, 2014; Reddy et al., 2021; Yoon & Ayalew, 2019)
Addressing Uncertainties	(Z. Huang, Wu, et al., 2023; Kolekar et al., 2020; Sun et al., 2019; Letian Wang et al., 2021)
Bayesian Inference	(C. Li et al., 2022; J. Wang et al., 2023; Letian Wang et al., 2021)
Behaviour Cloning	(Lingguang Wang et al., 2023a, 2023b)
Monte-Carlo Sampling	(Lingguang Wang et al., 2023a, 2023b)
Monte Carlo Tree Search	(C. Li et al., 2022)
Finite State Machine	(B. Wang et al., 2024)
Reasoning Graph	(D. Zhou et al., 2022)
Non-Convex Mixed-Integer Nonlinear Program	(Larsson et al., 2021)
Discrete Choice Model	(Kothari et al., 2021)
Minimising Counterfactual Perturbation	(Hirose et al., 2024)
Particle Filtering	(C. Xu et al., 2023)
Gaussian Process	(Valiente et al., 2024)
Genetic Algorithm	(J. Liu, Qi, et al., 2024)

coalitional game, potential game, and Partially Observable Stochastic Game (POSG)) to effectively model complex interactions between agents (e.g., AVs and HDVs), while a significant portion also utilises model predictive control (MPC) to refine and smooth control outputs following decision-making. In the realm of social preferences, a variety of social psychological terms, such as courtesy, coordination tendency, and Social Value Orientation (SVO), are used to encapsulate concepts related to social preferences. The targeting research objectives and tasks typically fall into three primary categories: behaviour generation, trajectory prediction, as well as interactive decision-making and control. Further, multiple-agent modelling is incorporated in some studies to simulate complex, interactive driving environments involving multiple road participants, e.g., in (Da & Hua, 2023; Peng et al., 2021; Toghi et al., 2021a, 2021b, 2022; Xue et al., 2023). These observations align with insights from the keyword network visualisation in **Figure 7-2** and are illustrated further in **Figure 7-3**.

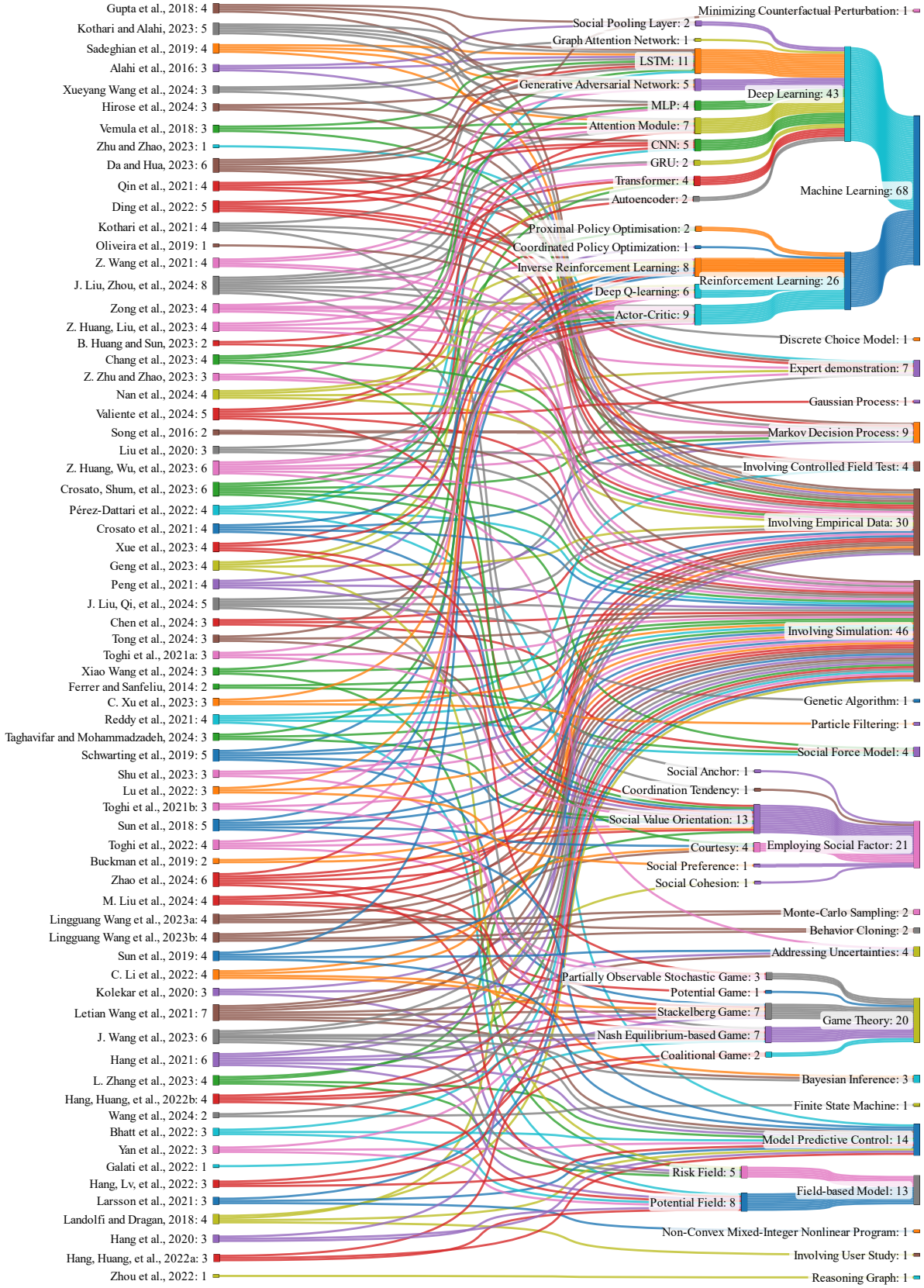


Figure 7-3. The identified methods adopted in each study

Note: A single paper may involve multiple methods (e.g., both Deep Learning and Reinforcement Learning), and may utilise multiple models within the same method category (e.g., both LSTM and CNN within the Deep Learning category).

Furthermore, while some interdisciplinary initiatives have been introduced, the majority of research continues to focus on combining approaches from computer science, physics, mathematics, transportation, and vehicular engineering. Although initial efforts to incorporate social psychology are emerging, they primarily centre around concepts like Social Value Orientation (SVO), coordination tendencies, and courtesy, which share common themes. The development of more advanced models grounded in social psychology and other relevant interdisciplinary fields is essential to deepen the understanding of human-AV interactions (Brown Et Al., 2023; Vinkhuyzen & Cefkin, 2016). Specifically, incorporating culturally sensitive social behaviours into AV decision-making to develop customised AVs for diverse cultural backgrounds remains a crucial area for further investigation (Dong et al., 2024).

Table 7-4 groups the reviewed papers based on simulation, data-driven, and empirical field testing approaches. From **Table 7-4** and **Figure 7-3**, it is revealed that more than half of the studies employed simulations to train, test, and verify their solutions. The commonly adopted simulation platforms and software tools include Highway-env (Leurent, 2018), SMARTS (M. Zhou et al., 2020), CARLA (Dosovitskiy et al., 2017), MetaDrive (Q. Li et al., 2023), PTV VISSIM, SUMO (Lopez et al., 2018), Universe simulator (D. Zhang, 2023), and Robot Operation System (ROS). Additionally, more than half of the studies incorporated empirical datasets collected from real-world environments to enhance model validation. Typical frequently used datasets include the Next Generation Simulation (NGSIM) dataset (U.S. Department of Transportation Federal Highway Administration, 2016), Waymo Open Motion dataset (Ettinger et al., 2021), INTERACTION dataset (Zhan et al., 2019), highD dataset (Krajewski et al., 2018), exiD dataset (Moers et al., 2022), inD dataset (Bock et al., 2020), rounD dataset (Krajewski et al., 2020), SinD dataset (Y. Xu et al., 2022), Argoverse Motion dataset (M. F. Chang et al., 2019), and Argoverse 2 Motion dataset (Wilson et al., 2023). Additional datasets, including the ETH (Pellegrini et al., 2009), UCY (Lerner et al., 2007), TrajNet++ (Kothari et al., 2022), PANDA (Xueyang Wang et al., 2020), Stanford Drone (Robicquet et al., 2016), and HuRoN (Hirose et al., 2024) datasets, are employed for scenarios and applications related to social robot navigation and human trajectory prediction.

Furthermore, as clearly illustrated in **Table 7-5** and **Figure 7-4**, regarding driving manoeuvres, the majority of studies focus on ones that require both longitudinal and lateral control. Various manoeuvres, e.g., driving through unsignalised intersections, performing unprotected left turns, lane changing, on-ramp merging, and overtaking, have been studied. The inherent complexity and dynamic nature of these scenarios, where both directional and speed-related aspects of control must be simultaneously managed, make them particularly well-suited for studying and examining social interactions between AVs and HDVs. Such scenarios provide robust "environments" for developing and validating socially compliant driving behaviours, as they compel AVs to navigate nuanced interactions, accommodate unpredictable human behaviours, convey their intentions, and adapt their decisions to align with various human social driving patterns. Interestingly, within the reviewed publications, only two studies specifically delve into manoeuvres involving only longitudinal control, i.e., car-following. This may stem from the fact that longitudinal manoeuvres are often already embedded within the broader, more complex scenarios mentioned above, thus, there is no need to specifically only target longitudinal manoeuvres.

Table 7-4. Grouping of reviewed papers based on simulation, data-driven, and empirical field testing approaches

Methods Adopted	Tools, Platforms, or Datasets	Related Publications
Simulation and simulator-related	Highway-env	(J. Liu, Zhou, et al., 2024; Toghi et al., 2021a, 2021b, 2022; Tong et al., 2024; Valiente et al., 2024; L. Zhang et al., 2023)
	SMARTS	(Z. Huang, Wu, et al., 2023; Xiao Wang et al., 2024)
	CARLA	(Bhatt et al., 2022; Lu et al., 2022; Pérez-Dattari et al., 2022; Z. Zhu & Zhao, 2023)
	MetaDrive	(Peng et al., 2021)
	Python-based	(Crosato et al., 2021; Crosato, Shum, et al., 2023; Da & Hua, 2023; L. Liu et al., 2020; Z. Wang et al., 2021)
	Python-Matlab	(Zhao et al., 2024)
	Matlab-Simulink	(Hang et al., 2021; Hang, Huang, et al., 2022a; Hang, Lv, et al., 2022; Hang, Lv, Huang, Cai, et al., 2020)
	Prescan	(Song et al., 2016)
	CarSim	(Chen et al., 2024)
	Matlab/Simulink-CarSim	(Yan et al., 2022)
	Prescan-MATLAB/Simulink-CarSim	(J. Wang et al., 2023)
	Robot Operation System (ROS)	(Pérez-Dattari et al., 2022; Letian Wang et al., 2021)
	SUMO-ROS	(Zong et al., 2023)
	PTV VISSIM	(Larsson et al., 2021)
	Julia	(Sun et al., 2018)
	MobileSim	(Reddy et al., 2021)
	Universe Simulator	(Xue et al., 2023)
	Fixed based Driving Simulator	(Kolekar et al., 2020)
	Human-in-the-loop driver simulator	(J. Liu, Qi, et al., 2024; C. Xu et al., 2023)
	Hardware-in-the-loop simulator	(Hang, Huang, et al., 2022b)
	Self-built upon datasets	(Lingguang Wang et al., 2023a)
	Not specified	(Buckman et al., 2019; Ferrer & Sanfeliu, 2014; Landolfi & Dragan, 2018; Schwarting et al., 2019; Shu et al., 2023; Sun et al., 2019; Taghavifar & Mohammadzadeh, 2024; B. Wang et al., 2024; D. Zhou et al., 2022)

Table 7-4. Continued

Methods Adopted	Tools, Platforms, or Datasets		Related Publications
Involving empirical data	Next Generation Simulation (NGSIM) Dataset		(Chen et al., 2024; Hang et al., 2021; M. Liu et al., 2024; Nan et al., 2024; Schwarting et al., 2019; Sun et al., 2018; J. Wang et al., 2023; Zhao et al., 2024)
	Waymo Open Motion Dataset		(W. J. Chang et al., 2023; Z. Huang, Liu, et al., 2023)
	INTERACTION Dataset		(B. Huang & Sun, 2023; C. Li et al., 2022; Shu et al., 2023; Tong et al., 2024; Letian Wang et al., 2021; Lingguang Wang et al., 2023b)
	highD Dataset		(Lingguang Wang et al., 2023a; C. Xu et al., 2023)
	exiD Dataset		(Lingguang Wang et al., 2023a)
	inD Dataset		(Geng et al., 2023; Lingguang Wang et al., 2023b)
	rounD Dataset		(Lingguang Wang et al., 2023b)
	SinD Dataset		(J. Liu, Qi, et al., 2024)
	Argoverse Motion Dataset		(Ding et al., 2022)
	Argoverse2 Motion Dataset		(J. Liu, Qi, et al., 2024)
	Beijing Jianguomen Flyover Area Dataset		(Z. Wang et al., 2021)
	Data collected by wheelchair testbed		(Qin et al., 2021)
	Data collected over 60 hours of driving from 10 drivers at 6 intersections		(Z. Zhu & Zhao, 2023)
	Datasets related to social robot navigation/ human trajectory prediction	PANDA	(Xueyang Wang et al., 2024)
		ETH	(Alahi et al., 2016; Gupta et al., 2018; Kothari & Alahi, 2023; Sadeghian et al., 2019; Vemula et al., 2018; Xueyang Wang et al., 2024)
		UCY	(Alahi et al., 2016; Gupta et al., 2018; Kothari & Alahi, 2023; Sadeghian et al., 2019; Vemula et al., 2018; Xueyang Wang et al., 2024)
		TrajNet ++	(Kothari et al., 2021; Kothari & Alahi, 2023)
		Stanford Drone Dataset	(Sadeghian et al., 2019)
		HuRoN	(Hirose et al., 2024)
Involving controlled field test			(Ding et al., 2022; Ferrer & Sanfeliu, 2014; Hirose et al., 2024; L. Liu et al., 2020; Oliveira et al., 2019; Reddy et al., 2021)
Involving survey questionnaire			(Galati et al., 2022)
Involving user study			(Landolfi & Dragan, 2018)

Table 7-5. Clustering of manoeuvres and applications identified in the reviewed papers

Use Cases		Related Publications
Intersection	Unsignalised intersection [a]	(Buckman et al., 2019; Geng et al., 2023; Hang, Huang, et al., 2022a; J. Liu, Qi, et al., 2024; M. Liu et al., 2024; Peng et al., 2021; Song et al., 2016; Valiente et al., 2024; Xia et al., 2022; Z. Zhu & Zhao, 2023; Zong et al., 2023)
	Unprotected left turn [a]	(Hang, Huang, et al., 2022b; Z. Huang, Wu, et al., 2023; J. Liu, Qi, et al., 2024; J. Liu, Zhou, et al., 2024; Schwarting et al., 2019; Shu et al., 2023; Xiao Wang et al., 2024; D. Zhou et al., 2022; Zong et al., 2023)
	Roundabout	(Z. Huang, Wu, et al., 2023; C. Li et al., 2022; Peng et al., 2021; Valiente et al., 2024; Letian Wang et al., 2021; L. Zhang et al., 2023)
	T-junction	(Oliveira et al., 2019; Pérez-Dattari et al., 2022; Tong et al., 2024)
Lane change	Highway driving	(Hang, Lv, Huang, Cai, et al., 2020; Larsson et al., 2021; Lingguang Wang et al., 2023a; C. Xu et al., 2023; Zhao et al., 2024)
	Urban driving	(W. J. Chang et al., 2023; Z. Wang et al., 2021)
	Two-lane road with large curvature	(Yan et al., 2022)
	Not specific	(Chen et al., 2024)
Merge	On-ramp merging	On-ramp merging: (Hang et al., 2021; Hang, Lv, et al., 2022; M. Liu et al., 2024; Nan et al., 2024; Schwarting et al., 2019; Toghi et al., 2021b, 2021a, 2022; Valiente et al., 2024; Lingguang Wang et al., 2023a; Xue et al., 2023)
	Intersection merging	(Xiao Wang et al., 2024)
Overtaking	Urban driving	(Lu et al., 2022; Zong et al., 2023)
	Highway driving	(Hang et al., 2021; Hang, Lv, Huang, Cai, et al., 2020; Zhao et al., 2024)
	Not specific	(Xiao Wang et al., 2024)
Highway exit		(Landolfi & Dragan, 2018; Toghi et al., 2022; Valiente et al., 2024; Lingguang Wang et al., 2023a)
Interact with pedestrian/ Pedestrian collision avoidance		(Bhatt et al., 2022; Crosato et al., 2021; Crosato, Shum, et al., 2023; Pérez-Dattari et al., 2022; Sun et al., 2019; Taghavifar & Mohammadzadeh, 2024)
Road cruising		(Xiao Wang et al., 2024)
Platoon		(B. Wang et al., 2024)
Bottleneck		(Peng et al., 2021; Xue et al., 2023)
Tollgate		(Peng et al., 2021)
Parking lot		(Peng et al., 2021)
Nudging parked cars on urban streets		(Bhatt et al., 2022)
Social occlusion inference		(B. Huang & Sun, 2023)
Oncoming traffic		(Kolekar et al., 2020; M. Liu et al., 2024)
Reacts to stalled car		(Landolfi & Dragan, 2018)
Reacts to speeding		(Landolfi & Dragan, 2018)
Reacts to ambulance		(Landolfi & Dragan, 2018)
Car-following		(Kolekar et al., 2020; Larsson et al., 2021)

[a] *Unsignalised intersection*: Here "unsignalised intersection" may include "unprotected left turn" or can be other scenarios (e.g., right-turning and going straight at unsignalised intersections), while the row of "unprotected left turn" is specifically about unprotected left turning through unsignalised intersections.

Table 7-5. Continued

Use Cases	Related Publications
Social robot navigating	(Da & Hua, 2023; Ferrer & Sanfeliu, 2014; Galati et al., 2022; Hirose et al., 2024; L. Liu et al., 2020; Reddy et al., 2021)
Human trajectory forecasting	(Alahi et al., 2016; Gupta et al., 2018; Kothari et al., 2021; Kothari & Alahi, 2023; Sadeghian et al., 2019; Vemula et al., 2018; Xueyang Wang et al., 2024)
Social reactions, feedbacks, and trust in AVs	(Joo & Kim, 2023; Oliveira et al., 2019; Othman, 2021; Schneble & Shaw, 2021)

From **Table 7-5** and **Figure 7-4**, it is also important to note that some studies focus primarily on social robot navigation and human trajectory forecasting for related applications, with 12 studies included in the review. While AVs can be considered a type of robot, and the insights from social robot navigation research could be beneficial for developing socially compliant driving, there are notable differences between human/pedestrian-robot interactions and the interactions between HDVs and AVs. These differences stem from the distinct speeds, operational environments, and interaction dynamics between the two scenarios. Social robot navigation often occurs at lower speeds and in more controlled environments, which facilitates the use of field test experiments to observe and refine socially aware behaviours. Insights gained from such experiments could serve as a foundation for adaptation to the more complex and high-speed interactions involved in AV driving. This study highlights some typical works related to pedestrian trajectory prediction and social robots navigating around humans, but does not aim to provide a comprehensive review of these domains. For further information, readers are encouraged to refer to (Singamaneni et al., 2024).

Lastly, some papers delve into the public's social perception, acceptance, and trust of AV technology, e.g., (Joo & Kim, 2023; Oliveira et al., 2019; Othman, 2021; Schneble & Shaw, 2021), recognising these aspects as critical for the broader adoption and integration of AVs into society. In particular, Joo and Kim (2023) conducted an online study to explore the influence of perceived collision algorithm types, i.e., selfish (prioritising passenger safety) versus utilitarian (minimising total damage by saving more lives, regardless of passenger status), and role of social approval of these algorithms on individuals' attitudes toward AVs. The study revealed a striking mismatch between societal and individual preferences. Participants rated utilitarian algorithms as more ethical and beneficial to society, aligning with broader social values. However, they expressed greater trust in, and a stronger personal preference for, selfish algorithms. Respondents were more willing to use and even pay a premium for AVs equipped with selfish algorithms, highlighting a significant divergence between ethical ideals and personal safety priorities. This discrepancy underscores the complexity of fostering public trust and acceptance of AV technology and suggests that designing and deploying SCAVs to balance societal ethics with individual user preferences is a crucial challenge for manufacturers and policymakers.

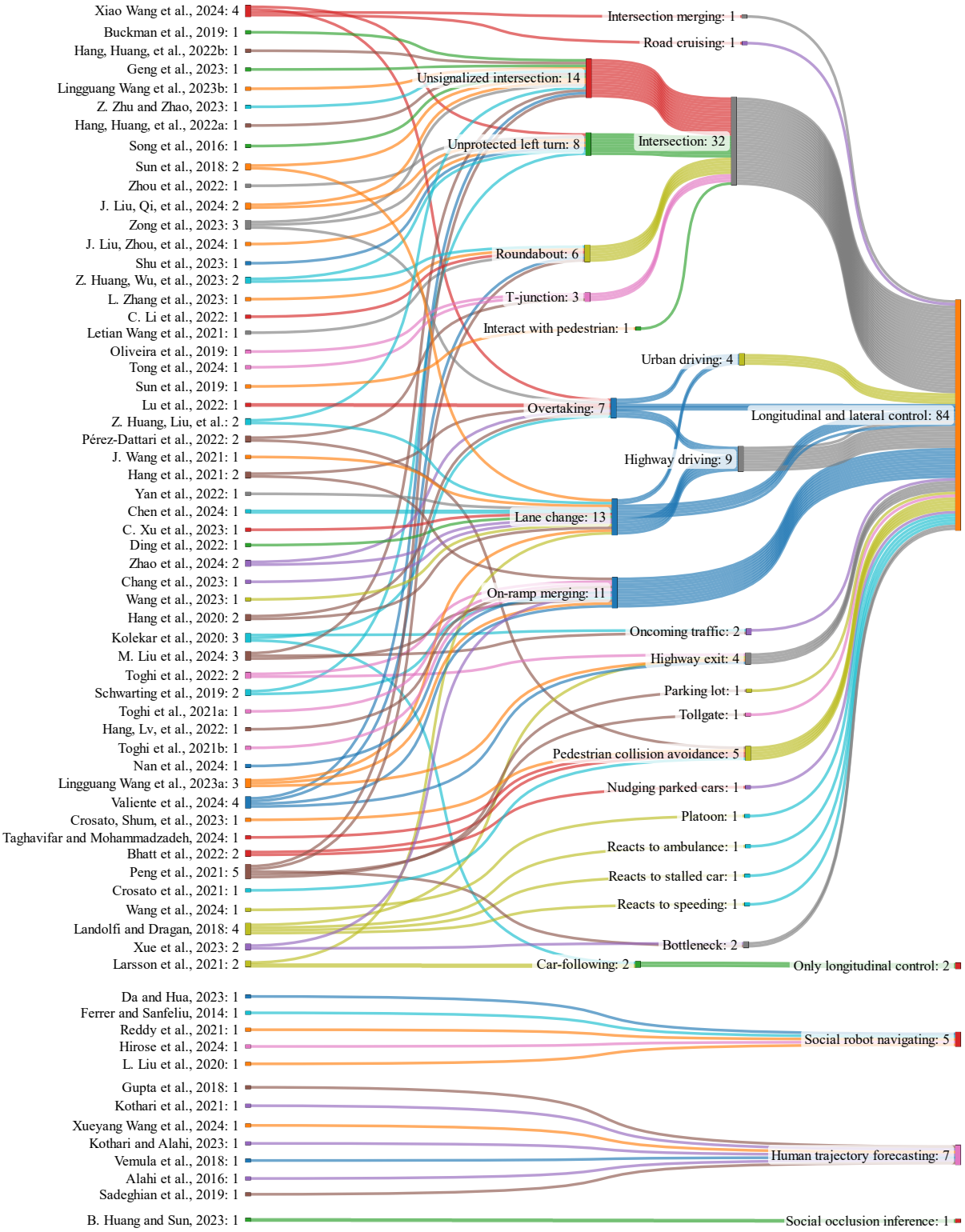


Figure 7-4. The identified involved manoeuvres in each study

Note: A single paper may involve multiple manoeuvres, thus the total number of manoeuvres can exceed the total number of reviewed papers (68).

7.3 Conceptual framework design

7.3.1 Expert interview

Building on the findings from the summarised literature review, an informal interview was conducted with ten experts representing diverse scientific and consultancy positions across research institutes, consulting firms, original equipment manufacturer (OEM) companies, and government sectors. The purpose of the interview was to gather expert perspectives through open-ended discussions on the current limitations of AVs, to further identify existing research gaps, and to understand their expectations for the development of SCAVs.

To facilitate insightful and meaningful discussions, the preliminary findings from the literature review were shared with the experts prior to the discussion. This ensured that the conversations were well-informed. The discussions were open-ended, allowing participants to elaborate on their views on the current limitations of AVs and provide in-depth observations on the challenges and opportunities in this field. The questions discussed include:

- *Do you have confidence in automated vehicles, particularly in mixed-traffic conditions?*
- *What are the current limitations and critical pain points of automated vehicles?*
- *Which scenarios do you perceive as particularly challenging for automated vehicles, and what scenarios, manoeuvres, or use cases would you like automated vehicles to address soon?*
- *What are your expectations for the short-term and long-term development of automated vehicles?*
- *What key efforts are necessary to drive the development and public acceptance of automated vehicles?*

Key insights derived from these expert interviews are summarised as follows:

Regarding the current practice and limitations of SCAVs, several critical shortcomings in the current generation of AVs were identified:

- ***Excessive Conservatism***: Most current AVs often adopt overly defensive driving strategies, which may significantly compromise traffic efficiency.
- ***Inability to Interpret Implicit Communications***: Most current AVs struggle to decode subtle signals to understand the implicit "communications" from human drivers, such as waving hands or a deceleration that implies yielding right of way.
- ***Challenges in Adapting to Various Driving Styles***: Most current AVs are unable to effectively adapt to the various driving styles, especially aggressive or assertive driving behaviours exhibited by surrounding HDVs.
- ***Limited Scenario Anticipation***: Unlike human drivers, current AVs lack robust capabilities to foresee, anticipate, and prepare for dynamic future scenarios.
- ***Cultural and Normative Inflexibility***: Current AVs are not yet designed to adapt their driving behaviours and styles to account for varying norms and driving cultures across different countries.

Regarding the research gaps and expectations, together with the literature review findings, the highlighted critical gaps and outlined priorities for advancing SCAVs are as follows:

- ***Integration of Sensing, Planning and Control***: Few studies connect AVs' sensing

capabilities, particularly considering sensor inaccuracies, to trajectory planning and control. Given the importance of this in real-world deployment, it warrants more in-depth exploration.

- ***Cultural and Normative Adaptation***: As limited research and development have incorporated cultural differences, driving norms, and implicit cues into automated driving models, this area deserves more attention.
- ***Development of AV Communication Pipelines***: There is a pressing need for AVs to express their intentions to other road users using, e.g., external human-machine interfaces (eHMI) such as colour-changing surfaces, signal lights, or LED panels on AVs.
- ***AV-Human Mutual Behavioural Adaptation***: The long-term and short-term adaptation of human drivers' behaviour when interacting with AVs and the corresponding adjustments AVs should make in response to those adaptations are seldom accounted for in the current development of AV driving models.
- ***Network-wide and Societal Benefits***: Few studies have considered the broader implications for overall network efficiency and societal benefits (e.g., total emissions across road networks) when deploying different AV driving strategies, styles, and behaviours.
- ***Interdisciplinary efforts***: Most research combines approaches from computer science, physics, mathematics, and engineering. Emerging efforts involving social psychology focus on adding concepts like SVO, coordination tendencies, and courtesy. More advanced frameworks incorporating social psychology and other interdisciplinary fields are needed to deepen the understanding of human-AV interactions.

These insights were the basis for the conceptual framework in the following *Section 7.3.2* to guide future research and development efforts in this area.

7.3.2 Proposed conceptual framework

Incorporating insights from the scoping review and addressing the identified gaps and research expectations from both the literature review and the expert interview, a conceptual framework, as illustrated in **Figure 7-5**, is proposed to guide future research and development on SCAVs.

Overall, this framework follows and adheres to the standard modular design for developing AVs, which includes sensing and perception modules, decision-making modules, planning modules, and control action modules. The differences and added values of the proposed conceptual framework are as follows:

a) **Socially Compliant Decision-Making Module**: The traditional decision-making module is enhanced and transformed into the proposed socially compliant decision-making module. This modification integrates social components (including culture, norms, and cues), which may influence implicit interactions, and consideration for various driving styles (e.g., aggressive, cautious, pro-social). The integration and embedding of these elements will help to address the aforementioned limitations of ***Cultural and Normative Inflexibility*** and ***Challenges in Adapting to Various Driving Styles***. Furthermore, the module incorporates mechanisms for bidirectional behavioural adaptation, enabling AVs to respond to human drivers' behavioural cues and adjust their responses accordingly, which will be illustrated later.

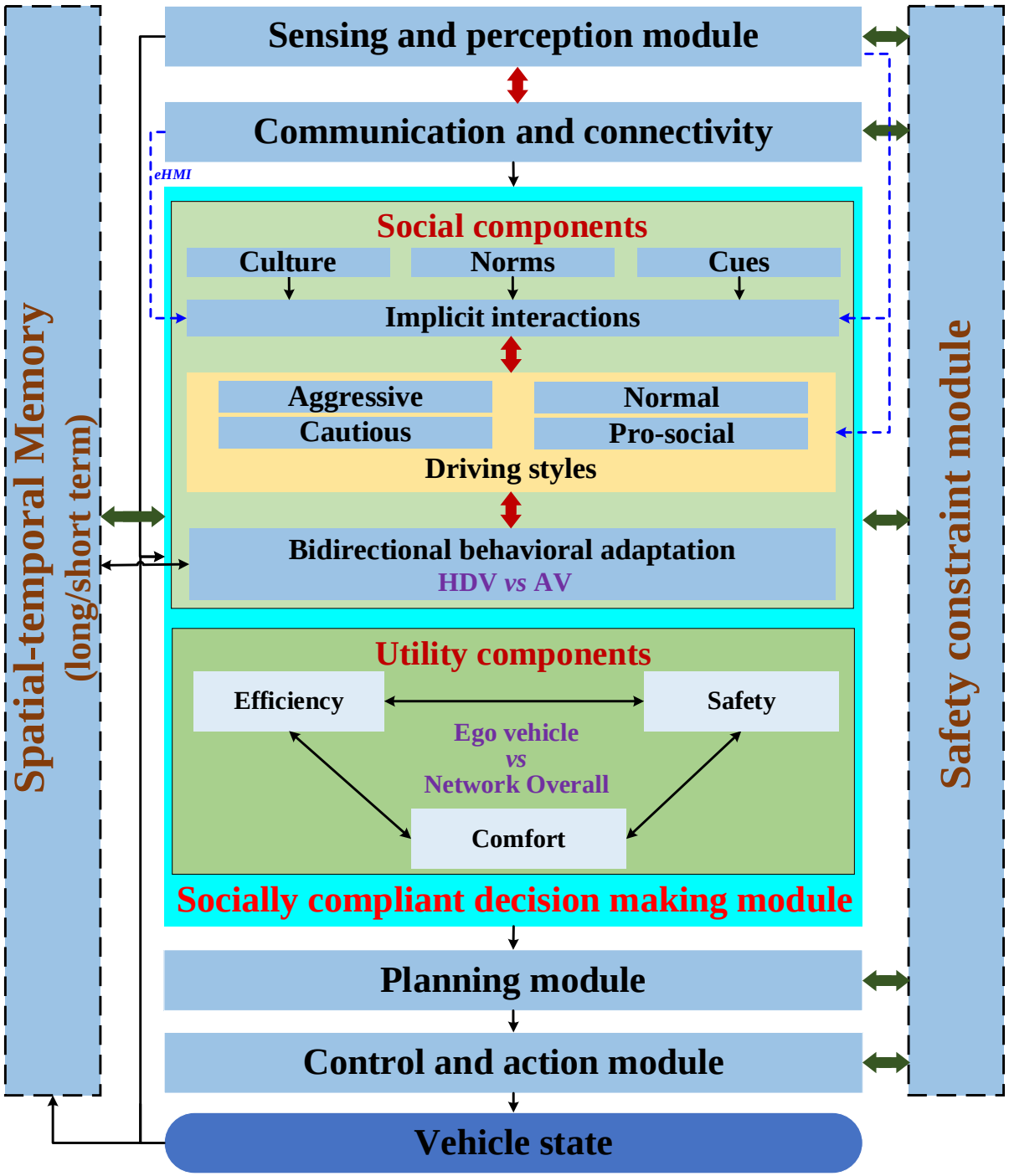


Figure 7-5. The proposed conceptual framework for developing SCAV

b) **Safety Constraint Module**: This module continuously monitors and enforces safety constraints to ensure that AVs operate within predefined safety boundaries. Although the socially compliant decision-making module should already incorporate safety metrics, the dedicated safety constraint module serves as a critical safeguard, ensuring that all actions taken by the AV are within the safety limits, thereby preventing undesirable outcomes. The planning module in this framework encompasses both high-level path planning and behaviour planning (e.g., lane changes, merging) as well as low-level motion planning (e.g., longitudinal and angular velocity, acceleration), all of which must adhere to the safety constraints outlined by this module.

c) **Trade-off between Ego and Network-Level Benefits**: A fundamental challenge (which is currently missing) in AV development is balancing the individual benefits of the ego vehicle (such as safety, comfort, and efficiency) with the broader benefits to the road network and other road users. The proposed framework emphasises the necessity of managing this trade-off (as shown in the Utility components within **Figure 7-5**), acknowledging that optimal performance for individual vehicles should not come at the expense of the overall network efficiency or societal benefits. It is suggested that this trade-off should be managed dynamically, on a case-by-case basis, to ensure a balanced approach that maximises both individual and collective outcomes (i.e., a more holistic, systems-level perspective). This requires close collaboration between AV developers, road operators, and regulatory authorities to align objectives and responsibilities. By managing the trade-off adaptively, this module will help meet the aforementioned expectation regarding ***Network-wide and Societal Benefits***.

d) **Bidirectional Behavioural Adaptation Module**: A key novel contribution of the proposed framework is the introduction of a bidirectional behavioural adaptation module. This module addresses the phenomenon where human drivers adapt their behaviour in response to the presence and actions of AVs in mixed traffic. For instance, drivers may exploit the defensive behaviour of AVs by engaging in more aggressive driving when interacting with them. To mitigate this, the AVs must adapt their behaviours in return, effectively responding to changes in human driving patterns and fostering a more balanced and cooperative interaction. The module is designed to facilitate a dynamic, iterative process of mutual adaptation, wherein both AVs and human drivers adjust their actions to optimise safety, traffic flow, and overall road network efficiency in mixed-traffic conditions. For successful real-world deployment, it is essential that the bidirectional behavioural adaptation module undergoes continuous updates, both in the short term and long term, to account for evolving traffic conditions and varied human driving behaviours. This ensures that the module remains responsive to a wide array of scenarios, thereby supporting the integration of AVs into diverse traffic contexts. This module will help to alleviate the aforementioned limitations of ***Excessive Conservatism*** and ***Challenges in Adapting to Various Driving Styles*** and help to meet the expectations of ***AV-Human Mutual Behavioural Adaptation***.

e) **Spatial-Temporal Memory Module**: The spatial-temporal memory module is designed to facilitate the long- and short-term updating of knowledge and driving rules, as well as to enhance the awareness of ongoing behavioural adaptations. This module enables AVs to incorporate historical interaction data and adapt their decision-making strategies over time. By maintaining a dynamic memory of past interactions, AVs can continuously refine their understanding of human-AV dynamics, ensuring that driving strategies incorporate lessons learned from prior experiences. This module is essential for the effective integration and implementation of bidirectional behavioural adaptation within the broader AV decision-making framework.

Explanations regarding the other remaining limitations, gaps, and expectations that are presented in *Section 7.3.1:*

The ***Limited Scenario Anticipation*** will be tackled by the *Sensing and Perception Module* as well as the *Communication and Connectivity Module*, which forms the backbone of the framework's ability to predict and respond to dynamic traffic scenarios. As illustrated by the

dashed arrows in **Figure 7-5**, the *Socially Compliant Decision-Making Module* depends on seamless integration with advanced sensing and communication systems. Sensing technologies, including cameras, LiDAR, and radar, deliver real-time data on the positions, velocities, and behavioural cues (e.g., accelerating, decelerating, and braking patterns) of surrounding road users. This data enables the *Socially Compliant Decision-Making Module* to interpret social norms, anticipate interactions, as well as estimate and adapt to diverse driving styles. Complementing this, communication and connectivity systems such as vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), vehicle-to-everything (V2X), and eHMI provide critical supplementary inputs, such as the signalled intentions or planned trajectories of other vehicles. Together, these systems enhance the AV's situational awareness, ensuring that social decisions are informed by a comprehensive understanding of the traffic environment, thereby mitigating ***Limited Scenario Anticipation*** and grounding the framework in operational reality.

The ***Inability to Interpret Implicit Communications*** can be alleviated through the proposed eHMI which connects the *Sensing and Perception Module* to the element of **Implicit Interactions** within the *Socially Compliant Decision-Making Module* (shown in blue text and dashed arrows in **Figure 7-5**). The eHMI allows AVs to convey their intentions (such as yielding or lane-changing) more effectively to other road users, facilitating mutual understanding and smoother interactions in mixed-traffic settings. This will also help meet expectations regarding the ***Development of AV Communication Pipelines***, fostering improved communication between AVs and surrounding HDVs, pedestrians, and cyclists.

Additionally, the ***Integration of Sensing, Planning, and Contro***l relies on an advanced, robust sensing and perception module capable of managing uncertainties and sensing failures. While such a module is integral to the framework's success, developing cutting-edge sensing and perception techniques exceeds the scope of this study and remains a broader research challenge itself, warranting further exploration.

Lastly, while vehicle connectivity, including V2V, V2I, and V2X communication with pedestrians, cyclists, and other road vehicles, plays a vital role in addressing several limitations and gaps, it is important to clarify that these aspects lie beyond the scope of this study. Connectivity is recognised as a crucial element in the broader ecosystem of autonomous driving, meriting its own dedicated line of research. Limited by space, this study could not delve deeply into that area.

7.4 Online questionnaire survey

To evaluate and verify the proposed framework for developing SCAVs, an online questionnaire-based survey was conducted. The survey was disseminated via targeted email distribution lists, including those of relevant expert groups such as the Universities' Transport Study Group (UTSG) and the TRAIL Research School. Additionally, the survey was actively promoted during key academic conferences, including the IEEE Intelligent Transportation Systems Conference (ITSC) and the IEEE Intelligent Vehicles Symposium (IV). The participants were asked to answer a sequence of questions, including multiple-choice questions, rank-order scale questions, rating scale questions, and open-ended questions. The questions are presented in seven subsections. The online survey takes approximately 15 minutes to fill out. The survey can be accessed at https://lnkd.in/evg6Dn9W. To promote experts' and professionals' participation in the survey, it was mentioned that every successful and qualified response would

result in a 5-euro donation to the United Nations Road Safety Fund (https://roadsafetyfund.un.org/). The survey questionnaire is provided in full for reference in Supplementary Attachment 2 at: https://lnkd.in/gpceU6gQ.

7.4.1 Respondents profile

A total of 99 responses were collected from experts across various nations and continents. 9 responses were excluded from the analysis due to contradictions in the answers or because the respondents self-identified as lacking confidence in their responses. Thus, 90 responses from experts were included in the final analysis. These experts represent a diverse range of roles in professional services, including researchers from universities, research institutes, and industry companies; developers from original equipment manufacturers (OEMs); policymakers; consultants; technicians; and professional drivers.

Figure 7-6 illustrates the distribution of respondents' profiles. The remaining 90 respondents were from 29 countries and across 6 continents. The majority of the respondents were from Europe and China, reflecting substantial representation in this study. Notably, only one respondent originates from the United States, a leading hub for AV technology development and deployment, resulting in its inclusion within the category of "Other" in **Figure 7-6 (a)**. Despite this limited presence of the United States, China's significant participation aligns with its own prominence in AV innovation and deployment, enriching the study with valuable insights from a key market.

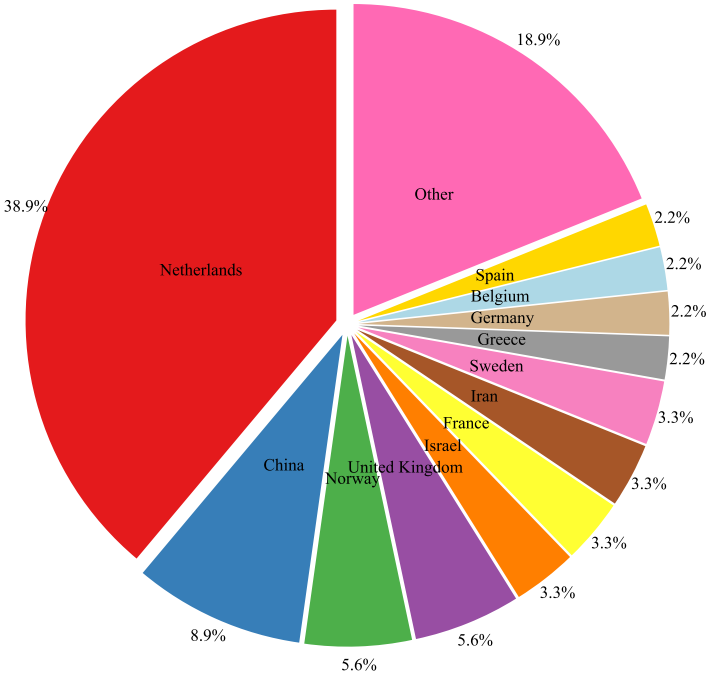
All 90 respondents claimed to be familiar with the concept and technology of automated vehicles, and more than half (54 out of the 90) of them are working in a field directly related to automated vehicles. Among them, 35 respondents are involved in developing AVs, 8 are engaged in testing automated driving functions, with 3 of them being qualified safety/test drivers, and 1 is researching human factors related to AVs.

In terms of professional roles, 49 respondents are researchers, 18 are consultants, 7 are policymakers, and 2 are developers or technicians at OEMs. Notably, one respondent claimed to be an associate editor for a relevant journal, one claimed to be responsible for the implementation of vehicle regulations by public authorities, and another one worked on the national strategy for the deployment of AVs. Furthermore, 86 out of the 90 respondents hold a driving license, with 6 claiming to have a professional driving qualification. These findings underscore the diverse expertise and perspectives that the respondents bring to the survey, enhancing the credibility of the survey results.

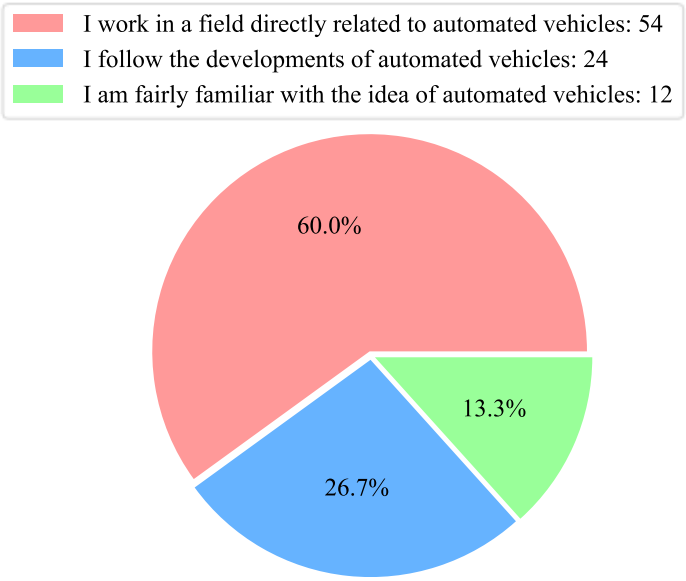
7.4.2 Benefits of SCAVs and willingness to purchase or use

Regarding the benefits of SCAVs, participants were asked to rate to what extent they think SCAVs will influence overall traffic safety and efficiency. The rating is based on a 7-point Likert scale with "-3" meaning strongly worsen; "0" standing for neutral/no influence; and "3" indicating strongly improve. As demonstrated in **Figure 7-7**, the majority of respondents believe that SCAVs contribute positively to both overall traffic safety and efficiency. The average rating for the potential improvement in safety is 1.04, while the average rating for efficiency is 0.54. These figures indicate that, on average, respondents perceive SCAVs as

having a greater potential to enhance safety than to improve efficiency, but both are seen as contributing positively.



(a)



(b)

Figure 7-6. The distribution of respondents' profiles: (a) residence countries, (b) familiarity with AV

Note: The "Other" category in (a) encompasses respondents from 17 countries (out of 29 total) not individually listed, including the United States, Canada, Australia, Italy, and India, among others beyond the 12 explicitly named nations (Netherlands, China, Norway, United Kingdom, Israel, France, Iran, Sweden, Greece, Germany, Belgium, Spain). Notably, the United States, a key AV technology hub, is grouped under "Other" due to its minimal representation (only one respondent post-preprocessing), insufficient for a distinct category.

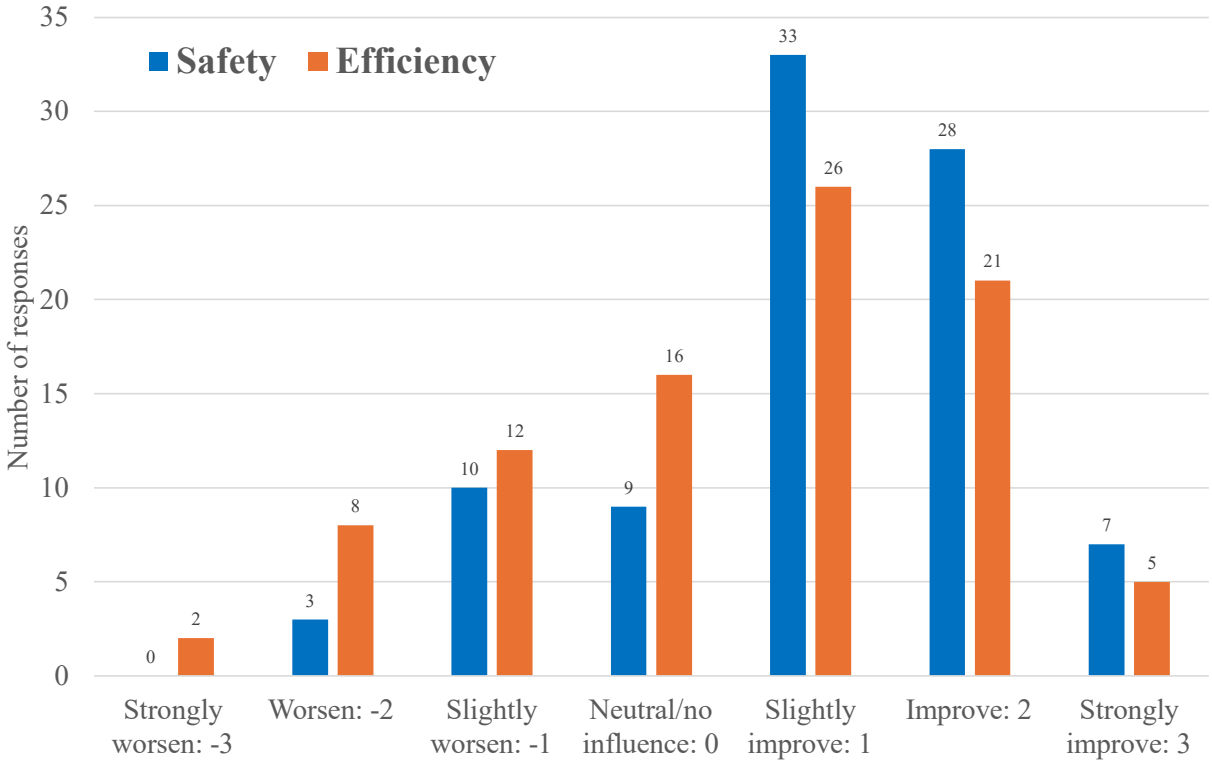


Figure 7-7. The rating distribution on to what extent the participants think SCAVs will influence overall traffic safety (light blue) and efficiency (orange)

Correspondingly, participants were asked about their willingness to purchase SCAVs when considering a vehicle purchase or their willingness to use them for on-demand mobility services during their travels. The majority responded positively, as shown in **Figure 7-8** and **Figure 7-9**. Specifically, 72 respondents indicated that they would like to buy an SCAV, while only 8 stated that they would never consider purchasing one, even if such AVs were cheaper.

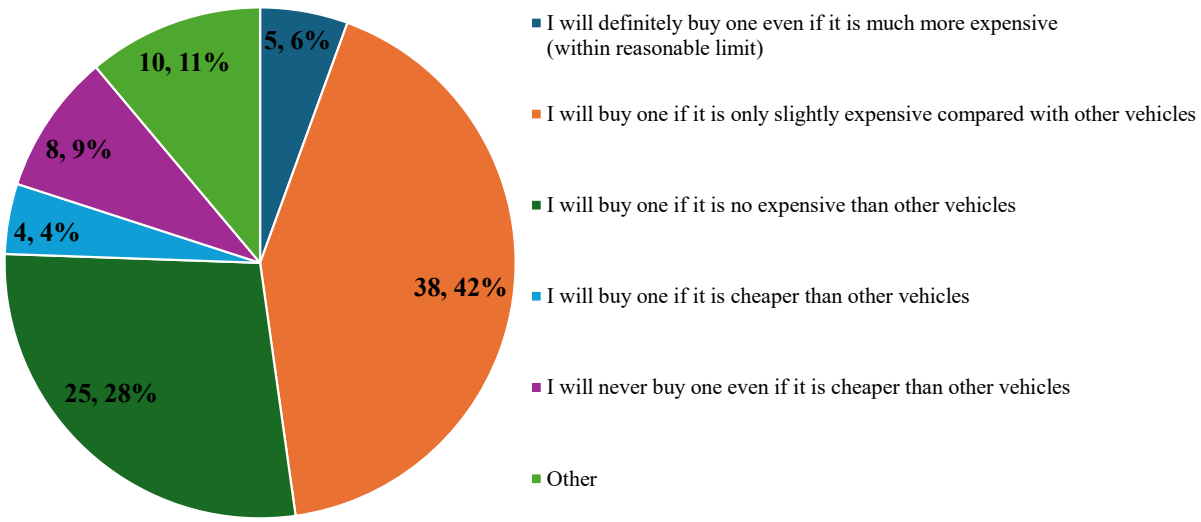


Figure 7-8. The distribution of willingness to buy one SCAV

Note: The "Other" category includes special responses beyond predefined options, e.g., preferences for cycling, walking, or public transit, explicitly rejecting car ownership.

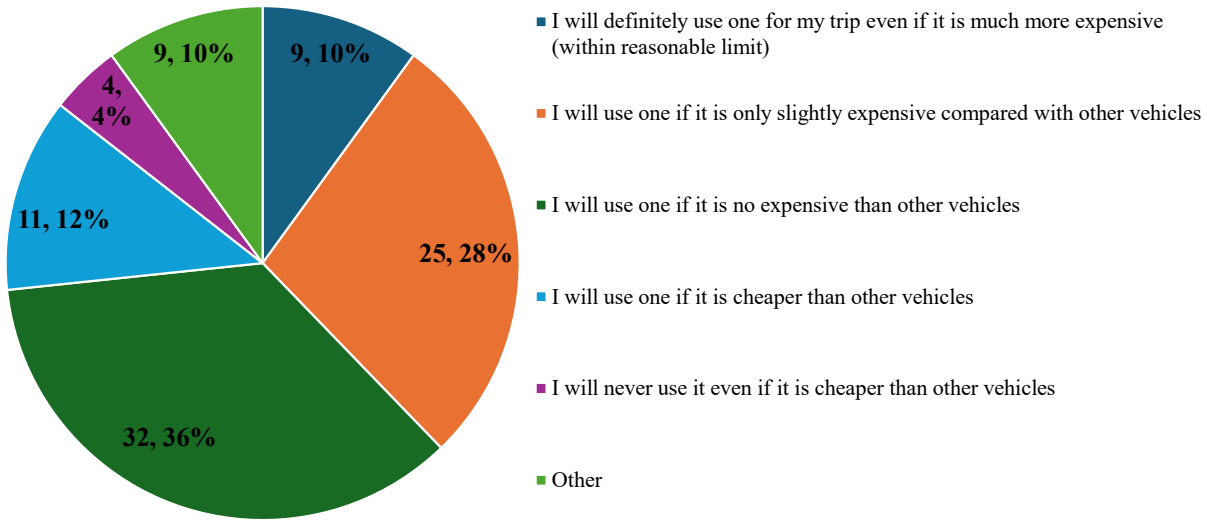


Figure 7-9. The distribution of willingness to use SCAVs for trips

Note: The "Other" category covers special responses outside listed options, e.g., bus travel, cycling, or car-sharing instead of car use.

Furthermore, acknowledging the suitability of SCAVs for shared on-demand travel services, 77 respondents expressed a willingness to use them for trips, while only 4 indicated that they would not use SCAVs, even if they were more affordable.

Notably, some participants emphasised that they prioritise functionality and performance over price, expressing a preference for public transport options that meet their specific needs; thus, they were categorised in the group of "Other".

7.4.3 Development of SCAVs

(1) Rating and ranking of the identified key technical capabilities

In the context of developing SCAVs, experts' opinions on the importance of various technical aspects required for AVs to exhibit socially compliant behaviours were assessed. Corresponding to the developed conceptual framework (**Figure 7-5**), respondents were asked to rate 9 key technical capabilities on a scale from 1 to 7, where 1 represented "Least Needed" and 7 represented "Strongly Needed." The evaluated technical aspects were:

- ***Anticipation Capability:*** The ability to anticipate the intended actions of other road users;
- ***eHMI Communication Capability:*** The ability to convey intended actions effectively through eHMI;
- ***Social and Cultural Alignment:*** The ability to adapt to different local cultures, social norms, and cues;
- ***User Acceptance:*** The ability to take consideration of acceptance levels among drivers, passengers, and nearby road users;
- ***Driving Style Adaptation:*** The ability to adjust to varying driving styles of surrounding human drivers, such as aggressive or defensive, and pro-social or egoistic;
- ***Bi-directional Behavioural Adaptation:*** The ability to enable mutual adaptation between AVs and human drivers over time;
- ***Multi-objective Optimisation:*** The ability to balance multiple goals such as safety, efficiency, energy consumption, and environmental impact;

- ***Trade-off Management:*** The ability to maintain trade-offs between the AV's benefits and those of surrounding traffic participants, between the ego AV's benefits and benefits at the network (regional) level;
- ***Spatial-temporal Memory Buffer Integration:*** Incorporating spatial-temporal memory buffers (short, medium, and long-term) to continually refine driving strategies.

As shown in **Figure 7-10**, respondents rated the extent to which they believe these properties should be integrated into SCAVs. All 9 key technical capabilities were rated as significant, with average ratings exceeding 4.8, which supports and verifies the elements proposed in the conceptual framework (**Figure 7-5**). Their ratings did not vary too much, with ***Anticipation Capability*** receiving the highest average rating (6.29), followed by the capabilities of ***Multi-objective Optimisation*** (5.76) and ***Trade-off Management*** (5.61).

As demonstrated in **Figure 7-11**, respondents were also asked to rank the top 3 most important aspects among 6 selected capabilities in the medium-term development (coming 1-3 years), supposing there are limited resources for developing SCAVs. The ranking results indicated that ***Anticipation Capability*** ranked first, followed by ***Multi-objective Optimisation***, which is consistent with the results shown in **Figure 7-10**.

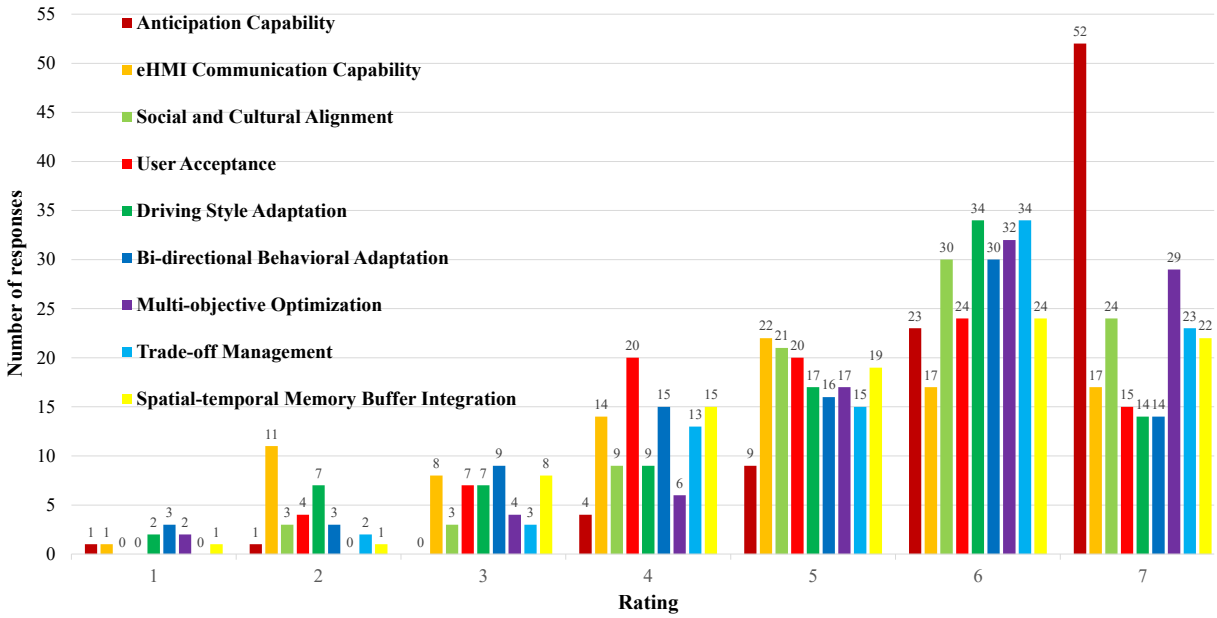
Furthermore, as illustrated in **Figure 7-12**, respondents were asked to rank the top 2 most important aspects among 4 selected capabilities for long-term development (in the coming 5-10 years or longer), again assuming limited resources for developing SCAVs. The results revealed that ***Bi-directional Behavioural Adaptation*** ranked first, followed by ***Spatial-temporal Memory Buffer Integration***, which is reasonable and aligns well with the proposed conceptual framework in **Figure 7-5**.

These ratings and rankings yield critical insights into which technical features are deemed essential and urgent for enabling AVs to navigate complex social interactions effectively. Such data-driven insights will be invaluable in guiding the prioritisation and future technical development of SCAVs.

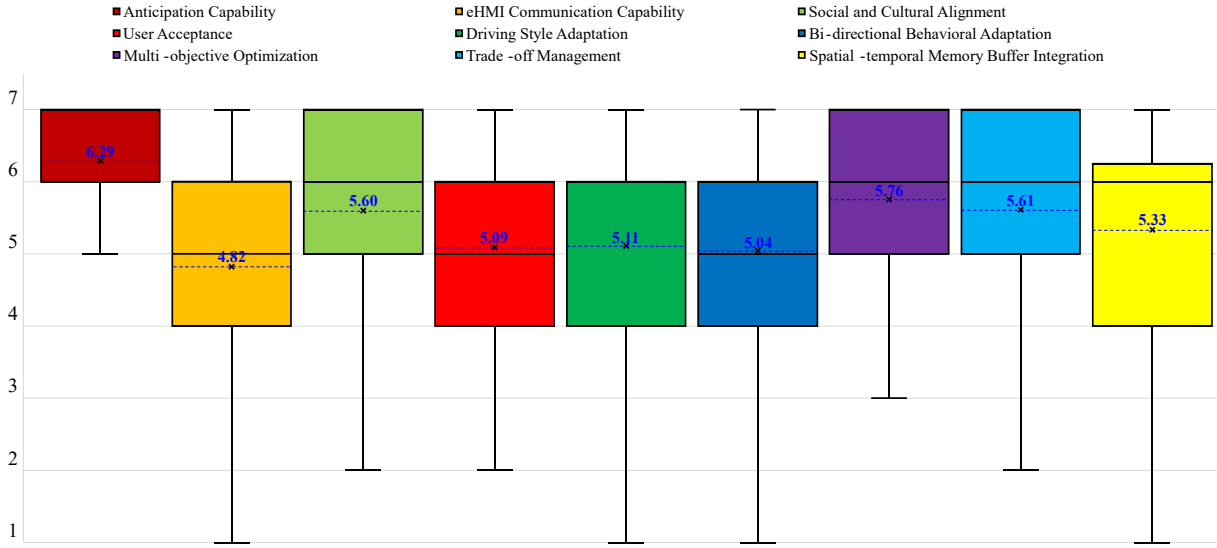
(2) Rating the possibility of mathematically modelling the identified key technical capabilities

Regarding the implementation of the identified key technical capabilities, the respondents were asked to rate the possibility and feasibility of mathematically modelling the six identified key technical capabilities of *Social and Cultural Alignment*, *Driving Style Adaptation*, *Bi-directional Behavioural Adaptation*, *Multi-objective Optimisation*, *Trade-off Management*, and *Spatial-temporal Memory Buffer Integration*. Ratings were provided on a scale from 1 to 7, where 1 represented "*Not Possible*" and 7 represented "*Highly Possible*". The results are depicted in **Figure 7-13**.

All the examined 6 key technical capabilities were found to be feasible for mathematical modelling. ***Multi-objective Optimisation*** was rated and deemed as the most feasible, followed by ***Trade-off Management***, which is expected, given that both of them could be modelled as typical optimisation problems. In contrast, ***Social and Cultural Alignment*** was identified as the most challenging and least feasible for mathematical modelling, followed by ***Bi-directional Behavioural Adaptation*** and ***Spatial-temporal Memory Buffer Integration***, which is also reasonable. This aligns with earlier recommendations for interdisciplinary cooperation, particularly drawing on knowledge and insights from social psychological domains alongside advancements in computer science.



(a)



(b)

Figure 7-10. Ratings on 9 key technical capabilities regarding their importance for developing SCAVs: (a) detailed rating distributions for each capability, (b) boxplot of the rating scales for each capability

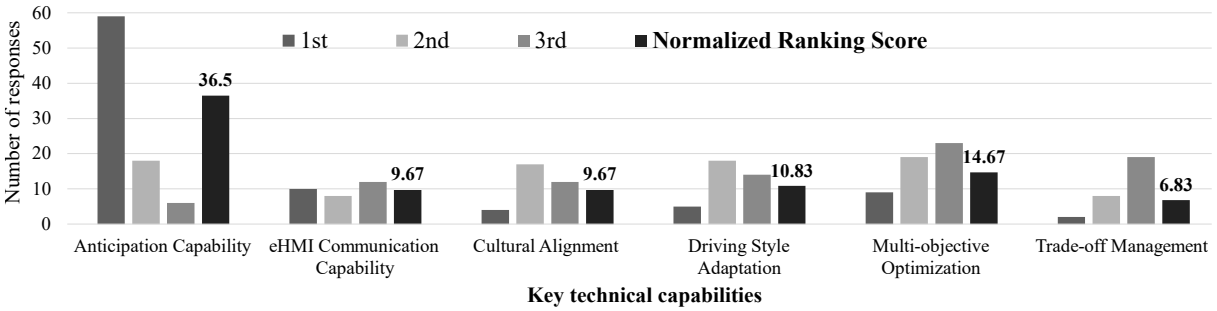


Figure 7-11. Ranking results for 6 selected technical capabilities regarding their priorities for developing SCAVs in the medium term

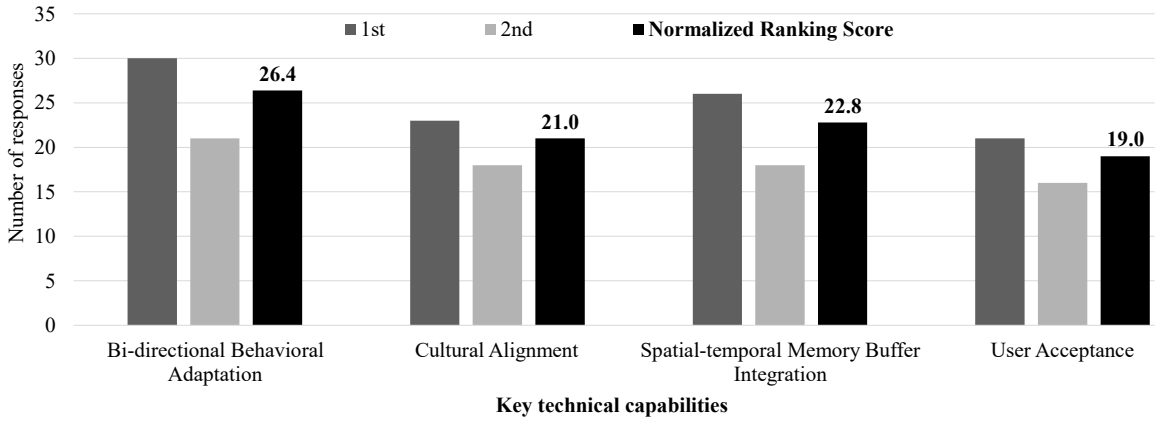
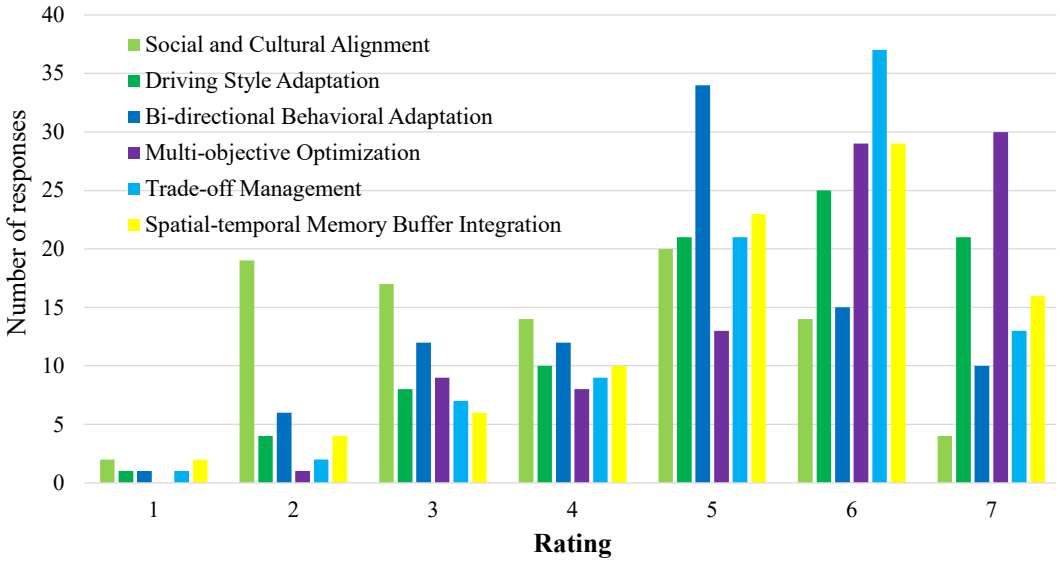
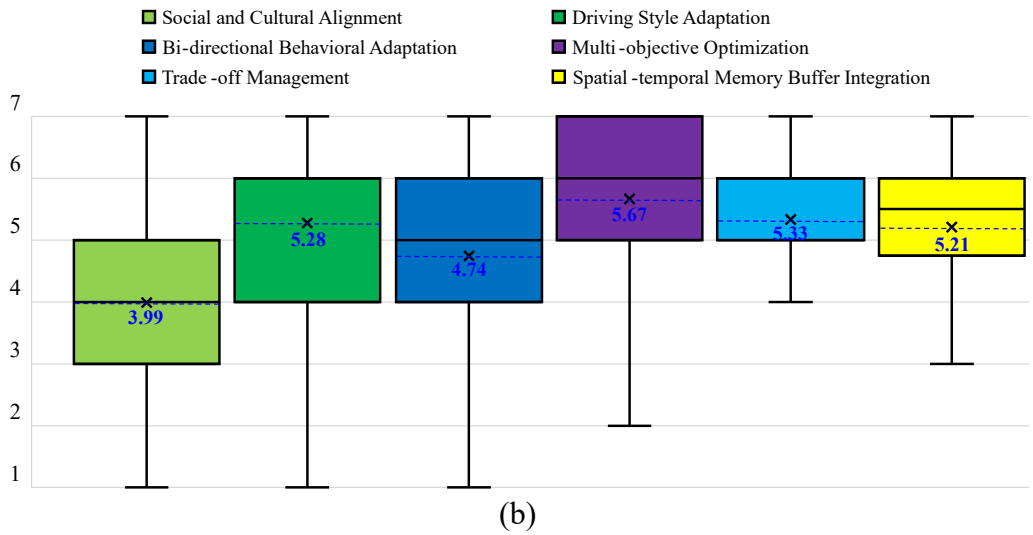


Figure 7-12. Ranking results for 4 selected technical capabilities regarding their priorities for developing SCAVs in the long term



(a)



(b)

Figure 7-13. Ratings on the feasibility of mathematically modelling the 6 identified key technical capabilities for developing SCAVs: (a) detailed rating distributions for each selected capability, (b) boxplot of the rating scales for each capability

(3) Suggestions from the respondents

Respondents were invited to share suggestions and insights through open-ended questions such as "*What else would you expect for the Socially Compliant Automated Vehicles?*" and "*Do you have any further comments for better development of Socially Compliant Automated Vehicles?*" A range of thoughtful responses was collected, which, after in-depth analysis, have been summarised, further upgraded, and polished as follows:

The development of SCAVs must prioritise **safety and trust** as core principles. Safety should remain paramount across all stages of development, and building trust between humans and SCAVs requires transparency, effective trust modelling, and clear communication of the vehicle's decision-making processes and intentions to its users and other road participants. Respondents emphasised the need for ML models to be trained using curated, unbiased datasets that reflect socially responsible driving behaviours rather than exceptional cases like those of professional drivers (e.g., F1 pilots). Additionally, initial deployment should focus on less complex environments, such as highways and provincial roads, before progressing to urban settings, where social compliance becomes more intricate and essential.

Infrastructure upgrades are also vital to support the successful deployment of SCAVs. This includes the development of dedicated AV lanes, vehicle-to-everything (V2X) communication networks, and robust systems with reliable backup mechanisms to prevent failures in smart traffic management systems. Respondents also highlighted the importance of balanced policy frameworks that encourage shared mobility solutions, such as controlled fleets of robotaxis, over private ownership of AVs. Collaboration among OEMs, regulators, and other stakeholders is deemed critical for fostering open communication, pooling knowledge, and advancing technical priorities strategically.

An **interdisciplinary and culturally sensitive approach** is required to reflect the diversity of societal needs in SCAV behaviours. Human factors must be central to design, ensuring that AVs can adapt to the social norms and behaviours of both drivers and other road users, such as cyclists and pedestrians, who are often overlooked. SCAVs should strike a balance between idealised performance and relatable, realistic behaviours that align with the imperfect nature of human driving.

Ultimately, the success of SCAVs hinges on the careful **prioritisation of technical and social efforts**, given the significant time and resources required for development. Transparent AI systems, robust infrastructure, and a focus on public acceptance and trustworthiness will be pivotal in ensuring SCAVs' seamless integration into society. These vehicles must not only navigate the immediate social and cultural contexts of their operation but also anticipate the long-term challenges of mixed-traffic environments and future scenarios dominated by automation. With thoughtful design and strategic planning, SCAVs can deliver safe, reliable, and socially aligned mobility solutions that meet the evolving needs of diverse communities.

Furthermore, as the deployment of AVs becomes increasingly widespread, a growing body of **empirical evidence** on real-world AV behaviour is emerging. This provides a valuable opportunity to investigate not only how AVs interact with human-driven vehicles but also how they respond to each other. Understanding interactions both within the same brand and between different brands of AVs is an area that remains underexplored but is critical for fostering interoperability, social compliance, and collaborative traffic systems. Such studies could reveal

how variations in algorithms, decision-making priorities, and communication protocols influence the dynamics of AV interactions. By fostering cross-brand standardisation and promoting cooperative driving behaviours among AVs, the industry can take a significant step toward realising the vision of a harmonised, intelligent transportation system that benefits all road users. Expanding research in this direction would further support the development of SCAVs that are not only socially compliant but also capable of thriving in increasingly complex and automated traffic environments.

7.5 Conclusion, limitation, and future research

This study represents the first comprehensive scoping review of the current state of the art in the development of socially compliant automated vehicles (SCAVs), systematically identifying key concepts, methodological approaches, and research gaps in the field. Through a rigorous review of existing literature and expert interviews, this study has elucidated critical pain points and research gaps while outlining vital research expectations essential for advancing SCAV development. Building on these insights, this study proposed a novel conceptual framework designed to address the multifaceted and interdisciplinary challenges of SCAVs in mixed-traffic environments. The framework outlines the key capability elements necessary for SCAVs and incorporates crucial considerations across technical, social, and cultural dimensions, effectively bridging theoretical insights with practical applications to achieve socially compliant automation.

To validate the conceptual framework, an online questionnaire-based survey was conducted, confirming the relevance of the framework's key elements and technical capabilities. Among these, ***Anticipation Capability*** emerged as the most significant and urgent requirement for medium-term implementation (1-3 years), reflecting its importance in enabling SCAVs to predict and adapt to dynamic road scenarios, especially regarding the interaction with HDVs. For long-term development (5-10 years or more), ***Bi-directional Behavioural Adaptation***—the ability to dynamically and mutually interact with and learn from other road users—and ***Spatial-Temporal Memory Buffer Integration*** were identified as the most critical priorities. These findings offer actionable insights for research and development (R&D) in both academia and industry, serving as a strategic roadmap for integrating social compliance into automated driving systems. They highlight research priorities and guide the creation of SCAVs that align with societal expectations. For researchers, the proposed conceptual framework identifies focus areas and key elements to be studied. For the industry, it provides actionable insights into developing and embedding social compliance in AV systems, enabling scalable and context-sensitive deployment. The developed framework can also foster collaboration among academia, industry, and policymakers, ensuring technical innovation aligns with societal needs and regulatory standards, accelerating the path toward SCAV and further towards safe and socially inclusive automated mobility solutions.

By providing a structured and interdisciplinary approach, this study contributes to the foundation of socially aware and ethically aligned AV technologies, laying the groundwork for safe, reliable, and socially compliant automated mobility solutions.

Despite its meaningful contributions, this study has several limitations that provide opportunities for further research. First, in the scoping review, as aforementioned, the scoping review did not analyse or summarise in detail the experiments, model performance, and results

from the reviewed studies. Furthermore, the study did not thoroughly investigate scenarios involving multi-vehicle interactions, particularly among multiple AVs. As AV penetration rates increase, understanding these interactions will become critical. Future reviews could address these gaps to provide a more comprehensive assessment of current research in this field.

Second, while the study emphasised the importance of anticipation capability, it did not extensively address its relationship with perception, particularly perception under uncertainty. This critical aspect, which includes managing ambiguous or incomplete information in real-world scenarios, represents a highly complex research domain that warrants dedicated research attention. Developing robust perception systems that can handle uncertainties will significantly enhance SCAVs' ability to navigate and interact socially in diverse and unpredictable environments. Similarly, connectivity, though recognised as an essential enabler, was not explored in depth. Future work could delve into the integration and benefits of vehicle-to-everything (V2X) technologies to support seamless communication between AVs, infrastructure, and road users for SCAV development.

Third, the study did not extensively examine interactions between AVs and vulnerable road users, such as cyclists and pedestrians. These interactions are crucial for ensuring SCAVs can operate safely and effectively in complex urban environments, where unpredictable behaviour from such road users often creates additional challenges. Addressing this limitation will not only enhance SCAVs' ability to anticipate and respond to the movements of vulnerable road users but also foster greater public trust and acceptance of AV technologies. Such efforts will help make SCAVs more inclusive and adaptable to diverse road user types, ultimately contributing to safer and more equitable urban mobility systems.

Additionally, while our framework is designed to be adaptable to mixed-traffic environments broadly, it does not explicitly investigate how social behaviours vary across specific settings such as urban, rural, and campus environments. Urban areas may require SCAVs to prioritise frequent, short-range interactions with diverse road users, whereas rural settings might involve adapting to less structured roads and unpredictable behaviours. Campus environments, with their mix of pedestrians, bicycles, and vehicles in confined spaces, could demand unique navigation strategies. Future research should explore these environmental differences to refine and validate our framework, tailoring social compliance strategies to context-specific challenges and enhancing the generalisability of SCAVs.

Moreover, the geographic distribution of respondents in the online survey of this study is predominantly concentrated in Europe and China, with only one participant from the United States. This imbalanced representation may introduce potential cultural and contextual biases into the study's findings. Social compliance in driving behaviours is influenced by regional norms, regulations, and infrastructure designs, meaning the current sample may not fully reflect global perspectives. Although China's substantial participation aligns with its prominence in AV development and deployment, the near absence of respondents from the United States, another global leader in this domain, may underrepresent critical insights from a major AV market, thereby limiting the findings' generalisability. This imbalance could particularly affect perceptions of social compliance expectations across diverse regional contexts. Future research should prioritise a more balanced sample, increasing representation from key AV markets like the United States to encompass diverse technological and cultural viewpoints.

Finally, this study does not address the broader systemic challenges of integrating SCAVs into existing infrastructure and ecosystems. Factors such as regulatory alignment, public acceptance, and economic feasibility remain critical to the successful deployment of SCAVs and must be explored further. In particular, balancing the needs of private and shared ownership models, addressing the environmental impact of SCAVs, and mitigating potential socioeconomic disparities should form part of future interdisciplinary research efforts.

As for future research, a significant barrier to SCAV research is the scarcity of real-world field data, which restricts much of the current literature to simulation-based methodologies. Although simulations provide a controlled setting for modelling social compliance, they struggle to capture the full spectrum of human unpredictability. This shortfall limits the validation of SCAV frameworks in authentic mixed-traffic contexts, potentially leading to overestimated performance and overlooked vulnerabilities. Overcoming this requires prioritising empirical field data collection, potentially through partnerships with AV testing initiatives (e.g., industry-led trials or regulatory pilot programs) or by harnessing data from controlled urban deployments. While such endeavours are resource-intensive, they are crucial for transitioning SCAV solutions from theoretical constructs to reliable, real-world applications, thereby significantly enhancing their practical robustness and relevance.

In conclusion, while this study provides a valuable foundation for SCAV development, it highlights the complexity and interdisciplinary nature of the challenges ahead. By addressing the identified limitations and advancing research in these critical areas, future efforts can build on the insights and framework presented here to create SCAV systems that are not only technically advanced but also socially responsible and globally inclusive.

Acknowledgements

This work was supported by the Applied and Technical Sciences (TTW), a subdomain of the Dutch Institute for Scientific Research (NWO) through the Project Safe and Efficient Operation of Automated and Human-Driven Vehicles in Mixed Traffic (SAMEN) under Contract 17187.

References

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2016.110

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. International Journal of Social Research Methodology: Theory and Practice, 8(1), 19–32. https://doi.org/10.1080/1364557032000119616

Benrachou, D. E., Glaser, S., Elhenawy, M., & Rakotonirainy, A. (2022). Use of social interaction and intention to improve motion prediction within automated vehicle framework: A review. IEEE Transactions on Intelligent Transportation Systems, 23(12), 22807–22837. https://doi.org/10.1109/TITS.2022.3207347

Bhatt, N. P., Khajepour, A., & Hashemi, E. (2022). MPC-PF: Social interaction aware trajectory prediction of dynamic objects for autonomous driving using potential fields. 2022

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), i, 9837–9844. https://doi.org/10.1109/iros47612.2022.9981046

Bock, J., Krajewski, R., Moers, T., Runde, S., Vater, L., & Eckstein, L. (2020). The inD dataset: A drone dataset of naturalistic road user trajectories at German intersections. IEEE Intelligent Vehicles Symposium, Proceedings. https://doi.org/10.1109/IV47402.2020.9304839

Buckman, N., Pierson, A., Schwarting, W., Karaman, S., & Rus, D. (2019). Sharing is caring: Socially-compliant autonomous intersection negotiation. IEEE International Conference on Intelligent Robots and Systems, 6136–6143. https://doi.org/10.1109/IROS40897.2019.8967997

Chang, M. F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., & Hays, J. (2019). Argoverse: 3D tracking and forecasting with rich maps. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2019.00895

Chang, W. J., Tang, C., Li, C., Hu, Y., Tomizuka, M., & Zhan, W. (2023). Editing driver character: socially-controllable behavior generation for interactive traffic simulation. IEEE Robotics and Automation Letters, 8(9), 5432–5439. https://doi.org/10.1109/LRA.2023.3291897

Chen, X., Zhang, W., Bai, H., Xu, C., Ding, H., & Huang, W. (2024). Two-dimensional following lane-changing (2DF-LC): A framework for dynamic decision-making and rapid behavior planning. IEEE Transactions on Intelligent Vehicles, 9(1), 427–445. https://doi.org/10.1109/TIV.2023.3324305

Crosato, L., Shum, H. P. H., Ho, E. S. L., & Wei, C. (2023). Interaction-aware decision-making for automated vehicles using social value orientation. IEEE Transactions on Intelligent Vehicles, 8(2), 1339–1349. https://doi.org/10.1109/TIV.2022.3189836

Crosato, L., Tian, K., Shum, H. P. H., Ho, E. S. L., Wang, Y., & Wei, C. (2023). Social interaction-aware dynamical models and decision-making for autonomous vehicles. Advanced Intelligent Systems, 2300575. https://doi.org/10.1002/aisy.202300575

Crosato, L., Wei, C., Ho, E. S. L., & Shum, H. P. H. (2021). Human-centric autonomous driving in an av-pedestrian interactive environment using SVO. Proceedings of the 2021 IEEE International Conference on Human-Machine Systems, ICHMS 2021, 1–6. https://doi.org/10.1109/ICHMS53169.2021.9582640

Da, L., & Hua, W. (2023). CrowdGAIL: A spatiotemporal aware method for agent navigation. Electronic Research Archive, 31(2), 1134–1146. https://doi.org/10.3934/era.2023057

Ding, W., Zhang, L., Chen, J., & Shen, S. (2022). EPSILON: An efficient planning system for automated vehicles in highly interactive environments. IEEE Transactions on Robotics, 38(2), 1118–1138. https://doi.org/10.1109/TRO.2021.3104254

Dong, Y., Liu, C., Wang, Y., & Fu, Z. (2024). Towards understanding worldwide cross-cultural differences in implicit driving cues: Review, comparative analysis, and research roadmap. 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), Edmonton, AB, Canada, 2024, pp. 1569-1575. http://dx.doi.org/10.1109/ITSC58415.2024.10919561

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. CoRL, 1–16. http://arxiv.org/abs/1711.03938

Du, Y., Chen, J., Zhao, C., Liu, C., Liao, F., & Chan, C. Y. (2022). Comfortable and energy-efficient speed control of autonomous vehicles on rough pavements using deep reinforcement learning. Transportation Research Part C: Emerging Technologies, 134(December 2021), 103489. https://doi.org/10.1016/j.trc.2021.103489

ElSamadisy, O., Shi, T., Smirnov, I., & Abdulhai, B. (2024). Safe, efficient, and comfortable reinforcement-learning-based car-following for AVs with an analytic safety guarantee and dynamic target speed. Transportation Research Record, 2678(1), 643–661. https://doi.org/10.1177/03611981231171899

Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C., Zhou, Y., Yang, Z., Chouard, A., Sun, P., Ngiam, J., Vasudevan, V., McCauley, A., Shlens, J., & Anguelov, D. (2021). Large scale interactive motion forecasting for autonomous driving: The Waymo Open Motion Dataset. Proceedings of the IEEE International Conference on Computer Vision. https://doi.org/10.1109/ICCV48922.2021.00957

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. Transportation Research Part A: Policy and Practice, 77, 167–181. https://doi.org/10.1016/j.tra.2015.04.003

Ferrer, G., & Sanfeliu, A. (2014). Proactive kinodynamic planning using the Extended Social Force Model and human motion prediction in urban environments. IEEE International Conference on Intelligent Robots and Systems, IROS, 1730–1735. https://doi.org/10.1109/IROS.2014.6942788

Fraedrich, E., Beiker, S., & Lenz, B. (2015). Transition pathways to fully automated driving and its implications for the sociotechnical system of automobility. European Journal of Futures Research, 3(1). https://doi.org/10.1007/s40309-015-0067-8

Galati, G., Primatesta, S., Grammatico, S., Macrì, S., & Rizzo, A. (2022). Game theoretical trajectory planning enhances social acceptability of robots by humans. Scientific Reports, 12(1), 1–18. https://doi.org/10.1038/s41598-022-25438-1

Geng, M., Cai, Z., Zhu, Y., Chen, X., & Lee, D. H. (2023). Multimodal vehicular trajectory prediction with inverse reinforcement learning and risk aversion at urban unsignalized intersections. IEEE Transactions on Intelligent Transportation Systems, 24(11), 12227–12240. https://doi.org/10.1109/TITS.2023.3285891

Greenblatt, J. B., & Shaheen, S. (2015). Automated vehicles, on-demand mobility, and environmental impacts. Current Sustainable/Renewable Energy Reports, 2(3), 74–81. https://doi.org/10.1007/s40518-015-0038-5

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2255–2264. https://doi.org/10.1109/CVPR.2018.00240

Hang, P., Huang, C., Hu, Z., & Lv, C. (2022a). Decision making for connected automated vehicles at urban intersections considering social and individual benefits. IEEE Transactions on Intelligent Transportation Systems, 23(11), 22549–22562. https://doi.org/10.1109/TITS.2022.3209607

Hang, P., Huang, C., Hu, Z., & Lv, C. (2022b). Driving conflict resolution of autonomous vehicles at unsignalized intersections: A differential game approach. IEEE/ASME Transactions on Mechatronics, 27(6), 5136–5146. https://doi.org/10.1109/TMECH.2022.3174273

Hang, P., Lv, C., Huang, C., Cai, J., Hu, Z., & Xing, Y. (2020). An integrated framework of decision making and motion planning for autonomous vehicles considering social behaviors. IEEE Transactions on Vehicular Technology, 69(12), 14458–14469. https://doi.org/10.1109/TVT.2020.3040398

Hang, P., Lv, C., Huang, C., Xing, Y., & Hu, Z. (2022). Cooperative decision making of connected automated vehicles at multi-lane merging zone: A coalitional game approach. IEEE Transactions on Intelligent Transportation Systems, 23(4), 3829–3841. https://doi.org/10.1109/TITS.2021.3069463

Hang, P., Lv, C., Huang, C., Xing, Y., Hu, Z., & Cai, J. (2020). Human-like lane-change decision making for automated driving with a game theoretic approach. 2020 4th CAA International Conference on Vehicular Control and Intelligence, CVCI 2020, 708–713. https://doi.org/10.1109/CVCI51460.2020.9338614

Hang, P., Lv, C., Xing, Y., Huang, C., & Hu, Z. (2021). Human-like decision making for autonomous driving: A noncooperative game theoretic approach. IEEE Transactions on Intelligent Transportation Systems, 22(4), 2076–2087. https://doi.org/10.1109/TITS.2020.3036984

Hirose, N., Shah, D., Sridhar, A., & Levine, S. (2024). SACSoN: Scalable autonomous control for social navigation. IEEE Robotics and Automation Letters, 9(1), 49–56. https://doi.org/10.1109/LRA.2023.3329626

Huang, B., & Sun, P. (2023). Social occlusion inference with vectorized representation for autonomous driving. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, 2634–2639. https://doi.org/10.1109/SMC53992.2023.10394619

Huang, Z., Liu, H., Wu, J., & Lv, C. (2023). Conditional predictive behavior planning with inverse reinforcement learning for human-like autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 24(7), 7244–7258. https://doi.org/10.1109/TITS.2023.3254579

Huang, Z., Wu, J., & Lv, C. (2023). Efficient deep reinforcement learning with imitative expert priors for autonomous driving. IEEE Transactions on Neural Networks and Learning Systems, 34(10), 7391–7403. https://doi.org/10.1109/TNNLS.2022.3142822

Jamson, H., Merat, N., Carsten, O., & Lai, F. (2011). Fully-automated driving: The road to future vehicles. In Driving Assessment Conference (Vol. 6, No. 2011). University of Iowa. https://doi.org/10.17077/drivingassessment.1370

Joo, Y. K., & Kim, B. (2023). Selfish but socially approved: The effects of perceived collision algorithms and social approval on attitudes toward autonomous vehicles. International Journal of Human-Computer Interaction, 39(19), 3717–3727. https://doi.org/10.1080/10447318.2022.2102716

Kolekar, S., de Winter, J., & Abbink, D. (2020). Human-like driving behaviour emerges from a risk-based driver model. Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-18353-4

Kothari, P., & Alahi, A. (2023). Safety-compliant generative adversarial networks for human trajectory forecasting. IEEE Transactions on Intelligent Transportation Systems, 24(4), 4251–4261. https://doi.org/10.1109/TITS.2022.3233906

Kothari, P., Kreiss, S., & Alahi, A. (2022). Human trajectory forecasting in crowds: A deep learning perspective. IEEE Transactions on Intelligent Transportation Systems, 23(7), 7386-7400. https://doi.org/10.1109/TITS.2021.3069362

Kothari, P., Sifringer, B., & Alahi, A. (2021). Interpretable social anchors for human trajectory forecasting in crowds. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 15551–15561. https://doi.org/10.1109/CVPR46437.2021.01530

Krajewski, R., Bock, J., Kloeker, L., & Eckstein, L. (2018). The highD dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2018-Novem, 2118–2125. https://doi.org/10.1109/ITSC.2018.8569552

Krajewski, R., Moers, T., Bock, J., Vater, L., & Eckstein, L. (2020). The rounD dataset: A drone dataset of road user trajectories at roundabouts in Germany. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC 2020. https://doi.org/10.1109/ITSC45102.2020.9294728

Landolfi, N. C., & Dragan, A. D. (2018). Social cohesion in autonomous driving. IEEE International Conference on Intelligent Robots and Systems, 8118–8125. https://doi.org/10.1109/IROS.2018.8593682

Larsson, J., Keskin, M. F., Peng, B., Kulcsár, B., & Wymeersch, H. (2021). Pro-social control of connected automated vehicles in mixed-autonomy multi-lane highway traffic. Communications in Transportation Research, 1, 100019. https://doi.org/10.1016/j.commtr.2021.100019

Lerner, A., Chrysanthou, Y., & Lischinski, D. (2007). Crowds by example. Computer Graphics Forum. (Vol. 26, No. 3, pp. 655-664). Oxford, UK: Blackwell Publishing Ltd. https://doi.org/10.1111/j.1467-8659.2007.01089.x

Leurent, E. (2018). An environment for autonomous driving decision-making. Accessed 2024-05-09 from https://github.com/eleurent/highway-env

Li, C., Trinh, T., Wang, L., Liu, C., Tomizuka, M., & Zhan, W. (2022). Efficient game-theoretic planning with prediction heuristic for socially-compliant autonomous driving. IEEE Robotics and Automation Letters, 7(4), 10248–10255. https://doi.org/10.1109/LRA.2022.3191241

Li, Q., Peng, Z., Feng, L., Zhang, Q., Xue, Z., & Zhou, B. (2023). MetaDrive: Composing diverse driving scenarios for generalizable reinforcement learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 3461–3475. https://doi.org/10.1109/TPAMI.2022.3190471

Liebrand, W. B. G., & McClintock, C. G. (1988). The ring measure of social values: A computerized procedure for assessing individual differences in information processing and social value orientation. European Journal of Personality. https://doi.org/10.1002/per.2410020304

Liu, J., Qi, X., Hang, P., & Sun, J. (2024). Enhancing social decision-making of autonomous vehicles: A mixed-strategy game approach with interaction orientation identification. IEEE Transactions on Vehicular Technology, PP, 1–14. https://doi.org/10.1109/TVT.2024.3385750

Liu, J., Zhou, D., Hang, P., Ni, Y., & Sun, J. (2024). Towards socially responsive autonomous vehicles: A reinforcement learning framework with driving priors and coordination awareness. IEEE Transactions on Intelligent Vehicles, 9(1), 827–838. https://doi.org/10.1109/TIV.2023.3332080

Liu, L., Dugas, D., Cesari, G., Siegwart, R., & Dube, R. (2020). Robot navigation in crowded environments using deep reinforcement learning. IEEE International Conference on Intelligent Robots and Systems, 5671–5677. https://doi.org/10.1109/IROS45743.2020.9341540

Liu, M., Tseng, H. E., Filev, D., Girard, A., & Kolmanovsky, I. (2024). Safe and human-like autonomous driving: A predictor-corrector potential game approach. IEEE Transactions on Control Systems Technology, 32(3), 834–848. https://doi.org/10.1109/TCST.2023.3332438

Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flotterod, Y. P., Hilbrich, R., Lucken, L., Rummel, J., Wagner, P., & Wiebner, E. (2018). Microscopic traffic simulation using SUMO. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2018-Novem, 2575–2582. https://doi.org/10.1109/ITSC.2018.8569938

Lu, H., Lu, C., Yu, Y., Xiong, G., & Gong, J. (2022). Autonomous overtaking for intelligent vehicles considering social preference based on hierarchical reinforcement learning. Automotive Innovation, 5(2), 195–208. https://doi.org/10.1007/s42154-022-00177-1

Moers, T., Vater, L., Krajewski, R., Bock, J., Zlocki, A., & Eckstein, L. (2022). The exiD dataset: A real-world trajectory dataset of highly interactive highway scenarios in Germany. IEEE Intelligent Vehicles Symposium, Proceedings. https://doi.org/10.1109/IV51971.2022.9827305

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Medical Research Methodology. https://doi.org/10.1186/s12874-018-0611-x

Murphy, R. O., & Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. Personality and Social Psychology Review, 18(1), 13–41. https://doi.org/10.1177/1088868313501745

Nan, J., Deng, W., Zhang, R., Wang, Y., Zhao, R., & Ding, J. (2024). Interaction-aware planning with deep inverse reinforcement learning for human-like autonomous driving in merge scenarios. IEEE Transactions on Intelligent Vehicles, 9(1), 2714–2726. https://doi.org/10.1109/TIV.2023.3298912

Oliveira, L., Proctor, K., Burns, C. G., & Birrell, S. (2019). Driving style: How should an automated vehicle behave? Information (Switzerland), 10(6), 1–20. https://doi.org/10.3390/INFO10060219

Orieno, O. H., Ndubuisi, N. L., Ilojianya, V. I., Biu, P. W., & Odonkor, B. (2024). The future of autonomous vehicles in the U.S. urban landscape: A review: Analyzing implications for traffic, urban planning, and the environment. Engineering Science & Technology Journal, 5(1), 43–64. https://doi.org/10.51594/estj.v5i1.721

Othman, K. (2021). Public acceptance and perception of autonomous vehicles: a comprehensive review. In AI and Ethics (Vol. 1, Issue 3). Springer International Publishing. https://doi.org/10.1007/s43681-021-00041-8

Pellegrini, S., Ess, A., Schindler, K., & Van Gool, L. (2009). You'll never walk alone: Modeling social behavior for multi-target tracking. Proceedings of the IEEE International Conference on Computer Vision. https://doi.org/10.1109/ICCV.2009.5459260

Peng, Z., Li, Q., Hui, K. M., Liu, C., & Zhou, B. (2021). Learning to simulate self-driven particles system with coordinated policy optimization. Advances in Neural Information Processing Systems, 13(NeurIPS), 10784–10797.

Pérez-Dattari, R., Brito, B., de Groot, O., Kober, J., & Alonso-Mora, J. (2022). Visually-guided motion planning for autonomous driving from interactive demonstrations. Engineering Applications of Artificial Intelligence, 116(August), 105277. https://doi.org/10.1016/j.engappai.2022.105277

Qin, L., Huang, Z., Zhang, C., Guo, H., Ang, M., & Rus, D. (2021). Deep imitation learning for autonomous navigation in dynamic pedestrian environments. Proceedings - IEEE International Conference on Robotics and Automation, 2021-May(ICRA), 4108–4115. https://doi.org/10.1109/ICRA48506.2021.9561220

Raju, N., Schakel, W., Reddy, N., Dong, Y., & Farah, H. (2022). Car-following properties of a commercial adaptive cruise control system: A pilot field test. In Transportation Research Record. https://doi.org/10.1177/03611981221077085

Reddy, A. K., Malviya, V., & Kala, R. (2021). Social cues in the autonomous navigation of indoor mobile robots. International Journal of Social Robotics, 13(6), 1335–1358. https://doi.org/10.1007/s12369-020-00721-1

Robicquet, A., Sadeghian, A., Alahi, A., & Savarese, S. (2016). Learning social etiquette: Human trajectory understanding in crowded scenes. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-319-46484-8_33

Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., & Savarese, S. (2019). SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June, 1349–1358. https://doi.org/10.1109/CVPR.2019.00144

Schneble, C. O., & Shaw, D. M. (2021). Driver's views on driverless vehicles: Public perspectives on defining and using autonomous cars. Transportation Research Interdisciplinary Perspectives, 11, 100446. https://doi.org/10.1016/j.trip.2021.100446

Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., & Rus, D. (2019). Social behavior for autonomous vehicles. Proceedings of the National Academy of Sciences of the United States of America, 116(50), 2492–24978. https://doi.org/10.1073/pnas.1820676116

Shu, K., Mehrizi, R. V., Li, S., Pirani, M., & Khajepour, A. (2023). Human inspired autonomous intersection handling using game theory. IEEE Transactions on Intelligent Transportation Systems, 24(10), 11360–11371. https://doi.org/10.1109/TITS.2023.3281390

Singamaneni, P. T., Bachiller-Burgos, P., Manso, L. J., Garrell, A., Sanfeliu, A., Spalanzani, A., & Alami, R. (2024). A survey on socially aware robot navigation: Taxonomy and future challenges. International Journal of Robotics Research, 0(0), 1–40. https://doi.org/10.1177/02783649241230562

Song, W., Xiong, G., & Chen, H. (2016). Intention-aware autonomous driving decision-making in an uncontrolled intersection. Mathematical Problems in Engineering, 2016. https://doi.org/10.1155/2016/1025349

Sun, L., Zhan, W., Chan, C. Y., & Tomizuka, M. (2019). Behavior planning of autonomous cars with social perception. IEEE Intelligent Vehicles Symposium, Proceedings, 2019-June(Iv), 207–213. https://doi.org/10.1109/IVS.2019.8814223

Sun, L., Zhan, W., Tomizuka, M., & Dragan, A. D. (2018). Courteous autonomous cars. IEEE International Conference on Intelligent Robots and Systems, 663–670. https://doi.org/10.1109/IROS.2018.8593969

Tafidis, P., Farah, H., Brijs, T., & Pirdavani, A. (2022). Safety implications of higher levels of automated vehicles: a scoping review. Transport Reviews, 42(2), 245–267. https://doi.org/10.1080/01441647.2021.1971794

Taghavifar, H., & Mohammadzadeh, A. (2024). Integrating deep reinforcement learning and social-behavioral cues: A new human-centric cyber-physical approach in automated vehicle decision-making. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering. https://doi.org/10.1177/09544070241230126

Talebpour, A., & Mahmassani, H. S. (2016). Influence of connected and autonomous vehicles on traffic flow stability and throughput. Transportation Research Part C: Emerging Technologies. https://doi.org/10.1016/j.trc.2016.07.007

Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R., & Fallah, Y. P. (2021a). Altruistic maneuver planning for cooperative autonomous vehicles using multi-agent advantage actor-critic. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)-Workshop on Autonomous Driving: Perception, Prediction and Planning. IEEE/CVF, 1–8.

Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R., & Fallah, Y. P. (2021b). Cooperative autonomous vehicles that sympathize with human drivers. IEEE International Conference on Intelligent Robots and Systems, 4517–4524. https://doi.org/10.1109/IROS51168.2021.9636151

Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R., & Fallah, Y. P. (2022). Social coordination and altruism in autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 23(12), 24791–24804. https://doi.org/10.1109/TITS.2022.3207872

Tong, Y., Wen, L., Cai, P., Fu, D., Mao, S., Shi, B., & Li, Y. (2024). Human-like decision making at unsignalized intersections using social value orientation. IEEE Intelligent Transportation Systems Magazine, 16(2), 55–69. https://doi.org/10.1109/MITS.2023.3342308

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. In Annals of Internal Medicine. https://doi.org/10.7326/M18-0850

U.S. Department of Transportation Federal Highway Administration. (2016). Next Generation Simulation (NGSIM) vehicle trajectories and supporting data. [Dataset]. Provided by ITS DataHub through Data.transportation.gov. Accessed 2024-05-09 from http://doi.org/10.21949/1504477

Valiente, R., Razzaghpour, M., Toghi, B., Shah, G., & Fallah, Y. P. (2024). Prediction-aware and reinforcement learning-based altruistic cooperative driving. IEEE Transactions on Intelligent Transportation Systems, 25(3), 2450–2465. https://doi.org/10.1109/TITS.2023.3323440

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. https://doi.org/10.1007/s11192-009-0146-3

Vasile, L., Dinkha, N., Seitz, B., Dasch, C., & Schramm, D. (2023). Comfort and safety in conditional automated driving in dependence on personal driving behavior. IEEE Open

Journal of Intelligent Transportation Systems, 4(June), 772–784. https://doi.org/10.1109/OJITS.2023.3323431

Vemula, A., Muelling, K., & Oh, J. (2018). Social attention: Modeling attention in human crowds. Proceedings - IEEE International Conference on Robotics and Automation, 4601–4607. https://doi.org/10.1109/ICRA.2018.8460504

Vinkhuyzen, E., & Cefkin, M. (2016). Developing socially acceptable autonomous vehicles. Ethnographic Praxis in Industry Conference Proceedings, 2016(1), 522–534. https://doi.org/10.1111/1559-8918.2016.01108

Wang, B., Su, R., Huang, L., Lu, Y., & Zhao, N. (2024). Distributed cooperative control and optimization of connected automated vehicles platoon against cut-in behaviors of social drivers. IEEE Transactions on Automatic Control, PP, 1–8. https://doi.org/10.1109/TAC.2024.3401082

Wang, J., Zhang, Y., Wang, X., & Li, L. (2023). A human-like lane-changing behavior model for autonomous vehicles in mixed traffic flow environment. IET Conference Proceedings, 2023(26), 107–112. https://doi.org/10.1049/icp.2023.3359

Wang, L., Fernandez, C., & Stiller, C. (2023a). High-level decision making for automated highway driving via behavior cloning. IEEE Transactions on Intelligent Vehicles, 8(1), 923–935. https://doi.org/10.1109/TIV.2022.3169207

Wang, L., Fernandez, C., & Stiller, C. (2023b). Learning safe and human-like high-level decisions for unsignalized intersections from naturalistic human driving trajectories. IEEE Transactions on Intelligent Transportation Systems, 24(11), 12477–12490. https://doi.org/10.1109/TITS.2023.3286454

Wang, L., Sun, L., Tomizuka, M., & Zhan, W. (2021). Socially-compatible behavior design of autonomous vehicles with verification on real human data. IEEE Robotics and Automation Letters, 6(2), 3421–3428. https://doi.org/10.1109/LRA.2021.3061350

Wang, W., Wang, L., Zhang, C., Liu, C., & Sun, L. (2022). Social interactions for autonomous driving: A review and perspectives. Foundations and Trends® in Robotics. https://doi.org/10.1561/2300000078

Wang, X., Chen, X., Jiang, P., Lin, H., Yuan, X., Ji, M., Guo, Y., Huang, R., & Fang, L. (2024). The group interaction field for learning and explaining pedestrian anticipation. Engineering, 34, 70–82. https://doi.org/10.1016/j.eng.2023.05.020

Wang, X., Tang, K., Dai, X., Xu, J., Du, Q., Ai, R., Wang, Y., & Gu, W. (2024). S4TP: Social-suitable and safety-sensitive trajectory planning for autonomous vehicles. IEEE Transactions on Intelligent Vehicles, 9(2), 3220–3231. https://doi.org/10.1109/TIV.2023.3338483

Wang, X., Zhang, X., Zhu, Y., Guo, Y., Yuan, X., Xiang, L., Wang, Z., Ding, G., Brady, D., Dai, Q., & Fang, L. (2020). Panda: A gigapixel-level human-centric video dataset. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR42600.2020.00333

Wang, Z., Gao, P., He, Z., & Zhao, L. (2021). A CGAN-based model for human-like driving decision making. IEEE Wireless Communications and Networking Conference, WCNC, 2021-March. https://doi.org/10.1109/WCNC49053.2021.9417336

Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J. K., Ramanan, D., Carr, P., & Hays, J. (2023). Argoverse 2: Next generation datasets for self-driving perception and forecasting. NeurIPS.

Xia, C., Xing, M., & He, S. (2022). Interactive planning for autonomous driving in intersection scenarios without traffic signs. IEEE Transactions on Intelligent Transportation Systems, 23(12), 24818–24828. https://doi.org/10.1109/TITS.2022.3205250

Xie, S., Chen, S., Tomizuka, M., Zheng, N., & Wang, J. (2020). To develop human-like automated driving strategy based on cognitive construction: Appraisal and perspective. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC 2020. https://doi.org/10.1109/ITSC45102.2020.9294591

Xu, C., Zhao, W., Wang, C., Cui, T., & Lv, C. (2023). Driving behavior modeling and characteristic learning for human-like decision-making in highway. IEEE Transactions on Intelligent Vehicles, 8(2), 1994–2005. https://doi.org/10.1109/TIV.2022.3224912

Xu, Y., Shao, W., Li, J., Yang, K., Wang, W., Huang, H., Lv, C., & Wang, H. (2022). SIND: A drone dataset at signalized intersection in China. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2022-Octob, 2471–2478. https://doi.org/10.1109/ITSC55140.2022.9921959

Xue, J., Zhang, D., Xiong, R., Wang, Y., & Liu, E. (2023). A two-stage based social preference recognition in multi-agent autonomous driving system. IEEE International Conference on Intelligent Robots and Systems, 5507–5513. https://doi.org/10.1109/IROS55552.2023.10341803

Yan, Y., Wang, J., Zhang, K., Liu, Y., Liu, Y., & Yin, G. (2022). Driver's individual risk perception-based trajectory planning: A human-like method. IEEE Transactions on Intelligent Transportation Systems, 23(11), 20413–20428. https://doi.org/10.1109/TITS.2022.3190521

Yaqoob, I., Khan, L. U., Kazmi, S. M. A., Imran, M., Guizani, N., & Hong, C. S. (2020). Autonomous driving cars in smart cities: Recent advances, requirements, and challenges. IEEE Network. https://doi.org/10.1109/MNET.2019.1900120

Yoon, D. D., & Ayalew, B. (2019). Social force aggregation control for autonomous driving with connected preview. Proceedings of the American Control Conference, 2019-July, 1388–1393. https://doi.org/10.23919/acc.2019.8814725

Zhan, W., Sun, L., Wang, D., Shi, H., Clausse, A., Naumann, M., Kummerle, J., Konigshof, H., Stiller, C., de La Fortelle, A., & Tomizuka, M. (2019). INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION dataset in interactive driving scenarios with semantic maps. http://arxiv.org/abs/1910.03088

Zhang, D. (2023). Universe simulator. Accessed 2024-05-09 from https://github.com/alibaba-damo-academy/universe

Zhang, L., Dong, Y., Farah, H., & Van Arem, B. (2023). Social-aware planning and control for automated vehicles based on driving risk field and model predictive contouring control: Driving through roundabouts as a case study. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, 3297–3304. https://doi.org/10.1109/SMC53992.2023.10394462

Zhang, T., Zhan, J., Shi, J., Xin, J., & Zheng, N. (2023). Human-like decision-making of autonomous vehicles in dynamic traffic scenarios. IEEE/CAA Journal of Automatica Sinica, 10(10), 1905–1917. https://doi.org/10.1109/JAS.2023.123696

Zhao, C., Chu, D., Deng, Z., & Lu, L. (2024). Human-like decision making for autonomous driving with social skills. IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 9, pp. 12269-12284. https://doi.org/10.1109/TITS.2024.3366699

Zhou, D., Ma, Z., Zhao, X., & Sun, J. (2022). Reasoning graph: A situation-aware framework for cooperating unprotected turns under mixed connected and autonomous traffic environments. Transportation Research Part C: Emerging Technologies, 143(July), 103815. https://doi.org/10.1016/j.trc.2022.103815

Zhou, M., Luo, J., Villella, J., Yang, Y., Rusu, D., Miao, J., Zhang, W., Alban, M., Fadakar, I., Chen, Z., Huang, A. C., Wen, Y., Hassanzadeh, K., Graves, D., Chen, D., Zhu, Z., Nguyen, N., Elsayed, M., Shao, K., … Wang, J. (2020). SMARTS: Scalable multi-agent reinforcement learning training school for autonomous driving. Proceedings of Machine Learning Research.

Zhu, M., Wang, Y., Pu, Z., Hu, J., Wang, X., & Ke, R. (2020). Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving. Transportation Research Part C: Emerging Technologies, 117, 102662. https://doi.org/10.1016/j.trc.2020.102662

Zhu, Z., & Zhao, H. (2023). Joint imitation learning of behavior decision and control for autonomous intersection navigation. IEEE International Conference on Intelligent Robots and Systems, 1564–1571. https://doi.org/10.1109/IROS55552.2023.10342405

Zong, Z., Shi, J., Wang, R., Chen, S., & Zheng, N. (2023). Human-like decision making and planning for autonomous driving with reinforcement learning. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, September, 3922–3929. https://doi.org/10.1109/ITSC57777.2023.10421908

8 Evaluation on deep reinforcement learning for automated driving in various manoeuvres and implementation of safe, efficient, comfortable, and energy-saving driving through roundabouts

Abstract

Developing and testing automated driving models in the real world might be challenging and even dangerous, while simulation can help with this, especially for challenging manoeuvres. Deep reinforcement learning (DRL) has the potential to tackle complex decision-making and controlling tasks through learning and interacting with the environment, thus it is suitable for developing automated driving while not being explored in detail yet. This study first conducted a comprehensive evaluation and implementation of DRL algorithms across diverse driving scenarios. Using the *highway-env* simulation platform, Deep Q-Network (DQN) and Trust Region Policy Optimisation (TRPO) were implemented and compared. Customised reward functions were developed, and models were evaluated based on lane accuracy, speed efficiency, safety from collisions, and driving comfort. Results indicated that TRPO-based models with tailored reward functions achieved superior performance across most metrics. To extend the scope beyond specific driving manoeuvres, this study expanded *highway-env* by developing a customised training environment, *ComplexRoads*, which integrates diverse road scenarios and manoeuvres, enabling models to generalise effectively across tasks.

Further exploration focused on the intricate challenges of driving through roundabouts, where state-space complexity and dynamic interactions complicate the driving modelling, planning and control. Here, three DRL algorithms, i.e., Deep Deterministic Policy Gradient (DDPG), Proximal Policy Optimisation (PPO), and TRPO, were implemented with reward functions prioritising safety, efficiency, comfort, and energy savings. Evaluation methods were refined using an Analytic Hierarchy Process (AHP) to weigh performance indicators. Experimental

results showed TRPO excelling in safety and efficiency while PPO optimised comfort for roundabout driving.

This integrated study demonstrates the versatility of DRL in addressing diverse automated driving challenges, providing a robust foundation for deploying DRL-based models in real-world traffic environments.

This chapter is based on the edited version of the two published research papers:

Dong, Y., Datema, T., Wassenaar, V., Van de Weg, J., Kopar, C. T., & Suleman, H. (2023). Comprehensive Training and Evaluation on Deep Reinforcement Learning for Automated Driving in Various Simulated Driving Maneuvers. In 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC) (pp. 6165-6170). IEEE. https://doi.org/10.1109/ITSC57777.2023.10422159

Yuan, H., Li, P., Van Arem, B., Kang, L., Farah, H., & Dong, Y. (2023). Safe, Efficient, Comfort, and Energy-saving Automated Driving through Roundabout Based on Deep Reinforcement Learning. In 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC) (pp. 6074-6079). IEEE. https://doi.org/10.1109/ITSC57777.2023.10422488*

8.1 Introduction

Artificial intelligence (AI) is making huge improvements in various fields, one of which is automated driving (Badue et al., 2021). One typical type of AI that is well-suitable for developing automated driving models is Deep Reinforcement Learning (DRL) (Rao & Frtunikj, 2018). DRL makes use of the advantage of deep neural networks regarding feature extraction and the advantage of reinforcement learning regarding learning from interacting with the environment. DRL exhibits excellent performance in various decision-making tasks, e.g., *GO* (Silver et al., 2016) and playing video games (Shao et al., 2019), and it has been employed in various automated driving tasks (Khalil & Mouftah, 2023; Kiran et al., 2022; Zhu & Zhao, 2022), e.g., lane-keeping, lane-changing, overtaking, ramp merging, and driving through intersections.

For the lane-keeping task, El Sallab et al. (2017) and Sallab et al. (2016) developed DRL-based methods for delivering both discrete policies using Deep Q-Network (DQN) and continuous policies using Deep Deterministic Actor-Critic Algorithm (DDAC) to follow the lane and to maximise the average velocity when driving on the curved race track on Open Racing Car Simulator (TORCS). Similarly, for the lane-changing task, Wang et al. (2018) trained a DQN-based model to perform decision-making of lane-keeping, lane changing to the left/right, and acceleration/deceleration, so that the trained agent can intelligently make a lane change under diverse and even unforeseen scenarios. Furthermore, Zhang et al. (2023) proposed a bi-level lane-change behaviour planning strategy using a DRL-based lane-change decision-making model and a negotiation-based right-of-way assignment model to deliver multi-agent lane-change manoeuvres. For the overtaking task, Kaushik et al. (2018) adopted Deep Deterministic Policy Gradients (DDPG) to learn overtaking manoeuvres for an automated vehicle (AV) in the presence of multiple surrounding cars in a simulated highway scenario. They verified that their curriculum learning resembled approach can learn to perform smooth overtaking manoeuvres, largely collision-free, and independent of the track and number of cars in the scene. For the ramp merging task, Wang and Chan (2018) employed a Long-Short Term Memory (LSTM) neural network to model the interactive environment conveying internal states containing historical driving information to a DQN which then generated Q-values for action selection regarding on-ramp merging. Additionally, for negotiating and driving through intersections, Isele et al. (2018) explored the effectiveness of the DQN-based DRL method in handling the task of navigating through unsignaled intersections. Finally, Guo and Ma (2021) developed a real-time learning and control framework for signalised intersection management, which integrated both vehicle trajectory control and signal optimisation using DDPG-based DRL learning directly from the dynamic interactions between vehicles, traffic signal control and traffic environment in the mixed connected and automated vehicle (CAV) environment.

It is observed that although many studies have utilised DRL for various driving tasks, most of them focus only on one specific driving manoeuvre. Seldom do they evaluate the DRL model performance across different manoeuvres and nor do they explore the adaptability of DRL models trained on one specific environment but tested in other various manoeuvres. This study first tries to fill this research gap by implementing, evaluating, and comprehensively comparing the performance of two DRLs, i.e., DQN and TRPO, in various driving scenarios. Customised effective reward functions were developed, and the implemented DRLs were evaluated in terms of various aspects, considering driving safety, efficiency, and comfort level. Then, this study typically customised and compared DDPG, PPO, and TRPO for the complex roundabout

driving scenarios and took energy savings into consideration. Results indicated that TRPO-based models with tailored reward functions achieved superior performance across most metrics.

To train a uniform driving model that can tackle various driving tasks, this study further constructed a new simulation environment, named "*ComplexRoads*" (shown in **Figure 8-1**), integrating various driving manoeuvres and multiple road scenarios. The *ComplexRoads* served to train a uniform driving model that can tackle various driving tasks. For verification, the models trained only on *ComplexRoads* were tested and evaluated in the specific driving manoeuvres. Intensive experimental results demonstrated the effectiveness of this customised training environment.

To advance the learning capability for the developed DRL-based AI models, i.e. encouraging relational insight, besides designing *ComplexRoads*, several built-in functions of the highway-env package were also upgraded. Notable modifications are summarised as follows: the tracking of the "current" lane with respect to the car (training agent) was upgraded to take into account the lane heading to eliminate confusing transitions when driving off-road. Furthermore, the distance between the car and its current lane was upgraded to a signed value to allow for orientation distinction. Similarly, the lane heading difference, LHD for short, was adjusted to also be a signed value. These improvements yield increased learning abilities for both on-road driving, returning to on-road driving when off-road, and a general sense of "awareness" given an arbitrary environment.

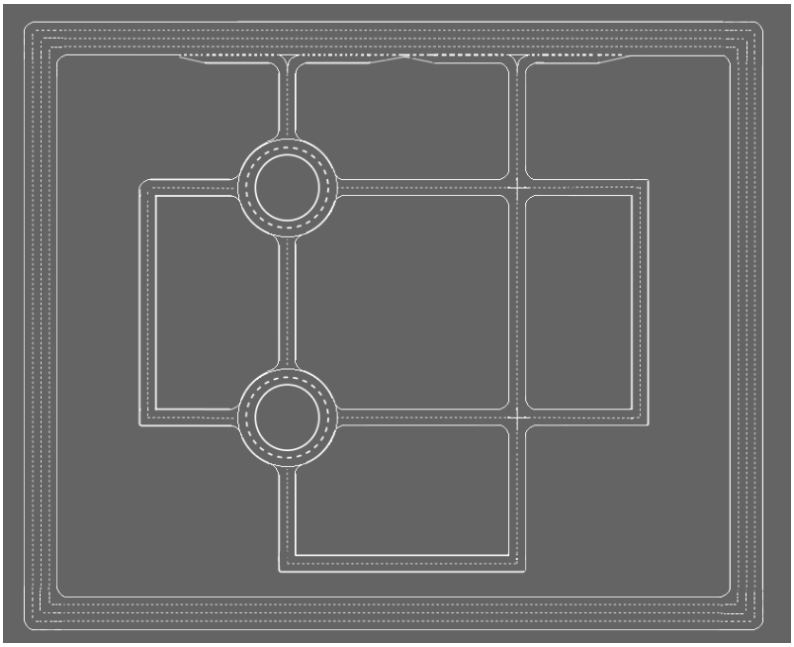


Figure 8-1. Illustration for the layout of the *ComplexRoads* environment

8.2 Methodology

8.2.1 System architecture

The general DRL learning cycle is an iterative learning process based on the agent's performance in the environment influenced by the agent's actions. In mathematical terms,

automated driving can be modelled as a Markov Decision Process (MDP) (Qiying Hu, 2007). MDP captures the features of sequential decision-making. The components of an MDP include environments, agents, actions, rewards, and states. In this study, the system framework, which illustrates the corresponding MDP, is depicted in **Figure 8-2**. The system generally consists of five main elements, i.e., environment, agent, action, state, and reward, which will be elaborated in detail in this section.

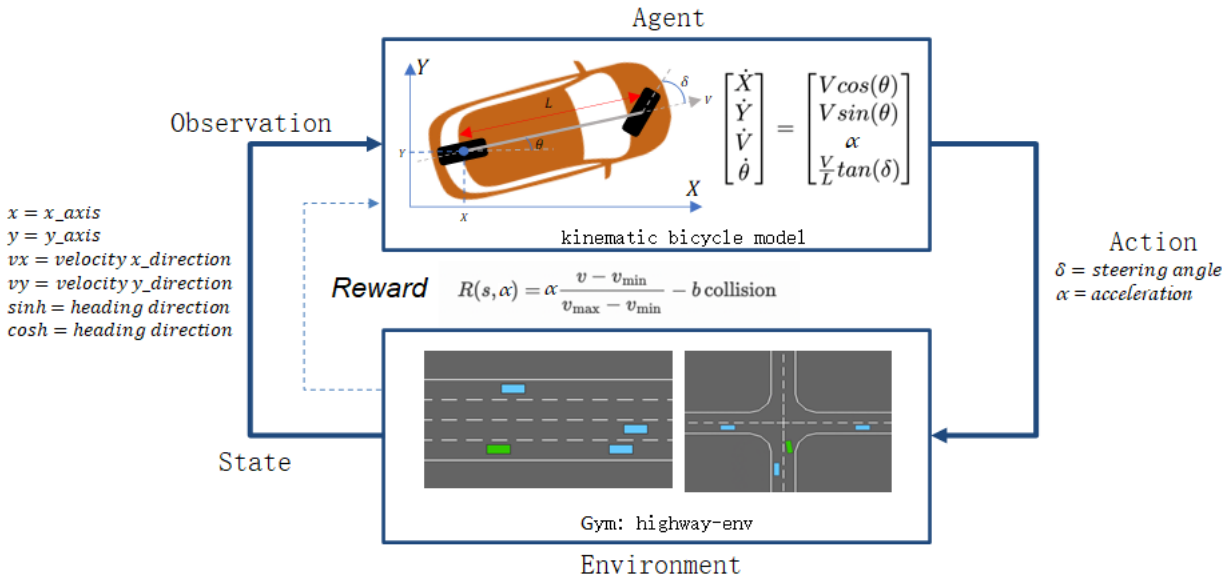


Figure 8-2. Illustration for the system framework of the DRL MDP

8.2.2 DRL MDP elements

Environment: To simulate the MDP, this study adopted the *highway-env* platform (Leurent, 2018), which is a Python-based package that offers a variety of driving environments. As a widely used platform, ample research has been conducted using the *highway-env*, such as (Alizadeh et al., 2019; Liu et al., 2022). In the *highway-env*, six dedicated driving scenarios are available, i.e., *Highway*, *Merge*, *Roundabout*, *Intersection*, *Racetrack*, and *Parking*. Users can also customise environments by specifying the number of lanes, the size of a given roundabout, and other parameters. In this study, all the driving scenarios, except for the *Highway* and *Parking*, are covered. For training and evaluating a uniform driving model, this study designed a new simulation environment, named "*ComplexRoads*" (shown in **Figure 8-1**). *ComplexRoads* integrates two highway merging scenarios, two four-way intersections, two roundabouts, and several segments of multi-straight lanes. The DRL models trained only on *ComplexRoads* were tested and evaluated in the specific driving manoeuvres originally available on *highway-env*.

Agent: A kinematic bicycle model is used to represent the vehicle as the agent of MDP. Despite its simplicity, a kinematic bicycle model is able to represent actual vehicle dynamics (Polack et al., 2017).

Action: An action taken by the agent in the proposed MDP is an element from the contracted *Action Space*. The *highway-env* environment offers three types of action spaces: Discrete Action, Discrete Meta Action, and Continuous Action. This study employs a hybrid approach using both discrete and continuous actions to train distinct driving tasks. In this study, the two dimensions

of the *Action Space* $\boldsymbol{A}$ are: acceleration (throttle) and steering angle of the front wheels. Depending on the DRL algorithm $\boldsymbol{A}$ is either of the form $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \times [-5, 5]$ for algorithms requiring a continuous action space, or $\{\delta_1, \cdots, \delta_n\} \times [\alpha_1, \cdots, \alpha_m]$ in the $n \times m$ discrete case. Hence, $(\delta, \alpha) \in \boldsymbol{A}$, where steering is denoted by δ and acceleration is denoted by α.

State: As illustrated in **Figure 8-2**, the state in the proposed MDP includes the ego AV's state, e.g., location (x, y), velocity (v_x, v_y), and heading direction, together with the surrounding vehicles state and road conditions and is directly accessible at each time frame to the ego car, either in absolute terms or relative to itself.

Reward: The customised Reward function is elaborated in detail in the following subsections of *8.2.3* and *8.2.4*.

8.2.3 General reward function

For training the models, this study used the reward function already present in the highway-env package (referred to as the baseline reward and illustrated in the middle of **Figure 8-2**) and the modified and upgraded reward function. The model performances were compared to demonstrate that the upgraded reward is better than the baseline reward. During the training, it was observed that in the early stages, the trained agent car would sometimes drive off the road. To make the training more efficient in handling off-road driving and stimulating the agent to return to driving on-road, one specific contribution in this study is to adjust the distance measure between the agent and the lane, in addition to constructing the lane heading difference measure illustrated in the following paragraphs. Let c denote the ego car agent and $\boldsymbol{\mathcal{L}}$ the corresponding lane. A lane is a collection of lane points $l \in \boldsymbol{\mathcal{L}}$. Now define l' as the lane point with the shortest Euclidean distance to the car, meaning

$$l' \coloneqq \arg\min_{l \in \boldsymbol{\mathcal{L}}} (c, l) \tag{8-1}$$

and define the orientation ω of the car c with respect to a lane point l as follows

$$\omega(c, l) = \begin{cases} 1 & if\ the\ car\ is\ located\ left\ of\ l^4 \\ -1 & otherwise \end{cases} \tag{8-2}$$

Then, this study defines the distance between the ego car and the lane as the shortest distance from the ego car c to any point l on lane $\boldsymbol{\mathcal{L}}$, meaning

$$d(c, \boldsymbol{\mathcal{L}}) = \omega(c, l')d(c, l') \tag{8-3}$$

The car heading and lane point heading are denoted by c_φ and l_φ respectively, both values are within the angle range $(-\pi, \pi]$. Then, the lane heading difference (LHD) is defined as

$$\text{LHD} = \begin{cases} l_\varphi - c_\varphi + 2\pi & if\ l_\varphi - c_\varphi < -\pi \\ l_\varphi - c_\varphi - 2\pi & if\ l_\varphi - c_\varphi > \pi \\ l_\varphi - c_\varphi & otherwise \end{cases} \tag{8-4}$$

[4] More precisely, if the car is located left of the tangent line for the lane segment containing l.

An important remark to this setup is the fact that if $sgn(\text{LHD}) \cdot d(c, \mathcal{L}) < 0$ then the car is heading for the lane. Similarly, if $sgn(\text{LHD}) \cdot d(c, \mathcal{L}) > 0$ the car is deviating (further) from the lane. Four different off-road scenarios are shown in **Figure 8-3**.

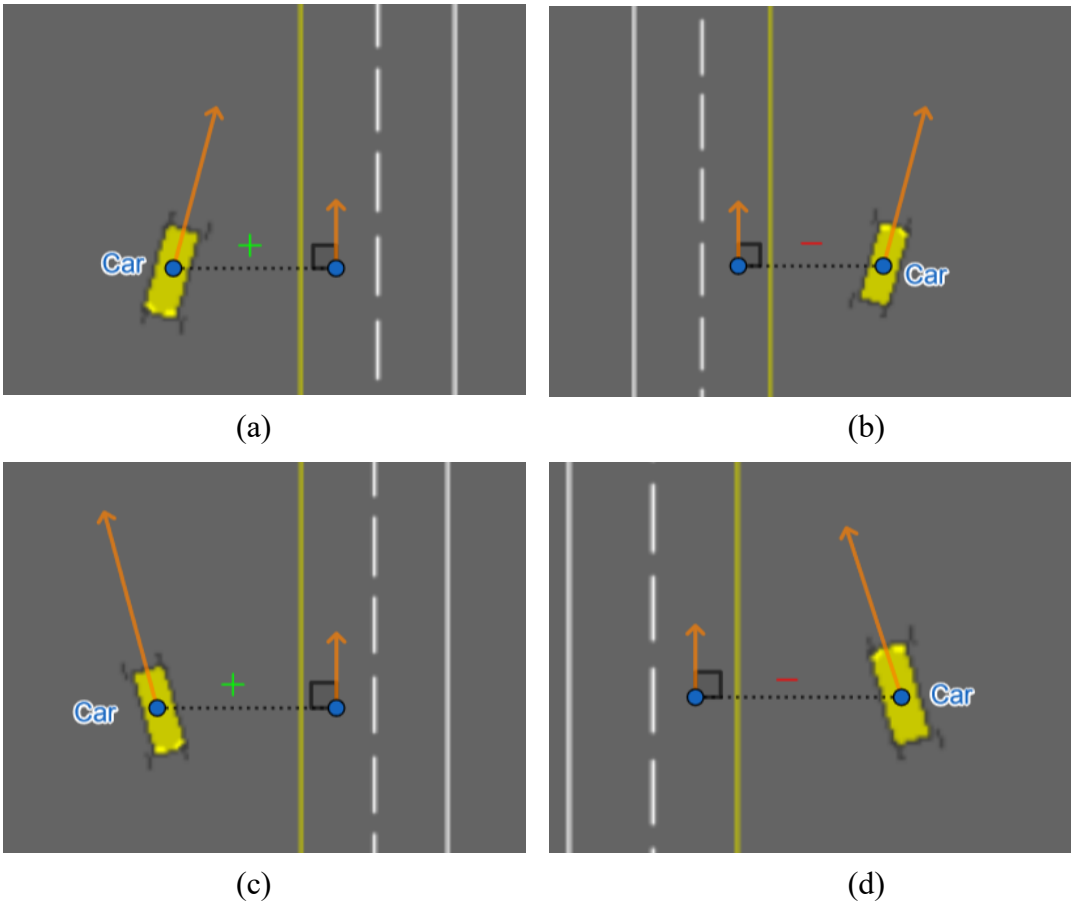


Figure 8-3. Four different off-road scenarios showcasing available environment observations of the ego car: (a) Off-road scenario with $d(c, l') > 0$ and $LHD < 0$; (b) Off-road scenario with $d(c, l') < 0$ and $LHD < 0$; (c) Off-road scenario with $d(c, l') > 0$ and $LHD > 0$; (d) Off-road scenario with $d(c, l') < 0$ and $LHD > 0$

In **Figure 8-3**, both lane heading and car heading are portrayed by vectors. The lane distance and LHD, for the ego car c with respect to the lane point l'. The sign is orientation-based: if the car is located left of the road, the Euclidean distance is perceived as positive, and negative if located right of the road.

Finally, denote the velocity of the ego car c by v_c, the reward function R with regard to the state S is defined as

$$R_S(c, \mathcal{L}) = \begin{cases} \frac{\cos(|\text{LHD}|) \cdot v_c}{20 \cdot max(1, |d(c, \mathcal{L})|)} & if\ v_c \geq 0 \\ 0 & otherwise \end{cases} \tag{8-5}$$

where LHD is the lane heading difference between the ego car and the closest lane point. However, if the car crashes during the simulation, the reward is automatically set as -10, regardless of the state.

The reward function, as defined in equation (8-5), rewards the car for its "effective" speed on the road, defined by the cosine of the angular difference between the direction the car is driving

in and the direction in which the road goes, multiplied by the speed of the car. With this design, both an increase in the driving speed and driving in line with the road heading will result in high rewards. Moreover, the value is divided by the lane offset to punish the car for driving off-road and also divided by 20 to scale the reward function to remain close to 1 under optimal circumstances.

8.2.4 Rewards customised for navigating through roundabouts

Specifically for AVs' navigating through roundabouts, the reward function is designed regarding driving safety, efficiency, comfort level, and energy consumption.

1) Safety reward

In the roundabout driving context, safety is primarily influenced by two factors, i.e., lane-centre positioning and time-to-collision (TTC). The lane-centring reward, indicated by R_{LC}., can be computed as

$$R_{LC} = 1 - \left(\frac{l_{lateral}}{l_{width/2}}\right)^2 \tag{8-6}$$

where $l_{lateral}$ is the vehicle's offset to the centre of the lane, and l_{width} is the lane width.

The TTC reward is computed as

$$R_{TTC} = 1 - \frac{3}{TTC} \tag{8-7}$$

If the Time-to-Collision (TTC) exceeds 3 seconds, the TTC reward will fall within the range of 0 to 1. A larger TTC results in a reward closer to 1. Conversely, when TTC is less than 3, the reward becomes negative. And in the event of an imminent collision, the TTC reward will approach $-\infty$.

The total safety reward is a weighted sum of the lane centre reward and the TTC reward. The TTC reward constitutes 70% of the R_{safe}, while the lane centre reward makes up the remaining 30%. The total safety reward can be expressed as:

$$R_{safe} = 0.7 \times R_{TTc} + 0.3 \times R_{LC} \tag{8-8}$$

2) Efficiency reward

The efficiency reward motivates the AV to move forward, avoiding stationary actions. It mainly rewards high speeds within set limits. When the vehicle's speed is less than or equal to the speed limit, the efficiency reward is set to the ratio of the vehicle's current speed to the speed limit as

$$R_{efficient} = \frac{v_{ego}}{v_{limit}} \tag{8-9}$$

When the vehicle's speed is greater than the speed limit, the reward value decreases as the speed increases.

$$R_{efficient} = 1 - \frac{v_{ego} - v_{limit}}{v_{max} - v_{limit}} \tag{8-10}$$

where v_{ego} is the current speed, v_{limit} is the speed limit on the road, and v_{max} is the maximum achievable speed value of the vehicle.

3) Comfort level reward

Vehicle comfort, a key performance indicator for automated driving, significantly impacts user experience. This study focuses on smooth acceleration, deceleration, and steering. The reward function considers the rate of change in acceleration/braking and steering. Lower rates of change, indicating smoother movements, yield higher rewards, while higher rates of change result in lower rewards. The calculation of the comfort level reward value is as follows

$$diff_{throttle} = \frac{d\ throttle_t}{dt} \tag{8-11}$$

$$diff_{steering} = \frac{d\ a_{steering}}{dt} \tag{8-12}$$

$$R_{comfort} = 1 - \frac{diff_{throttle} + diff_{steering}}{4} \tag{8-13}$$

where $diff_{throttle}$ is the rate of change of the throttle or brake, $a_{steering}$ is the input value of the throttle or brake, $diff_{steering}$ is the rate of change of the steering wheel, and $a_{steering}$ is the input value of the steering wheel.

4) Energy consumption reward

Jiménez Palacios (1999) indicates that Vehicle Specific Power (VSP) can indirectly reflect vehicle energy consumption, demonstrating a roughly linear positive correlation with specific power. Hence, specific power values can be used to approximate energy consumption. Parameters for this model were calibrated by (Jiménez Palacios, 1999). In this study, the slope resistance term is omitted since road slope is not considered.

$$VSP = v \times (1.1a + 0.132) + 0.000302v^3 \tag{8-14}$$

For the setting of the reward function, this study considers the maximum specific power value of the vehicle and uses it as a standard to normalise the value of the specific power at the current moment to the range from 0 to 1, and thus

$$R_{energy} = 1 - \frac{VSP}{VSP_{max}} \tag{8-15}$$

5) Total integrated rewards

In the roundabout setting, AVs will enter from any of the four entrances with a predefined exit destination. A destination reward is implemented for the agent to learn to navigate towards its objective when performing continuous actions. This reward is Boolean, i.e., it is set to 1 if the vehicle reaches the target exit and 0 otherwise:

$$R_{arrive} = \begin{cases} 1 & if\ the\ \text{vehicle}\ has\ reached\ the\ target\ exit \\ 0 & else \end{cases} \tag{8-16}$$

The total integrated reward function combines the aforementioned sub-reward functions through a weighted sum. Having closely similar weights for all four sub-reward functions would overcomplicate the reward function and hinder satisfactory model training. Emphasis is placed on safety and efficiency by assigning larger weights, as they are critical elements. The total reward function is calculated as

$$R_{total} = 0.6\,R_{safe} + 0.25\,R_{efficient} + 0.1\,R_{comfort} + 0.05\,R_{energy} + R_{arrive} \quad (8\text{-}17)$$

8.2.5 DRL algorithms

DRL is a specialised machine learning algorithm designed to aid agents in decision-making. Through interactive training between the agent and its environment, DRL can enhance the agent's decision-making capacities. In this study, the agent, i.e., an automated vehicle, is trained in various simulated driving environments. Through DRL training, the agent iteratively updates its policy (controlling the throttle and steering) to maximise the obtained reward (encompassing safety, efficiency, comfort, and energy consumption). The DRL-based approach enables the model to optimise the decision-making strategy and determine subsequent actions.

Regarding selected DRL algorithms, TRPO (Schulman et al., 2015), DDPG (Lillicrap et al., 2016), PPO (Schulman et al., 2017), and DQN (Fan et al., 2020) were customised and implemented. The DRL was implemented through the PyTorch deep learning framework. The DRL algorithms are instantiated via the stable-baselines3 (Raffin et al., 2021) reinforcement learning library. Details of the DRLs, including hyperparameter settings, are elaborated in the supplementary at https://lnkd.in/gSb92UcR and https://lnkd.in/gft8fscf.

8.2.6 Evaluation of the models

To evaluate and compare the model performance, one needs a set of indicators and metrics, for which this study implemented a performance logger that measures and stores various indicators when testing a model in a given environment. These indicators are measured for a set number of runs, and the logger then prints the average values over all the runs. The measured indicators are: 1) Speed, 2) Peak jerk, 3) Total jerk, 4) Total distance, 5) Total steering, 6) Running time, 7) Onlane rate (rate of time the cars are running within the road), and 8) Rate of collision.

The jerk is defined as the difference between the current and the previous action of a vehicle, consisting of both the steering angle and the acceleration. The magnitude of the total jerk reflects the degree to which the vehicle's motion changes abruptly and frequently, where a higher value of the total jerk implies less comfortable driving. The jerk is defined by equations in (8-18)-(8-20):

$$J_{acceleration} = \frac{\alpha_t - \alpha_{t-1}}{\alpha_{max} - \alpha_{min}} \quad (8\text{-}18)$$

$$J_{steering} = \frac{\delta_t - \delta_{t-1}}{\delta_{max} - \delta_{min}} \quad (8\text{-}19)$$

$$J_{total} = \frac{J_{acceleration} + J_{steering}}{2} \quad (8\text{-}20)$$

The total steering is defined as the total sum of steering the car performs in the course of an evaluation, measured in angles. A higher amount of steering could, to a certain extent, imply less efficient driving with unnecessary steering. The onlane rate is defined as the amount of time the evaluated car spends driving on the lane, divided by the total amount of time the car spends driving. The collision rate is defined as the total amount of collisions the car makes, divided by the total amount of evaluation trials.

While evaluating the specifically selected challenging scenario of AVs' navigating through roundabouts, the average collision rate, lane-centring loss value, efficiency, comfort, and energy consumption level were selected as the metrics. The impact of the above five evaluation indicators on automated driving is different, and thus the weight of each indicator needs to be further analysed. For that, this study utilised the Analytic Hierarchy Process (AHP) method to determine the weights of the five testing indicators. Details of the AHP process are provided in the supplementary materials available at https://lnkd.in/gSb92UcR. The final estimated weight values are shown in **Table 8-1**.

Table 8-1. Estimated weight values

Indicator	Weight Value
Average collision rate score test value	0.4764
Average lane-centring loss	0.2853
Average efficiency	0.1428
Average comfort level	0.0634
Average energy consumption level	0.0320

8.3 Experiments

Firstly, for the general comparison, this study conducted intensive experiments to train and evaluate DRL models using TRPO and DQN algorithms on four environments provided by highway-env, and also the newly self-designed *ComplexRoads*. The models were trained using both the original standard reward function provided by *highway-env* (which served as the baseline) and the customised reward function. The hyperparameters used for training can be found in the supplementary materials at https://lnkd.in/gSb92UcR and https://lnkd.in/gft8fscf. The models were trained on the supercomputer Delft Blue (*DelftBlue Supercomputer (Phase 1)*, 2022). For every environment, ten models were trained and saved for 10,000 and 100,000 iterations. When finishing training, the model performance was tested for 10 runs. During the performance testing, constraints such as a maximum running time, minimum speed, and whether a crash had occurred were adopted. To obtain an overall assessment, the average of all these 10 testing results was calculated. To get an idea of how well the models perform regarding a uniform driving model, they were not only tested in their trained environments but also cross-evaluated in other different environments. With the cross-evaluation, the effectiveness of the newly designed environment *ComplexRoads* can be verified.

Then, regarding the specific focus of the roundabout driving case, this study implemented DDPG, TRPO, and PPO, and evaluated and compared their performances. In the implementation, model fine-tuning and hyperparameter optimisation play a vital role in enhancing the performance of reinforcement learning algorithms. Model fine-tuning adjusts the algorithm model's specifics and structure, while hyperparameter optimisation involves selecting and adjusting the hyperparameters within the algorithm to improve performance.

Typical techniques for model fine-tuning include neural network structure adjustment, e.g., tweaking the number of layers, neurons, and activation function, to boost the algorithm's efficacy. In this research, all these three DRL algorithms adopt similar network structures.

Specifically, both the actor and critic networks of DDPG are designed with two hidden layers, each containing 64 neurons. TRPO and PPO utilise the Multi-Layer Perceptron (MLP) neural network with two hidden layers, each containing 64 neurons.

In reinforcement learning, hyperparameters are parameters that cannot be optimised iteratively during training and need to be set manually beforehand. Hyperparameter tuning involves adjusting these parameters to enhance algorithm performance. This study employs grid search to optimise hyperparameter values, preserving or excluding hyperparameter combinations based on the decrease or increase of the reward function during training.

8.4 Results and discussion

8.4.1 Comprehensive comparison in various scenarios

Tables 8-2, **8-3**, **8-4**, **8-5**, and **8-6** present the average performances of the DRL models trained on five environments and evaluated on the same respective environment. For every model variant in one specific environment, this study trained it for 10 times and also evaluated it for 10 times to get the average performance indicators. This study writes "1*" when the number is rounded to 1, but not quite equal to 1. With the letters "B" and "M", this study refers to whether the baseline reward function or the modified reward function was used in training the model.

Meanwhile, **Tables 8-7**, **8-8**, **8-9**, **8-10**, **8-11**, and **8-12** present the average performances of the implemented DRL models trained in their own environment but evaluated in other different environments. This is for evaluating how adaptive these models are.

One needs to note that for the environment of *Merge* and the self-designed *ComplexRoads*, no baseline reward functions are available, so only the models trained by the modified and upgraded reward (indicated with "-M") were evaluated. Also, for cross-environment evaluation, only models with the modified reward were evaluated.

Table 8-2. **DRL model performances in *ComplexRoads***

Indicator	DQN-M	TRPO-M
Speed	16.1	16.2
Peak jerk	0.990	0.799
Total jerk	221.0	13.2
Total distance	661	547
Total steering	263	46
Runtime	607	492
Onlane rate	0.999	0.999
Collision rate	0.07	0.09

Table 8-3. DRL model performances in *Roundabout* scenario

Indicator	DQN-B	DQN-M	TRPO-B	TRPO-M
Speed	8.30	8.60	7.88	8.25
Peak jerk	1.070	1.090	0.704	0.775
Total jerk	71.0	159.0	15.4	20.7
Total distance	185	278	214	229
Total steering	128	210	92.8	114
Runtime	318	479	382	394
Onlane rate	0.384	0.783	0.341	0.693
Collision rate	0.71	0.68	0.51	0.62

Table 8-4. DRL model performances in *Intersection* scenario

Indicator	DQN-B	DQN-M	TRPO-B	TRPO-M
Speed	9.89	10.10	9.74	10.30
Peak jerk	0.892	1.040	0.545	0.637
Total jerk	24.3	32.6	6.2	6.5
Total distance	38.6	62.8	65.0	68.3
Total steering	29.9	41.4	18.7	18.5
Runtime	59	93	101	100
Onlane rate	0.988	0.999	0.999	1*
Collision rate	0.38	0.49	0.33	0.19

Table 8-5. DRL model performances in *Merge* scenario

Indicator	DQN-M	TRPO-M
Speed	30.9	29.1
Peak jerk	0.863	0.607
Total jerk	47.8	11.2
Total distance	491	487
Total steering	86.7	82.1
Runtime	226	253
Onlane rate	0.875	0.836
Collision rate	0.5	0.4

Table 8-6. DRL model performances in *Racetrack* scenario

Indicator	DQN-B	DQN-M	TRPO-B	TRPO-M
Speed	7.12	9.44	10.30	7.59
Peak jerk	0.956	0.756	0.518	0.962
Total jerk	70.6	43.1	8.4	127.0
Total distance	229	207	254	222
Total steering	183	68	85	181
Runtime	449	346	362	471
Onlane rate	0.225	0.991	0.943	0.992
Collision rate	0.13	0.84	0.74	0.29

Table 8-7. DQN-M trained on *ComplexRoads* evaluated in other environments

Indicator	Racetrack	Roundabout	Merge	Intersection
Speed	10.2	8.3	30.6	10.0
Total distance	180	200	377	59
Runtime	275	349	185	89
Onlane rate	0.998	0.602	0.935	0.998
Collision rate	0.92	0.79	0.30	0.52

Table 8-8. TRPO-M trained on *ComplexRoads* evaluated in other environments

Indicator	Racetrack	Roundabout	Merge	Intersection
Speed	10.0	9.0	29.8	10.3
Total distance	130	195	339	59.7
Runtime	222	289	172	87
Onlane rate	1*	0.647	0.996	0.999
Collision rate	0.82	0.76	0.10	0.51

Table 8-9. DQN-M trained on *Roundabout* evaluated in other environments

Indicator	Racetrack	Merge	Intersection
Speed	10.7	30.6	10.1
Total distance	156	335	22
Runtime	224	164	33
Onlane rate	0.954	0.955	0.968
Collision rate	0.97	0.20	0.05

Table 8-10. **TRPO-M trained on *Intersection* evaluated in other environments**

Indicator	Racetrack	Merge	Roundabout
Speed	9.00	30.90	8.91
Total distance	137	477	236
Runtime	253	228	345
Onlane rate	0.999	0.970	0.527
Collision rate	0.57	0.10	0.68

Table 8-11. **TRPO-M trained on *Merge* evaluated in other environments**

Indicator	Intersection	Racetrack	Roundabout
Speed	9.87	9.85	9.38
Total distance	14	437	349
Runtime	22	632	486
Onlane rate	0.886	0.399	0.159
Collision rate	0.06	0.16	0.38

Table 8-12. **TRPO-M trained on *Racetrack* evaluated in other environments**

Indicator	Intersection	Merge	Roundabout
Speed	9.69	29.7	7.38
Total distance	51	304	113
Runtime	79	154	239
Onlane rate	0.996	0.970	0.849
Collision rate	0.67	0.60	0.76

While there might be various ways to express that one model outperforms another, it is important to prioritise safety as the main concern. Therefore, the measured values considered the most important in the comparison here are the onlane rate and the collision rate, which reflect driving safety. Other values, such as speed or jerk, are less important but can be compared in cases where the onlane and collision rates are similar.

From **Tables 8-2**, **8-3**, **8-4**, **8-5**, and **8-6**, one can see that in most cases the DQN with modified reward function (DQN-M) and the TRPO with modified reward function (TRPO-M) outperform the DQN and TRPO models with the baseline reward functions, especially with regards to the onlane rate. Between the DQN and TRPO models, the models trained by TRPO tend to perform better in most cases.

Furthermore, looking at **Tables 8-7**, **8-8**, **8-9**, **8-10**, **8-11**, and **8-12**, it is observed that the models trained on *ComplexRoads* indeed tend to perform better than the other models in the cross-

evaluation, especially in keeping a high onlane rate. This is due to various traffic situations represented in the *ComplexRoads* environment, as well as the fact that the starting location of the car during training on *ComplexRoads* was randomised, meaning that the car can experience various driving situations. This will also prevent the model from merely "memorising" the environment, but instead learning better to master the manoeuvres to interact with the randomly generated environments.

Due to the size of *ComplexRoads*, training on it was very computationally intensive, especially with a large amount of simulated surrounding cars. Non-ego cars get destinations assigned randomly and drive around scripted, meaning they follow deterministic driving rules to drive "perfectly" and receive a new destination upon reaching the previous one. Thus, this study opted to train the model with relatively few surrounding cars, meaning that the model does not get to interact with other cars as often as in the other environments. Due to this, it resulted in a higher collision rate when evaluated in other environments with more surrounding cars. When the computational resource is abundant, by adding more surrounding cars into the *ComplexRoads* environment, this reduced awareness of the ego car can be reduced.

All in all, it is verified that the designed *ComplexRoads* training environment indeed contributes to the training of a more flexible and adaptive driving model. All the testing scenarios and results are better demonstrated in the supplementary materials with the demo videos also provided at https://lnkd.in/gft8fscf.

8.4.2 Comparison for navigating through roundabout scenarios

Specifically, for the evaluation of AVs' navigating through roundabouts regarding driving safety, efficiency, comfort level, and energy consumption, the comparison results of the selected three DRL models (i.e., DDPG, TRPO, and PPO) are shown in **Table 8-13**.

Table 8-13. **Model performance comparison in *Roundabout* scenarios**

Indicator / Algorithm	Collision Rate Score	Lane-centring	Efficiency	Comfort	Energy Consumption	Total Test Score
DDPG	0.43	0.8653	0.8872	0.8846	0.8058	0.6606
PPO	0.68	0.8385	0.8784	**0.9836**	**0.8103**	0.7769
TRPO	**0.73**	**0.9322**	**0.9295**	0.8627	0.7995	**0.8267**

In **Table 8-13**, the average collision rate score is calculated as (8-21):

$$Collision\ Rate = 1 - \frac{num_{collision}}{T} \times 10^3 \tag{8-21}$$

where $num_{collision}$ is the number of vehicle collisions during the entire simulating test, T is the total simulation time step of the 50 rounds of testing (larger than 5000). This calculation converts the collision performance into a score of 0 to 1.

The results show that TRPO outperforms the other two compared DRLs in collision rate score, lane-centring loss, and efficiency metrics, though it lags slightly in comfort level and energy consumption compared to the other two algorithms. Overall, TRPO achieved the highest integrated test score, surpassing both DDPG and PPO.

DDPG, while defective in terms of collision rate score, demonstrates better lane-centring and efficiency performance than PPO, yet falls behind TRPO. While PPO excels in comfort and energy consumption, it lags behind TRPO in terms of the other three metrics. Despite individual algorithm strengths in certain aspects, overall, TRPO performs the best.

For model characteristics and their verification, DDPG uses a deep Q-network to estimate the optimal action-value function, differing from TRPO and PPO, which utilise natural policy gradient algorithms with distinct optimisation constraints. For exploration, DDPG applies noise-induced action perturbations suitable for continuous action spaces, although possibly resulting in slower convergence. In contrast, TRPO and PPO use stochastic policies, usually providing more effective global optimal solutions. Unlike DDPG's instability due to hyperparameter sensitivity, TRPO and PPO exhibit robustness and stability thanks to their conservative optimisation strategies.

To sum up, for the specifically evaluated scenarios of navigating roundabouts, TRPO excels in collision rate score, lane-centring, and efficiency, and delivers the best overall testing score; PPO is distinct in comfort and energy consumption, and follows TRPO regarding the overall testing score; while DDPG may be hampered by its sensitivity to hyperparameters and less effective exploration strategies leading to the worst overall testing performance.

8.5 Conclusion

This study first summarised the utilisation of DRL in every specific automated driving task, e.g., lane-keeping, lane-changing, overtaking, and ramp merging, then customised and implemented two widely used DRLs, i.e., DQN and TRPO, to tackle various driving manoeuvres, and finally specifically compared three DRLs, i.e., TRPO, DDPG, and PPO, regarding safe, efficient, comfortable and energy-saving navigating through roundabouts, and carried out a comprehensive evaluation and comparison on the model performance. Based on the *highway-env* simulation platform, a modified and upgraded reward function was designed for training the DRL models in general. Furthermore, a new integrated training environment, *ComplexRoads*, was constructed, together with several built-in functions were upgraded. Through various experiments, it is verified that the models trained using the modified reward generally outperformed those with the original baseline reward, and the newly constructed *ComplexRoads* demonstrated effective performance in training a uniform model that can tackle various driving tasks rather than one specific manoeuvre. As a preliminary study, the findings will provide meaningful and instructive insights for future studies towards developing automated driving in complex and real traffic environments with DRL and simulation.

This study approached the challenge of training a uniform driving model from the perspective of designing an integrated training environment. However, future research should prioritise the development of a uniform driving model from an algorithmic standpoint. Additionally, exploring alternative direct navigation reward designs that seamlessly integrate strategic planning with low-level control presents a promising avenue for further investigation.

References

Alizadeh, A., Moghadam, M., Bicer, Y., Ure, N. K., Yavas, U., & Kurtulus, C. (2019). Automated lane change decision making using deep reinforcement learning in dynamic

and uncertain highway environment. 2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019, 1399–1404. https://doi.org/10.1109/ITSC.2019.8917192

Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., Jesus, L., Berriel, R., Paixão, T. M., Mutz, F., de Paula Veronese, L., Oliveira-Santos, T., & De Souza, A. F. (2021). Self-driving cars: A survey. Expert Systems with Applications, 165(September 2019), 113816. https://doi.org/10.1016/j.eswa.2020.113816

DelftBlue supercomputer (Phase 1). (2022). https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1

El Sallab, A., Abdou, M., Perot, E., & Yogamani, S. (2017). Deep reinforcement learning framework for autonomous driving. IS and T International Symposium on Electronic Imaging Science and Technology, 70–76. https://doi.org/10.2352/ISSN.2470-1173.2017.19.AVM-023

Fan, J., Wang, Z., Xie, Y., & Yang, Z. (2020). A theoretical analysis of deep Q-learning. In Learning for dynamics and control (pp. 486-489). PMLR.

Guo, Y., & Ma, J. (2021). DRL-TP3: A learning and control framework for signalized intersections with mixed connected automated traffic. Transportation Research Part C: Emerging Technologies. https://doi.org/10.1016/j.trc.2021.103416

Isele, D., Rahimi, R., Cosgun, A., Subramanian, K., & Fujimura, K. (2018). Navigating occluded intersections with autonomous vehicles using deep reinforcement learning. Proceedings - IEEE International Conference on Robotics and Automation, 2034–2039. https://doi.org/10.1109/ICRA.2018.8461233

Jiménez Palacios, J. L. (1999). Understanding and quantifying motor vehicle emissions with vehicle specific power and TILDAS remote sensing. Massachusetts Institute of Technology, Cambridge.

Kaushik, M., Prasad, V., Krishna, K. M., & Ravindran, B. (2018). Overtaking maneuvers in simulated highway driving using deep reinforcement learning. IEEE Intelligent Vehicles Symposium, Proceedings, 2018-June (IV), 1885–1890. https://doi.org/10.1109/IVS.2018.8500718

Khalil, Y. H., & Mouftah, H. T. (2023). Exploiting multi-modal fusion for urban autonomous driving using latent deep reinforcement learning. IEEE Transactions on Vehicular Technology, 72(3), 2921–2935. https://doi.org/10.1109/TVT.2022.3217299

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Sallab, A. A. A., Yogamani, S., & Perez, P. (2022). Deep reinforcement learning for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems, 23(6), 4909–4926. https://doi.org/10.1109/TITS.2021.3054625

Leurent, E. (2018). An environment for autonomous driving decision-making. Accessed 2024-05-09 from https://github.com/eleurent/highway-env

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2016). Continuous control with deep reinforcement learning. 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings.

Liu, Q., Dang, F., Wang, X., & Ren, X. (2022). Autonomous highway merging in mixed traffic using reinforcement learning and motion predictive safety controller. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2022-Octob, 1063–1069. https://doi.org/10.1109/ITSC55140.2022.9921741

Polack, P., Altche, F., DAndrea-Novel, B., & De La Fortelle, A. (2017). The kinematic bicycle model: A consistent model for planning feasible trajectories for autonomous vehicles? IEEE Intelligent Vehicles Symposium, Proceedings. https://doi.org/10.1109/IVS.2017.7995816

Qiying Hu, W. Y. (2007). Markov decision processes with their applications. Springer Science & Business Media.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. Journal of Machine Learning Research.

Rao, Q., & Frtunikj, J. (2018). Deep learning for self-driving cars : Chances and challenges. 2018 IEEE/ACM 1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIAS), 35–38.

Sallab, A. El, Abdou, M., Perot, E., & Yogamani, S. (2016). End-to-end deep reinforcement learning for lane keeping assist. NIPS, 1–9. http://arxiv.org/abs/1612.04340

Schulman, J., Levine, S., Moritz, P., Jordan, M., & Abbeel, P. (2015). Trust region policy optimization. 32nd International Conference on Machine Learning, ICML 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. 1–12. http://arxiv.org/abs/1707.06347

Shao, K., Tang, Z., Zhu, Y., Li, N., & Zhao, D. (2019). A survey of deep reinforcement learning in video games. ArXiv Preprint ArXiv:1912.10944, 1–13. http://arxiv.org/abs/1912.10944

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature. https://doi.org/10.1038/nature16961

Wang, P., & Chan, C. Y. (2018). Formulation of deep reinforcement learning architecture toward autonomous driving for on-ramp merge. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2018-March, 1–6. https://doi.org/10.1109/ITSC.2017.8317735

Wang, P., Chan, C. Y., & De La Fortelle, A. (2018). A reinforcement learning based approach for automated lane change maneuvers. IEEE Intelligent Vehicles Symposium, Proceedings. https://doi.org/10.1109/IVS.2018.8500556

Zhang, J., Chang, C., Zeng, X., & Li, L. (2023). Multi-agent DRL-based lane change with right-of-way collaboration awareness. IEEE Transactions on Intelligent Transportation Systems, 24(1), 854–869. https://doi.org/10.1109/TITS.2022.3216288

Zhu, Z., & Zhao, H. (2022). A survey of deep RL and IL for autonomous driving policy learning. IEEE Transactions on Intelligent Transportation Systems, 23(9), 14043–14065. https://doi.org/10.1109/TITS.2021.3134702

9 Social-aware planning and control for automated vehicles based on driving risk field and model predictive contouring control: Driving through roundabouts as a case study

Abstract

The gradual deployment of automated vehicles (AVs) results in mixed traffic where AVs will interact with human-driven vehicles (HDVs). Thus, social-aware motion planning and control while considering interactions with HDVs on the road is critical for AVs' deployment and safe driving under various manoeuvres. Previous research mostly focuses on the trajectory planning of AVs using Model Predictive Control or other relevant methods, while seldom considering the integrated planning and control of AVs altogether to simplify the whole pipeline architecture. Furthermore, there are very limited studies on social-aware driving that make AVs understandable and expected by human drivers, and none when it comes to the challenging manoeuvre of driving through roundabouts. To fill these research gaps, this study develops an integrated social-aware planning and control algorithm for AVs' driving through roundabouts based on Driving Risk Field (DRF), Social Value Orientation (SVO), and Model Predictive Contouring Control (MPCC), i.e., DRF-SVO-MPCC. The proposed method is tested and verified with simulations on the open-sourced highway-env platform. Compared with the baseline method using purely Nonlinear Model Predictive Control, the DRF-SVO-MPCC can achieve better performance under various manoeuvres of driving through roundabouts with and without surrounding HDVs.

This chapter is based on the published research paper:

Zhang, L., Dong, Y., Farah, H., & Van Arem, B. (2023). Social-Aware Planning and Control for Automated Vehicles Based on Driving Risk Field and Model Predictive Contouring Control: Driving Through Roundabouts as a Case Study. In 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 3297-3304). IEEE. https://doi.org/10.1109/SMC53992.2023.10394462 (Co-first authors and corresponding author).*

9.1 Introduction

Purely fully autonomous vehicles on roads are demonstrated to be beneficial to road safety and efficiency (Yaqoob et al., 2020). However, the gradual development and deployment of automated vehicles (AVs) and advanced driver assistance systems (ADAS) at various levels results in mixed-traffic conditions where AVs need to interact with human-driven vehicles (HDVs). Thus, making AVs' behaviour understandable, expected, and accepted by human drivers through so-called social-aware driving models is critical for road safety and efficiency under various manoeuvres, especially challenging ones, e.g., driving on weaving sections, highly curved roads, and driving through roundabouts.

There are some preliminary studies regarding social-aware driving (Wang et al., 2022). These studies usually focus on social cooperation for AVs' path planning, whose methods can be mainly divided into two categories, i.e., learning-based and model-based methods. Reinforcement learning methods, such as Deep Q-Network (DQN), Actor-Critic (A2C), and Proximal Policy Optimisation (PPO), integrating Partially Observable Stochastic Games (POSG) can factor surrounding HDVs' influence into the AVs' path planning and then connect to proportional–integral–derivative (PID) as a low-level controller for path tracking (Toghi et al., 2021a, 2021b). Since the reward for social compliance is difficult to quantify, many researchers employed inverse reinforcement learning (IRL) to learn and mimic how human drivers act in the real world using empirical driving data (Li et al., 2020; Schwarting et al., 2019; Sun et al., 2018; L. Wang et al., 2021). In addition to reinforcement learning-based approaches, some studies adopted deep learning, e.g., Social Long Short Term Memory (LSTM) (Alahi et al., 2016) and Social Generative Adversarial Network (GAN) (Gupta et al., 2018), incorporating social factors for trajectory prediction of the surrounding HDVs, and then designed socially aware path-planning for AVs correspondingly. These are all learning-based methods. Regarding model-based methods, in (Hang et al., 2020) and (Hang et al., 2021), a game-theoretic-based decision-making approach is combined with Model Predictive Control (MPC) under the dynamic bicycle model (Kong et al., 2015) to build a complete architecture tackling scenarios such as lane changing, overtaking, etc. This approach requires the estimation of the model parameters for different environments and is not robust to different scenarios. Another model-based approach is to build a field model to estimate the dangers around AVs (Ji et al., 2017; Kolekar et al., 2020; Mullakkal-Babu et al., 2020). Ji et al. (2017) created 3D risk fields and combined them with MPC for path planning and tracking to ensure a collision-free path for AV. Kolekar et al. (2020) developed a driving risk field (DRF) model to quantify the risks perceived by drivers. And by coupling DRF to a controller that can maintain the perceived risk below a threshold, they generated human-like driving behaviour. The required model parameters of the human driver were obtained through simulation. In addition, the model does not require real-time parameter estimation, improving the robustness regarding different environments. Although field-based planning and control can reduce the occurrence of hazards and is highly robust across different scenarios, little consideration is given to social cooperation and the impact of different driving styles on social compliance with surrounding HDVs.

On the other hand, MPC's capability to handle multiple-input multiple-output (MIMO) systems with various constraints makes it particularly suitable for real-world autonomous vehicle planning and control. Thus, it is also necessary to review relevant research in this domain. MPC can be traced back to the 1980s when engineers in the process industry first started deploying it in real practice (Garriga & Soroush, 2010). MPC methods assume a finite look-ahead horizon

for which control signals are calculated to optimise an objective function. MPC allows direct planning and control of the vehicle, whether driving on the highway or parking in low-speed scenarios with different predicted models (Yoon et al., 2009; Zhang et al., 2018, 2021). In (Buyval et al., 2017; Obayashi et al., 2016), the HDV was simply seen as an obstacle, and the optimisation goal of the MPC is to move away from the obstacle on the highway. This can lead to unexpected scenarios where vehicles are seen as dangerous objects even if they are driving in the same direction with no conflicts, and it is hard to tackle uncertain environments such as an intersection. For this, Ulfsjoo & Axehill (2022) additionally employed the partially observable Markov decision process (POMDP) for decision-making before MPC, allowing it to handle more uncertain scenarios. At the same time, to encourage vehicles to collaborate, MPCs that can control multiple AVs within a scenario were developed (Faris et al., 2022; Pauls et al., 2022). The problem is that it only enables cooperation between AVs, and it is difficult to consider other users on the road, needless to say, delivering social-aware driving.

From the previous reviews, it is identified that the disadvantage of MPC is that it is difficult to take into account the risks faced by other vehicles on the road, while purely using the aforementioned social cooperation-based path planning method alone can result in a less flexible and less reliable path. Furthermore, few studies implemented integrated planning and control together, and seldom did they cover the challenging manoeuvre of driving through roundabouts. To fill these research gaps, this research studies the suitability of utilising MPC incorporating the DRF method to generate a social-aware driving algorithm that can safely control the motion of a vehicle driving through a roundabout while being able to handle potential conflicts with surrounding HDVs and considering different levels of interests of other road users. There are several challenges. The first one is to ensure the safety and comfort of all users on the road. It is important to understand the intention of human drivers correctly and try to work with the HDVs correspondingly. Machines and humans do not understand the danger/risk in the same way. Thus, what AVs need is to "think" more like humans and anticipate possible dangers to interact with other HDVs safely. Furthermore, for social-aware driving, it is necessary to modify the AV's original objective by balancing its own benefits versus the benefits of other surrounding HDVs, considering the different driving styles and characteristics of human drivers, thus making the AV accepted by HDVs. Different human drivers possess different priorities concerning safety, efficiency, and attitudes toward other vehicles, reflecting their different driving styles, e.g., aggressive, and defensive(W. Wang et al., 2022). Also, the driving style of AVs determined by the needs of the passengers may vary from time to time, and case by case. For example, for daily commuters and those in a hurry, the efficiency of their journey should be assigned with a higher priority. While, if there is an elderly or sick person in the vehicle, he/she probably will place more weight on comfort level and be more willing to give precedence to others to ensure safety. Finally, it is challenging for the model to maintain robustness in tackling different scenarios and handling different driving styles.

To tackle these challenges, this study develops an integrated social-aware planning and control algorithm incorporating DRF, Social Value Orientation (SVO) (Liebrand & McClintock, 1988), and Model Predictive Contouring Control (MPCC). DRF is adopted to model the surrounding drivers' perceived risk when interacting with the AV. The SVO, a social psychology-derived approach, is utilised to measure how individuals make the trade-off between personal benefits and the benefits to others (Liebrand & McClintock, 1988). Then, the model-based DRF-SVO is packaged into the MPC framework connecting to the specific MPCC algorithm to implement

the integration of both planning and control. The integration avoids approaching the motion planning and feedback control hierarchically, and therefore brings more stability to the system. With the proposed DRF-SVO-MPCC algorithm, this study implements two types of driving styles, i.e., egoistic and prosocial, where an egoistic vehicle will not tolerate any increase in its own cost, a prosocial vehicle will prefer a minor increase in its own cost or the surrender of part of its benefits to reduce the danger of other vehicles. Lastly, the proposed model is verified on complex manoeuvres, i.e., driving through roundabouts with large curvature, which is one of the most accident-prone scenarios. Both single-lane and two-lane roundabouts, which are common in most countries, are tested to verify the robustness and generalisation ability of the proposed method.

In short, the main contributions of this study are:

1. A social-aware MPC is developed by combining MPC and DRF using SVO as the bridge to consider both the accuracy of controls and the perceived danger of other vehicles. Integration with SVO also makes it possible to balance the benefits of ego AV versus those of surrounding HDVs.
2. Different driving styles are generated under the proposed DRF-SVO-MPCC method, especially with the help of SVO. SVO can also determine the desired driving style of the AV under different situations.
3. The proposed DRF-SVO-MPCC integrates motion planning and feedback control simultaneously, improving the stability of the vehicle control system.
4. The performance of the proposed DRF-SVO-MPCC is validated on challenging manoeuvres, i.e., driving through both single-lane and two-lane roundabouts with two different driving styles implemented.

9.2 Basic theory

9.2.1 Model predictive control

In this study, the MPC aims to minimise the cost function for the system based on the non-linear prediction model on the vehicle and system constraints. The general formulation of the non-linear MPC can be written as follows:

$$min \ \sum_{k=0}^{N_P-1} J_k\left(X_k, U_k, X_k^{ref}\right) \tag{9-1a}$$

$$\text{s.t.:} \quad X_{k+1} = f(X_k, U_k), k = 0, \dots, N_P - 1 \tag{9-1b}$$

$$G(X_k, U_k) \leq g_b, k = 0, \dots, N_P - 1 \tag{9-1c}$$

$$X_0 = X_{init} \tag{9-1d}$$

In (9-1), U_k and X_k are the input and state of the system, respectively. The function J_k is the cost function that determines the cost of the whole system, and the function G comprises all constraints, with g_b being the bound value. These constraints ensure the system state and inputs are within a set boundary. Currently, the constraints are only defined as box constraints; however, they are flexible to be expanded. N_P is the prediction horizon for the MPC. The

predicted mode $X_{k+1} = f(X_k, U_k)$ is based on the kinematic bicycle model (Kong et al., 2015), which is written as:

$$\dot{x} = vcos(\psi + \beta) \tag{9-2a}$$

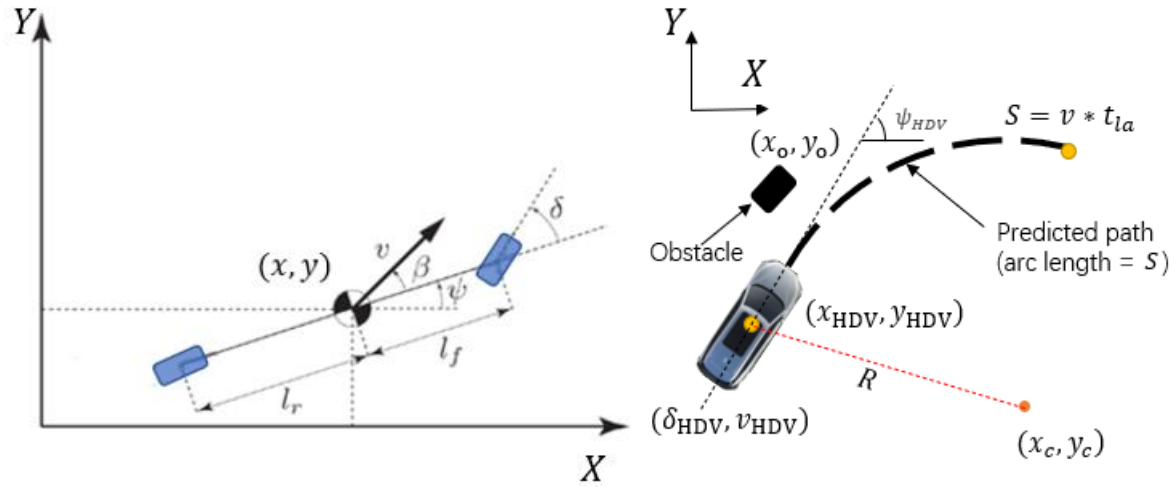
$$\dot{y} = vsin(\psi + \beta) \tag{9-2b}$$

$$\dot{\psi} = \frac{v}{l_r}\sin(\beta) \tag{9-2c}$$

$$\dot{v} = a \tag{9-2d}$$

$$\beta = \tan^{-1}(\frac{l_r}{l_r + l_f}\tan(\delta)) \tag{9-2e}$$

As in **Figure 9-1**, the x and y are the longitudinal and lateral positions of the vehicle, respectively. ψ is the heading angle of the vehicle, and v is the velocity of the vehicle. $[x, y, \psi, v]$ are the state variables of the kinematic bicycle model. The distance from the centre of gravity to the front and rear wheels are l_f and l_r, respectively. β is the angle of the current velocity of the centre of mass with respect to the longitudinal axis of the vehicle. The control input parameters are the front steering angle and the acceleration, which are $[\delta, a]$. This model is a non-linear model, which means that this study concentrated on Nonlinear Model Predictive Control (NMPC).



(a) (b)

Figure 9-1. Illustration of (a) the kinematic bicycle model and (b) the predicted path in the DRF model

The continuous space model is discretised to $X_{k+1} = f(X_k, U_k) = X_k + \Delta_t f^c(X_k, U_k)$ with a discretisation time Δ_t.

Several steps should be followed when using the MPC formulation described above. Firstly, the measured or estimated current state should be obtained as the initial state. The second step is to solve the optimal control formula. Then the optimal control input sequence (N_P elements) will be obtained. Finally, only the first element in the sequence will be applied to the system and then move to the next MPC round.

9.2.2 Driving risk field

The Driving Risk Field (DRF) (Kolekar et al., 2020) represents the driver's belief about the probability of the risk occurring. The value of a DRF can change with the vehicle's different velocities and steering angles. Since the kinematic bicycle model is used as the prediction model in MPC, to maintain consistency, it is also adopted to calculate the vehicle's path in DRF:

$$R = \frac{l_r + l_f}{\tan(\delta)} \tag{9-3}$$

As shown in **Figure 9-1 (b)**, in (9-3), the radius of the arc (R) of the vehicle's preceding trajectory and the centre of the turning circle (x_c, y_c) can be determined by the HDV's position (x_{HDV}, y_{HDV}), HDV's heading ψ_{HDV}, and HDV's steering angle δ_{HDV}. The DRF of a vehicle is modelled as a torus with a Gaussian cross-section, which can be written as:

$$DRF(x_o, y_o) = a \exp\left(\frac{-\left(\sqrt{(x_o - x_c)^2 + (y_o - y_c)^2} - R\right)^2}{2\sigma^2}\right) \tag{9-4}$$

The coordinate of a risk obstacle to the HDV is (x_o, y_o). The height (a) of the Gaussian is modelled as a parabola, and the width (σ) of the Gaussian is modelled as a linear function which is a simplification of the parabolic function:

$$a(s) = p(s - vt_{la})^2 \tag{9-5}$$

$$\sigma = (m + k_i|\delta|)s + c \tag{9-6}$$

$$i = 1\ (inner\ \sigma), or\ 2\ (outter\ \sigma)$$

The t_{la} is a fixed look-ahead time. Based on it, the look-ahead distance increases linearly with the velocity of the vehicle. And p is a parameter that defines the parabola's steepness. The width of DRF at the location of the vehicle (c) is related to the car width and m defines the slope of widening of the DRF when driving straight. Then, k_1 and k_2 which represent the parameters of the inner and outer edges of the DRF, respectively, can affect the width of the DRF, and they can help to generate asymmetric DRFs. With this modelling method, the risk grows linearly with the increasing steering angle. It is similar to a human when the driver controls the steering of the vehicle, which simulates the driver paying more attention to the environment in the direction turned, resulting in a higher risk presented in the other direction. The increase in DRF is proportional to δ, leading to higher risk when driving through sharp curves with cumulatively smaller radii.

So, all the hyperparameters in DRF are related to the driver's status instead of the environment. In this work, the DRF is utilised to obtain the possible risk of the HDVs interacting with AVs. Therefore, the coordinates in (9-4) are from the HDV's perspective, while all other parameters represent those of the driver in the HDV. The human driver parameters are identified through a simulation, and referring to (Kolekar et al., 2020), in this study, the parameters are from a 25-year-old male volunteer driver, shown in **Table 9-1**. This will allow AVs to put themselves in the shoes of other drivers to be informed of what they perceive as the probability of danger, which will also better reflect the consideration for social-aware driving.

9.2.3 Social value orientation

Social Value Orientation (SVO), a metric from social psychology (Liebrand & McClintock, 1988), is a parameter that describes how much a person is willing to consider the benefits of other people versus his/her own. In psychology, each individual wants to maximise the reward and minimise the cost when considering only himself or herself. However, as social road users, some of our planning needs to take into account the welfare of others. The SVO term conducts us to model each individual's social preferences by expressing their cost function as a combination of two terms, the cost to self J_{self} and the cost to others J_{other}:

$$J_{total} = \cos\,\alpha\, J_{self} + \sin\alpha\; J_{other} \tag{9-7}$$

where α, as an angle, indicates the value of SVO. It reflects the selfishness or altruism of each individual. Just like in **Figure 9-2**, when this angle is 0^0, it means that the system is completely individualistic; while when the angle is 90^0, it means that the system is completely altruistic to other systems. In Fig. 2, it is noticed that most people's SVOs are between 0^0 and 60^0 illustrated by the blue points. In this work, to motivate AVs to behave with different personality traits like human drivers, two different styles, i.e., prosocial and egoistic, are implemented. Furthermore, it should be ensured that the lower limit of SVO is set so as not to completely ignore the risk of colliding with other vehicles. As a result, regarding the two driving styles, α is set as 60^0 for prosocial driving and 15^0 for egoistic driving.

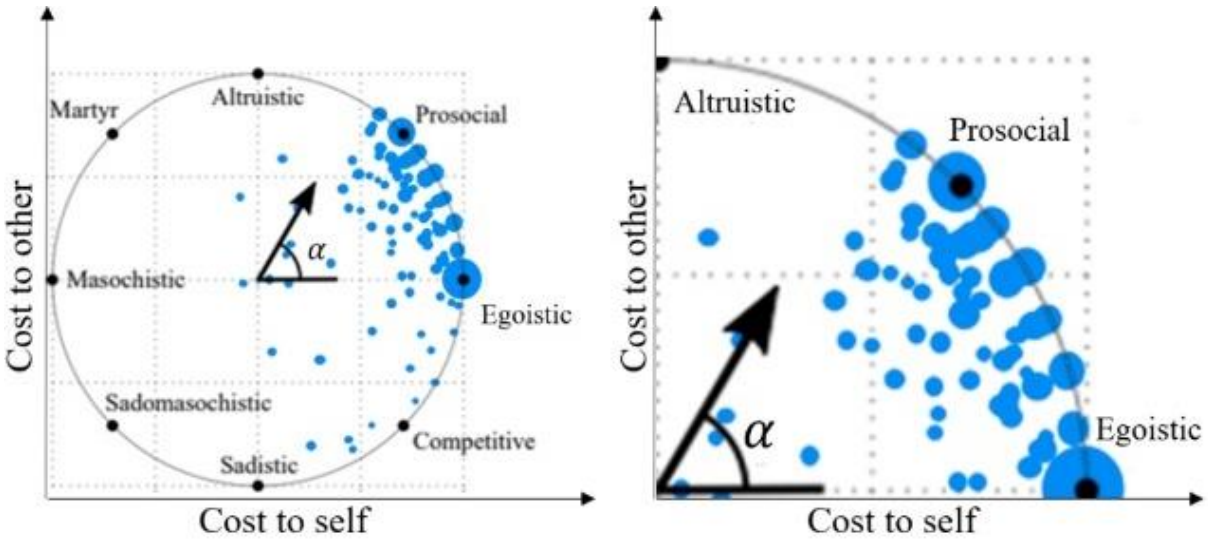


Figure 9-2. Illustration of SVO and its distribution in the population (Buckman et al., 2019)

9.3 Social-aware DRF-SVO-MPCC implementation

9.3.1 Quantifying perceived risk

In this section, the proposed method is introduced in detail. Firstly, the overall architecture of the proposed four-phase pipeline is illustrated and briefly explained. Then, each of the four phases, i.e., image pre-processing, self-supervised pre-training, fine-tuning classification, and post-processing, is depicted with comprehensive delineations sequentially.

Referring to the previous study by Kolekar et al. (2020), the perceived risk is the product of the subjective probability of an event occurring and the consequences of that event. In this study, the DRF captures the probability of collision with the AV at the next timestep t as perceived by other drivers at the current position. According to (Abu-Zidan & Eid, 2015), the consequence of the collision should be represented by the impulse as:

$$I = m_{total}\ (|v_1 - v_2|) \tag{9-8}$$

where m_{total} is the total weight of the two colliding vehicles, and v_1 and v_2 are the relative velocities of the two vehicles before and after the collision. This study simplifies the collision of the two vehicles as a rigid body collision so that the relative velocity after the collision is $0\ m/s$ ($v_2 = 0\ m/s$).

The risk perceived by other vehicles can be seen as a cost to them. Therefore, the cost to others in (9-7) is obtained as follows:

$$J_{other} = I * DRF_{other} \tag{9-9}$$

With (9-3)-(9-6) and (9-8)-(9-9), this study calculates the DRF risk perceived by HDVs. Connecting with the SVO, the calculated DRF will be embedded into the MPC cost function, enabling AV to consider the benefits/costs of HDV in its planning and control.

9.3.2 Cost function and social-aware MPCC formulation

The basis of the cost function is provided by the model predictive contouring control (MPCC) formulation (Lam et al., 2010), which has been utilised in the AVs field for motion planning (Faris et al., 2022; Ferranti et al., 2019), or path generation and tracking (Liniger et al., 2015). The main idea of this approach is to track the position of the vehicle regarding a reference point on the path and to introduce a new state quantity, i.e., progress, so that it is intuitively possible to balance the maximisation of progress along the path with the minimisation of lateral, longitudinal and angular offset from the path. Furthermore, this study introduces a "far point", which is used mainly as a second reference point to only minimise contouring error which is similar to lateral error from the reference path.

The progress variable θ can be seen as the distance that the vehicle had moved. Compared with MPC, the state vector in MPCC is updated to $x_{mpcc} = [x, y, \psi, v, \theta]^T$ and the input of the model is updated by the progress rate as: $u_{mpcc} = \left[a, \delta, \dot{\theta}\right]^T$. The goal of MPCC is to maximise the progress θ and track the reference trajectory.

The contouring error E_c and the longitudinal error E_l are also linked to progress. To improve the efficiency, an approximation is adopted to calculate the two errors:

$$\widehat{E_c} = -\left(x - x_{ref}\right)\sin\left(\psi_{ref}\right) + \left(y - y_{ref}\right)\cos\left(\psi_{ref}\right) \tag{9-10a}$$

$$\widehat{E_l} = \left(x - x_{ref}\right)\cos\left(\psi_{ref}\right) + \left(y - y_{ref}\right)\sin\left(\psi_{ref}\right) \tag{9-10b}$$

$[x_{ref}, y_{ref}, \psi_{ref}]$ means the reference point on the centre line of the roundabout, which is obtained by the perception module. In addition to these two types of error, an orientation error as a penalty term is added to ensure that not only is the car positioned in the middle of the road,

but also the prediction of the vehicle movement is close to the centre line. The orientation error can be written as follows:

$$\widehat{E_o} = 1 - \left|\cos(\psi_{ref})\cos(\psi) + \sin(\psi_{ref})\sin(\psi)\right| \tag{9-11}$$

In parallel to the current reference point, the "near point" information is considered. The information from the "far point" (x_{la}, y_{la}), illustrated in green colour in **Figure 9-3**, also needs to be used as a reference to correct the contouring error of the vehicle and expand the vehicle's forward visibility. The upgraded formula can be established by referring to the previous formula for contouring error:

$$\widehat{E_{la,c}} = -(x - x_{la})\sin(\psi_{la}) + (y - y_{la})\cos(\psi_{la}) \tag{9-12}$$

In (9-12), the $[x_{la}, y_{la}, \psi_{la}]$ provides the far-point's information. This study finally combines all the errors with a linear progress maximisation reward on $\dot{\theta}$ (which is the derivation of θ) in the MPCC cost function:

$$J_{mpcc} = \sum_{k=2}^{N_P+1}\left(q_c\hat{E}_{ck}^2 + q_l\hat{E}_{lk}^2 + q_o\hat{E}_{ok}^2 + q_{la,ck}\hat{E}_{la,ck}^2\right) - \sum_{k=1}^{N_P} q_v\dot{\theta}_k \tag{9-13}$$

This part of the cost function ensures that the vehicle can follow the reference path and maximise the progress as much as possible, and $(q_c, q_l, q_o, q_{la,ck}, q_v)$ are weighting factors for every part. Minimising this J_{mpcc} loss enables the ego vehicle to track the reference trajectory accurately.

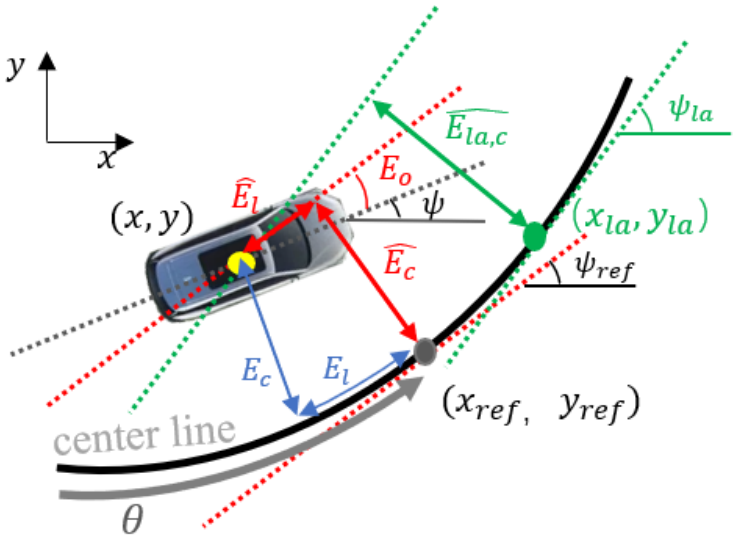


Figure 9-3. Illustration of MPCC

In addition, AVs also need to ensure the comfort of the passengers in the vehicle. The main cause of discomfort in the car is the steering wheel swinging back and forth from side to side, followed by sudden acceleration and deceleration of the AV. So, the variation in the system inputs is set to be as small as possible and the weight of δ should be bigger than the other parts. Thus, the comfort cost J_{comf} is demonstrated as

$$J_{comf} = \sum_{k=1}^{N_P}\|u_k - u_{k-1}\|_S^2 \tag{9-14}$$

Combining the two cost functions, i.e., J_{mpcc} in (9-13) and J_{comf} in (9-14), the total cost to the self-AV is obtained, which considers safety, efficiency, and comfort, as a function of (9-15):

$$J_{self} = J_{mpcc} + J_{comf} \tag{9-15}$$

Thus, according to SVO, this study combines J_{self} and J_{other} using (9-7) to obtain the total cost function J_{total} and adopts it as the objective function for social-aware MPC as in (9-1a). The inequality constraints in MPC are mainly based on the mechanical limits of the vehicle and traffic regulations, for example, the speed limit on the road, the maximum acceleration that the engine can provide, and the maximum steering angle that the steering gear can provide. This study adapts the driving style and social characteristics of AVs by adjusting the desired velocity, the weighting of the individual costs, and the SVO.

9.4 Simulation experiments and results

In this study, the architecture of the social-aware DRF-SVO-MPCC is the same as NMPC but with the redefined cost function. Moreover, this study takes the benefits/costs of surrounding vehicles into consideration tackling the risks faced by HDVs. At the same time, the MPCC was used in defining the proposed own cost, and the "far point" was introduced to make the vehicle more stable over curves with large curvature. Since the proposed DRF-SVO-MPCC integrates and outputs both planning and control simultaneously, in the simulation experiments, two test cases are carried out. Firstly, this study compares the control accuracy of the developed social-aware DRF-SVO-MPC with regard to two baselines, i.e., the pure NMPC and the well-established tracking trajectory methods pure pursuit controller (Coulter, 1992) combined with PID controller, which is simply referred to as the PP controller in this study (since the pure pursuit controller is the main part of this method). This is done by testing on the single-lane roundabout scenario with no HDVs. Secondly, this study also verifies whether the proposed method can consider other vehicles' benefits/costs and whether it can generate different driving styles under different SVOs and other parameter settings. This is done by testing on single-lane and two-lane roundabout scenarios with AVs interacting with HDVs in two different situations.

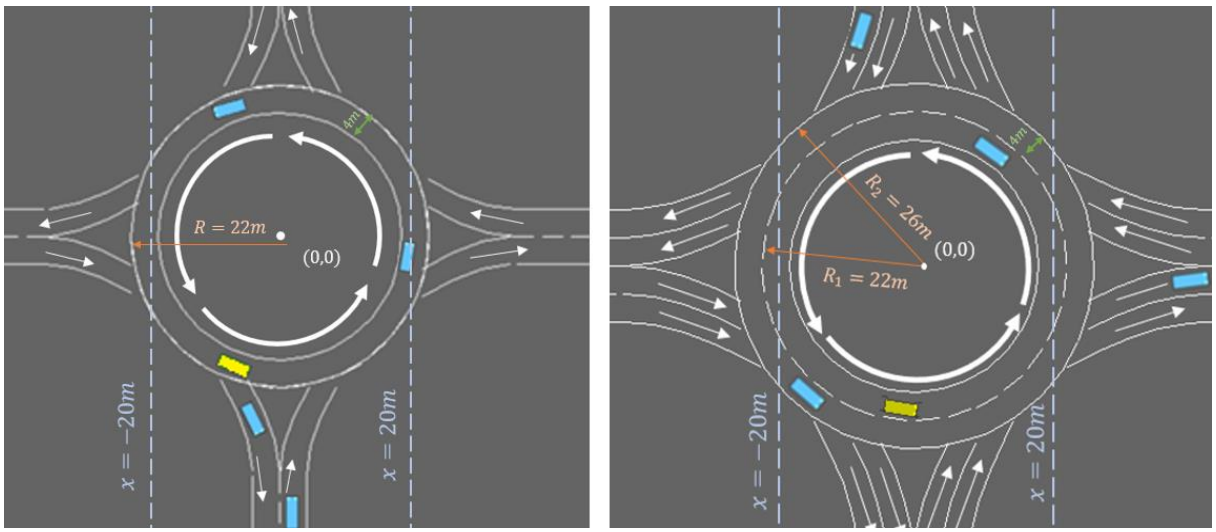
9.4.1 Controller and simulation setups

This study adopts *highway-env* (Leurent, 2018) simulation (a platform widely used in relevant publications) with Python to test the proposed approach. The examined scenarios are presented in **Figure 9-4**. In the simulation, the radius of the roundabout is 22 m, while the connection between the straight road and the roundabout is made with a curve fitted by a sine function, which is shown in **Figure 9-4**. In the simulation, the AV, indicated in the yellow colour, travels from west to east (left to right), while the HDV, indicated in the blue colour, travels from south to north (bottom to up) randomly at 3~7 m/s. The parameters of the vehicles that appear in all the scenarios are shown in **Table 9-2**. Because of the road peculiarities of roundabouts, vehicles are generally not allowed to pass through them at very high speeds, so the maximum velocity limit in the simulation is 15 m/s. The initial speed of the vehicle v_0 is set randomly within 0~3 m/s.

In the simulations, two baseline controllers, i.e., PP and NMPC controllers, together with the proposed society-aware DRF-SVO-MPCC were tested. In the PP controller, there is only a look-ahead distance that needs to be sited, and it is sited to 5 m. The parameters of NMPC and DRF-

SVO-MPCC are set as shown in **Table 9-3**. These two MPCs are solved by the optimisation solver framework CasADi (Andersson et al., 2019).

To test and verify the performance of the social-aware planning and control of the developed DRF-SVO-MPCC, three main scenarios are implemented. The first scenario focuses on only comparing the control performance of the three controllers with no other HDVs present in the roundabout, and thus, the developed DRF-SVO-MPCC will not consider social factors. In the second scenario, there will be HDV merging from other lanes of the roundabout. In the last scenario, the HDV travels from north to south (up to down) and enters the roundabout first. This study considers two different driving styles of ego AVs and compares their differences in motion planning. The common parameters of DRF are shown in **Table 9-1**, and the different parameters corresponding to the different driving styles are shown in **Table 9-4**. The bottom line in both driving styles is that no collisions can occur, so the AV driving model needs to at least consider HDV's safety cost, which means that the SVO cannot be set to 0°. Furthermore, manoeuvres of driving through both single-lane and two-lane roundabouts are simulated (**Figure 9-4**).



(a) (b)

Figure 9-4. Illustration of (a) single-lane roundabout and (b) two-lane roundabout

Table 9-1. Parameters of DRF

Parameter	p	m	k_1	k_2	t_{la}	c
Value	0.0064	0.001	0	1.3	$3s$	$0.5m$

Table 9-2. Parameters of the vehicle

Parameter	l_r	l_f	$mass$	$width$
Value	$2.46\ m$	$2.49\ m$	$2020\ Kg$	$2.0\ m$

Table 9-3. Parameters of MPC controller

Parameter	$v_{road,max}$	a_{lim}	δ_{lim}	$\Delta\delta_{lim}$	N_p
Value	$15.0\ m/s$	$3.0\ m/s^2$	30°	$30°/s$	15

Table 9-4. Parameters of MPCC in different styles

Driving Style	SVO	Desire Velocity
Prosocial	$\alpha = 60°$	$v_{ref} = 5.0\ m/s$
Egoistic	$\alpha = 15°$	$v_{ref} = 6.8\ m/s$

9.4.2 Analysis and results

In the first testing scenario, this study focuses on comparing the control accuracy and performance of the three controllers: PP controller, NMPC, and the social-aware DRF-SVO-MPCC. **Figure 9-5** shows all the trajectories controlled by the three controllers. It is easy to identify that all three AVs can follow the reference path, the centerline, to pass the roundabout. However, the PP controller gets the worst tracking performance, with a large error from the reference path (shown in **Table 9-5)**. The maximum positional error is about 3 m, which means that the bodywork of the AV is partly outside of the lane. As can be seen in **Figure 9-6**, due to the large curvature of the roundabout, the PP controller gets difficulties in trajectory tracking, resulting in large fluctuations in δ, especially when $x = \pm 20\ m$. Compared to the PP controller, the optimisation-based method, NMPC, delivers a much better tracking of the reference trajectory, except for two instances of inappropriate steering around $x = \pm 20\ m$ due to the lack of proper judgments of the future path, as shown by **Figure 9-5 (b)**. Unlike the PP controller, the NMPC is a lateral and longitudinal coupled control, and therefore a will experience waves during steering at $x = 20\ m$ and $x = -18\ m$ as shown in **Figure 9-6 (a)**. The proposed social-aware DRF-SVO-MPCC demonstrates a good solution to the above problems. As the roads are stitched together using aggregate shapes, they are not completely smooth at the road joints, however, as shown in **Figure 9-5 (c)**, the proposed DRF-SVO-MPCC not only tracks the reference trajectory well but also comes out with a smoother curve than the reference trajectory. At the same time, **Figure 9-6 (a)** shows that the social-aware DRF-SVO-MPCC can still maintain a smooth acceleration during steering with high curvature at around $x = \pm 20\ m$.

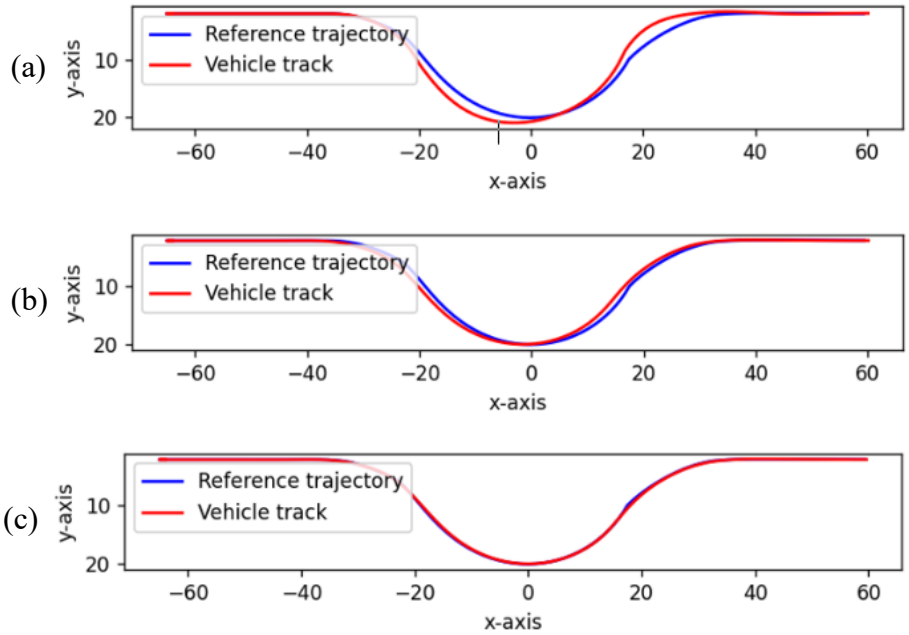


Figure 9-5. The paths obtained by using (a) PP controller, (b) NMPC, and (c) social-aware DRF-SVO-MPCC in comparison to the reference trajectory

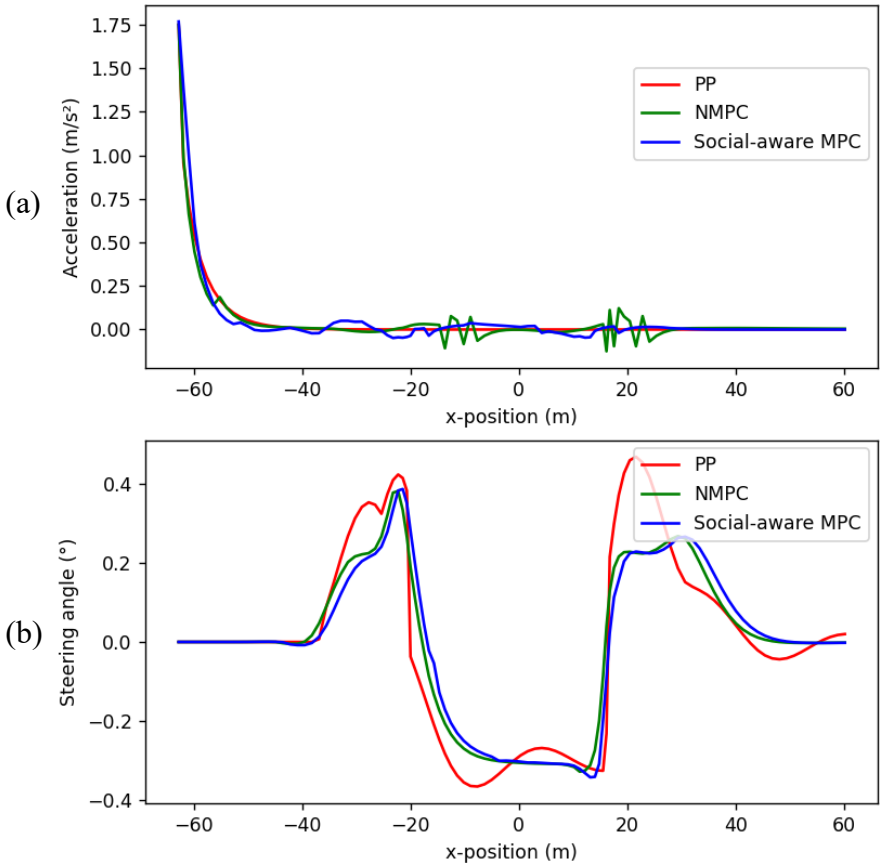
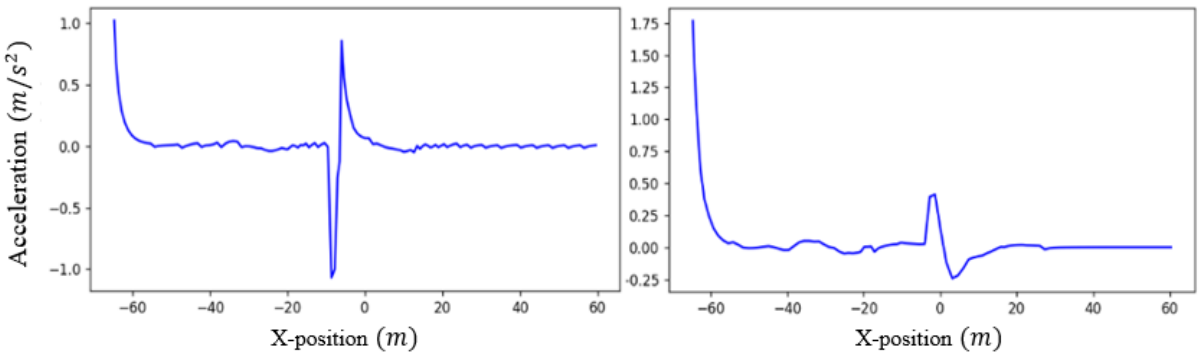


Figure 9-6. Comparison of the control inputs, i.e., (a) acceleration and (b) steering angle, in different controllers when passing the roundabout

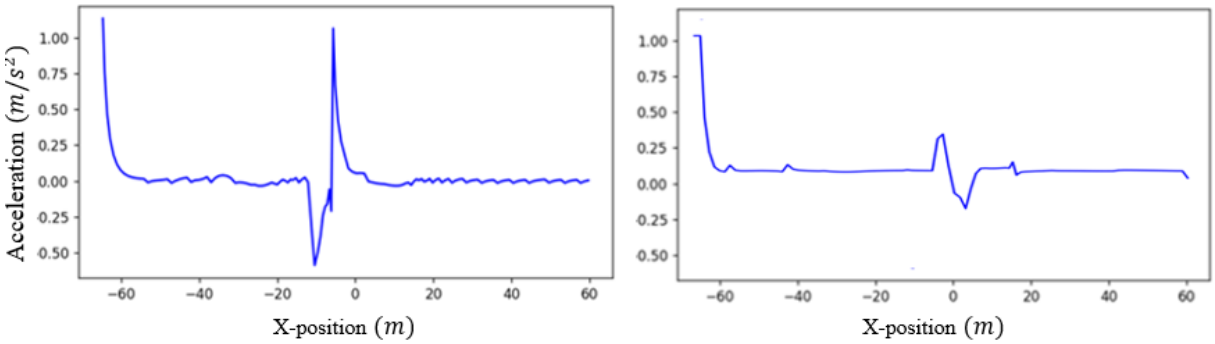
Having demonstrated the control performance of the developed social-aware DRF-SVO-MPC outperforms the two baselines, this study further compares the effects of different driving styles on the planning of the AV. In the second scenario, an aggressive HDV is added, which attempts to enter the roundabout even if the AV is already inside and running from its left. Two driving styles, i.e., prosocial and egoistic, are tested with **Figure 9-7** showing the acceleration of the AV under the two driving styles. As shown in **Figure 9-7 (a)**, under the prosocial driving style, AV will first actively slow down with $a = -1.02\mathrm{m/s^2}$ to avoid the HDV, minimising the risk to which the HDV is exposed, and then it will accelerate to v_{ref}. Conversely, an egoistic AV with a small SVO (e.g., 15^0), will be more biased to consider minimising its own costs. Thus, as in **Figure 9-7 (b)**, the AV decides to accelerate with $a = 0.43\mathrm{m/s^2}$ driving through the junction before the HDV to avoid collision and improve its efficiency through the roundabout. These statistics show that the proposed DRF-SVO-MPC can generate different driving styles while all maintaining safety. As shown in **Figure 9-4 (b)**, this study further sets up a two-lane roundabout to test the performance of the proposed DRF-SVO-MPC when the two vehicles are in different lanes. An extra lane is added with AV driving in the inner lane and HDV driving in the outer lane. **Figure 9-8 (a)** shows that the prosocial AV will still give precedence to the HDV by braking with $a = -0.52\ \mathrm{m/s^2}$, waiting to maintain a safe distance from the HDV before accelerating back to the v_{ref} to pass the roundabout safely. The choice of braking behind the HDV was made because it was calculated that there would be a greater risk to the HDV if the v_{ref} was maintained. Comparing **Figure 9-8 (a)** and **Figure 9-7 (a)**, it can be seen that, compared to HDV running in the near lane, the AV will brake more sharply when the HDV

wants to merge into the same lane. This is caused by the HDV blocking the AV's trajectory when in the same lane, which potentially poses a greater risk to both HDV and AV. The simulation demonstrates the proposed DRF-SVO-MPCC's capability to handle interacting with HDVs in different lanes separately. Similar to the single-lane roundabout case, when the driving style is egoistic, the AV will accelerate aggressively, try to change to the right lane just before the HDV, and then exit the two-lane roundabout without any deceleration throughout the whole process. This helps the AV maintain a low cost and high benefits while sacrificing the benefits of the HDV. Furthermore, it will be dangerous if the HDV is more egoistic and more aggressive, which will cause a collision.



(a) (b)

Figure 9-7. Illustration of the acceleration in different driving styles when passing the single-lane roundabout: (a) prosocial driving and (b) egoistic driving



(a) (b)

Figure 9-8. Illustration of the acceleration in different driving styles when passing the two-lane roundabout (a) prosocial driving and (b) egoistic driving

In the last scenario, HDVs enter the roundabout first, and the AV plans to merge into the roundabout afterwards. Because of safety and traffic rules, AVs in both driving styles will brake to avoid collision with HDVs, and this study compares the planning of the different driving styles. As shown in **Table 9-6**, the egoistic AV will slow down as late as possible, keeping only a minimum of 3.65 m from the HDV for safety and maintaining a higher velocity compared to the prosocial driving style. While the prosocial AV starts slowing down earlier at 18.22 m from the HDV and keeps a longer distance to the HDV of 8.49 m. The results show that the prosocial AV focuses more on minimising the risk, and it places more weight on the benefit of HDVs. On the contrary, the egoistic AV aims to minimise its own costs while ensuring the safety of both vehicles.

All the quantitative results are wrapped up and shown in **Table 9-5** and **Table 9-6**. And all the testing scenarios are better demonstrated in the supplementary video with a description document, which can be viewed at https://lnkd.in/g_MDNs5F.

Table 9-5. Quantitative results of the experiments (AV enters the roundabout first)

Scenarios	Method	Driving Styles	Max Positional Error	Average Positional Error	Collision
Single-lane roundabout with no HDV	PP Controller	---	3.08 m	1.37 m	---
	NMPC	---	1.27 m	0.65 m	---
	DRF-SVO-MPCC	---	**0.23 m**	**0.12 m**	---
Single-lane roundabout interacting with an HDV	NMPC	---	---	---	Yes
	DRF-SVO-MPCC	Prosocial	**0.19 m**	**0.09 m**	No
		Egoistic	0.28 m	0.16 m	No
Two-lane roundabout interacting with an HDV	NMPC	---	---	---	Yes
	DRF-SVO-MPCC	Prosocial	**0.26 m**	**0.17 m**	No
		Egoistic	0.34 m	0.22 m	No

Table 9-6. Quantitative results of the experiments (HDV enters the roundabout first)

Scenarios	Method	Driving Styles	Start Braking Distance	Min. Distance to HDV	Min. Velocity
Two-lane roundabout interacting with an HDV	DRF-SVO-MPCC	Prosocial	18.22 m	8.49 m	1.47 m/s
		Egoistic	13.87 m	3.65 m	3.17 m/s

9.5 Conclusion

This study develops an integrated social-aware planning and control algorithm, i.e., DRF-SVO-MPCC, which incorporates Driving Risk Field (DRF), Social Value Orientation (SVO), and Model Predictive Contouring Control (MPCC) to enable AVs to consider HDVs' risk and balance their own benefits with regards to the benefits of HDVs. The DRF is used to model the perceived risk, and SVO is adopted to measure how AVs make the trade-off between their own benefits and the benefits of other HDVs. Using the SVO-based DRF and MPCC costs, together with the desired velocity, this study implements two types of driving styles, i.e., prosocial and egoistic. The model-based DRF-SVO is packaged into the cost function established by MPCC to deliver integrated planning and control. The proposed DRF-SVO-MPCC model is tested and verified on various simulation experiments, comparing with two baselines, which demonstrates its good planning and control performance driving through both single-lane and two-lane roundabouts with or without interacting with HDVs. Future research directions could focus on the estimation of model parameters using learning-based methods. For example, the driving

style of HDVs can be estimated using a reinforcement learning approach, leading to different DRF-SVO-MPCC models to better perceive risks under the proposed framework. Furthermore, it is suggested to validate the model on other challenging driving manoeuvres (e.g., on-ramp merging, highway lane changing, or overtaking) and scenarios involving interactions with more surrounding vehicles to verify the model's robustness.

References

Abu-Zidan, F. M., & Eid, H. O. (2015). Factors affecting injury severity of vehicle occupants following road traffic collisions. Injury. https://doi.org/10.1016/j.injury.2014.10.066

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2016.110

Andersson, J. A. E., Gillis, J., Horn, G., Rawlings, J. B., & Diehl, M. (2019). CasADi: a software framework for nonlinear optimization and optimal control. Mathematical Programming Computation. https://doi.org/10.1007/s12532-018-0139-4

Buckman, N., Pierson, A., Schwarting, W., Karaman, S., & Rus, D. (2019). Sharing is caring: Socially-compliant autonomous intersection negotiation. IEEE International Conference on Intelligent Robots and Systems, 6136–6143. https://doi.org/10.1109/IROS40897.2019.8967997

Buyval, A., Gabdulin, A., Mustafin, R., & Shimchik, I. (2017). Deriving overtaking strategy from nonlinear model predictive control for a race car. IEEE International Conference on Intelligent Robots and Systems. https://doi.org/10.1109/IROS.2017.8206086

Faris, M., Falcone, P., & Sjoberg, J. (2022). Optimization-based coordination of mixed traffic at unsignalized intersections based on platooning strategy. IEEE Intelligent Vehicles Symposium, Proceedings, 2022-June(Iv), 977–983. https://doi.org/10.1109/IV51971.2022.9827149

Ferranti, L., Brito, B., Pool, E., Zheng, Y., Ensing, R. M., Happee, R., Shyrokau, B., Kooij, J. F. P., Alonso-Mora, J., & Gavrila, D. M. (2019). SafeVRU: A research platform for the interaction of self-driving vehicles with vulnerable road users. IEEE Intelligent Vehicles Symposium, Proceedings. https://doi.org/10.1109/IVS.2019.8813899

Garriga, J. L., & Soroush, M. (2010). Model predictive control tuning methods: A review. In Industrial and Engineering Chemistry Research. https://doi.org/10.1021/ie900323c

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2255–2264. https://doi.org/10.1109/CVPR.2018.00240

Hang, P., Lv, C., Huang, C., Cai, J., Hu, Z., & Xing, Y. (2020). An integrated framework of decision making and motion planning for autonomous vehicles considering social behaviors. IEEE Transactions on Vehicular Technology, 69(12), 14458–14469. https://doi.org/10.1109/TVT.2020.3040398

Hang, P., Lv, C., Xing, Y., Huang, C., & Hu, Z. (2021). Human-like decision making for autonomous driving: A noncooperative game theoretic approach. IEEE Transactions on

Intelligent Transportation Systems, 22(4), 2076–2087. https://doi.org/10.1109/TITS.2020.3036984

Ji, J., Khajepour, A., Melek, W. W., & Huang, Y. (2017). Path planning and tracking for vehicle collision avoidance based on model predictive control with multiconstraints. IEEE Transactions on Vehicular Technology. https://doi.org/10.1109/TVT.2016.2555853

Kolekar, S., de Winter, J., & Abbink, D. (2020). Human-like driving behaviour emerges from a risk-based driver model. Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-18353-4

Kong, J., Pfeiffer, M., Schildbach, G., & Borrelli, F. (2015). Kinematic and dynamic vehicle models for autonomous driving control design. IEEE Intelligent Vehicles Symposium, Proceedings. https://doi.org/10.1109/IVS.2015.7225830

Lam, D., Manzie, C., & Good, M. (2010). Model predictive contouring control. Proceedings of the IEEE Conference on Decision and Control. https://doi.org/10.1109/CDC.2010.5717042

Leurent, E. (2018). An environment for autonomous driving decision making. https://github.com/Farama-Foundation/HighwayEnv

Li, J., Sun, L., Zhan, W., & Tomizuka, M. (2020). Interaction-aware behavior planning for autonomous vehicles validated with real traffic data. ASME 2020 Dynamic Systems and Control Conference, DSCC 2020. https://doi.org/10.1115/DSCC2020-3328

Liebrand, W. B. G., & McClintock, C. G. (1988). The ring measure of social values: A computerized procedure for assessing individual differences in information processing and social value orientation. European Journal of Personality. https://doi.org/10.1002/per.2410020304

Liniger, A., Domahidi, A., & Morari, M. (2015). Optimization-based autonomous racing of 1:43 scale RC cars. Optimal Control Applications and Methods. https://doi.org/10.1002/oca.2123

Mullakkal-Babu, F. A., Wang, M., He, X., van Arem, B., & Happee, R. (2020). Probabilistic field approach for motorway driving risk assessment. Transportation Research Part C: Emerging Technologies, 118(October). https://doi.org/10.1016/j.trc.2020.102716

Obayashi, M., Uto, K., & Takano, G. (2016). Appropriate overtaking motion generating method using predictive control with suitable car dynamics. 2016 IEEE 55th Conference on Decision and Control, CDC 2016. https://doi.org/10.1109/CDC.2016.7799032

Pauls, J. H., Boxheimer, M., & Stiller, C. (2022). Real-time cooperative motion planning using efficient model predictive contouring control. IEEE Intelligent Vehicles Symposium, Proceedings, 2022-June(Iv), 1495–1503. https://doi.org/10.1109/IV51971.2022.9827063

Coulter, R. C. (1992). Implementation of the pure pursuit tracking algorithm. Carnegie Mellon University.

Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., & Rus, D. (2019). Social behavior for autonomous vehicles. Proceedings of the National Academy of Sciences of the United States of America, 116(50), 2492–24978. https://doi.org/10.1073/pnas.1820676116

Sun, L., Zhan, W., & Tomizuka, M. (2018). Probabilistic prediction of interactive driving behavior via hierarchical inverse reinforcement learning. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC. https://doi.org/10.1109/ITSC.2018.8569453

Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R., & Fallah, Y. P. (2021a). Altruistic maneuver planning for cooperative autonomous vehicles using multi-agent advantage actor-critic. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)-Workshop on Autonomous Driving: Perception, Prediction and Planning. IEEE/CVF. http://arxiv.org/abs/2107.05664

Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R., & Fallah, Y. P. (2021b). Cooperative autonomous vehicles that sympathize with human drivers. IEEE International Conference on Intelligent Robots and Systems, 4517–4524. https://doi.org/10.1109/IROS51168.2021.9636151

Ulfsjoo, C. H., & Axehill, D. (2022). On integrating POMDP and scenario MPC for planning under uncertainty - with applications to highway driving. IEEE Intelligent Vehicles Symposium, Proceedings, 2022-June (IV), 1152–1160. https://doi.org/10.1109/IV51971.2022.9827005

Wang, L., Sun, L., Tomizuka, M., & Zhan, W. (2021). Socially-compatible behavior design of autonomous vehicles with verification on real human data. IEEE Robotics and Automation Letters, 6(2), 3421–3428. https://doi.org/10.1109/LRA.2021.3061350

Wang, W., Wang, L., Zhang, C., Liu, C., & Sun, L. (2022). Social Interactions for Autonomous Driving: A review and perspectives. Foundations and Trends® in Robotics. https://doi.org/10.1561/2300000078

Yaqoob, I., Khan, L. U., Kazmi, S. M. A., Imran, M., Guizani, N., & Hong, C. S. (2020). Autonomous driving cars in smart cities: Recent advances, requirements, and challenges. IEEE Network. https://doi.org/10.1109/MNET.2019.1900120

Yoon, Y., Shin, J., Kim, H. J., Park, Y., & Sastry, S. (2009). Model-predictive active steering and obstacle avoidance for autonomous ground vehicles. Control Engineering Practice. https://doi.org/10.1016/j.conengprac.2008.12.001

Zhang, X., Liniger, A., & Borrelli, F. (2021). Optimization-based collision avoidance. IEEE Transactions on Control Systems Technology. https://doi.org/10.1109/TCST.2019.2949540

Zhang, X., Liniger, A., Sakai, A., & Borrelli, F. (2018). Autonomous parking using optimization-based collision avoidance. Proceedings of the IEEE Conference on Decision and Control. https://doi.org/10.1109/CDC.2018.8619433

10 Discussion, conclusions, perspectives, and recommendations

Abstract

This thesis aims to broaden the Operational Design Domain (ODD) to augment the capabilities of automated vehicles (AVs), thereby enabling the realisation of safe, efficient, and socially compliant automated driving within mixed-traffic environments. This thesis addresses the overall objective through three main pillars (i.e., sensing and perception, anomaly detection, as well as planning and control) and by answering the three main research questions corresponding to these three pillars. This chapter recaps the research questions, summarises the key research findings, discusses the limitations and future research recommendations with regard to each pillar, and finally highlights the overall implications and recommendations for various relevant stakeholders.

10.1 Sensing and perception

Sensing and perception module

RQ1: *How can spatial-temporal features and correlations be effectively utilised to enhance vision-based sensing and perception capabilities (e.g., lane detection), and to what extent can these capabilities be improved?*

Sub research questions:
RQ 1-1: How to develop effective sequential deep neural network architecture or mechanism to effectively capture spatial-temporal correlations?
RQ 1-2: How to speed up the training of sequential deep neural network models? What strategies can be employed?
RQ 1-3: How to make efficient use of the available data, especially the unlabelled ones?

10.1.1 Key findings and summary

To address the above research questions, vision-based lane detection was selected as the main focus. Available lane detection methods presented in the literature either focus on feature extraction in single image extraction, e.g., in (Pan et al., 2018) or employing multiple image frames to make use of the correlations among image sequences, e.g., in (Zou et al., 2020). Chapter 2 proposes a novel hybrid spatial-temporal sequence-to-one deep learning architecture to integrate the spatial convolutional neural network (SCNN) (Pan et al., 2018) for single-image feature extraction with spatial-temporal Recurrent Neural Network (RNN) modules to capture correlations and dependencies among continuous images. Under this architecture, various sequential encoder-decoder based deep neural network (DNN) model variants are developed. They utilise multiple continuous image frames as input and detect the lane lines in the last image frame. Extensive experiments that were conducted on both normal and challenging driving scenes verify the effectiveness of the designed architecture. Under the proposed architecture, even the light version of the model variants with fewer model parameters and less computational complexity outperformed existing state-of-the-art models. Post-explanations based on visualisation of the extracted low-level features further validate the proposed model architecture. It is concluded that strengthening spatial relation abstraction in every single image, combined with the employment of spatial-temporal correlations among multiple continuous image frames simultaneously, boosts vision-based lane detection performance.

In accordance with Chapter 2, Chapter 3 addresses the need for further optimisation in vision-based sensing and perception by focusing on the development of customised spatial-temporal attention mechanisms. Three attention mechanisms were designed, namely temporal attention, spatial-temporal attention, and spatial-temporal attention with fully connected layers. The designed attention mechanisms aim to enhance the utilisation of spatial-temporal correlations among different image regions in continuous frames, thus improving the accuracy and robustness of lane detection. Leveraging linear Long Short Term Memory (LSTM) neural networks (Hochreiter & Schmidhuber, 1997) connected with the proposed attention blocks, the thesis demonstrates the feasibility of lightweight and computationally efficient solutions for possible real-time detection applications. Through rigorous experimentation on four large-scale datasets and comparative analysis, the effectiveness of the proposed attention mechanisms is validated, showcasing significant improvements in lane detection performance compared to

conventional methods. The findings underscore the importance of incorporating spatial-temporal attention mechanisms to effectively capture relevant information and correlation across consecutive image frames, ultimately enhancing the reliability of vision-based sensing and perception systems in AVs.

Chapter 4 further extends the exploration of enhancing vision-based sensing and perception by introducing a self-supervised pretraining method using masked sequential autoencoders (MSAE). This method aims to make efficient use of the available data, including the unlabelled ones, to improve detection accuracy and expedite the training process of deep neural network (DNN) models developed for lane detection tasks. Additionally, a customised Focal Loss based PolyLoss is proposed to further enhance detection accuracy by tackling the defects of the commonly adopted cross-entropy based loss in handling the extreme imbalance between lane points and the background points. Through comprehensive experimentation and comparative analysis, the efficacy of the proposed pretraining method and loss function is demonstrated, showcasing significant improvements in lane detection performance under various driving scenarios and dramatically reducing the total training time. Typically, employing the MSAE-based pre-training and utilising the customised PolyLoss, the proposed model featuring the spatial-temporal attention mechanism developed in Chapter 3 demonstrates superior performance in terms of accuracy, precision, and F1-measure. It outperforms other DNN-based counterparts by a significant margin. The findings highlight the importance of leveraging self-supervised learning techniques and tailored loss functions to enhance the robustness and efficiency of vision-based sensing and perception systems in AVs.

To summarise, in essence, the hybrid spatial-temporal DNN architecture, which combines robust single-image feature extraction with spatial-temporal modules to capture correlations and dependencies among continuous images, along with customised spatial-temporal attention mechanisms aimed at enhancing the extraction and utilisation of spatial-temporal correlations among different image regions in continuous frames, together with MSAE-based self-supervised pre-training and tailored PolyLoss, collectively contribute to the advancement of vision-based sensing and perception capabilities.

10.1.2 Discussion of limitations and recommendations

Regarding the limitations in the sensing and perception task, it is important to acknowledge that while the proposed models exhibit promising performance, they may still encounter challenges when faced with scenarios significantly different from those present in the training dataset. These challenges can stem from scenarios that pose difficulties even for human annotators to correctly identify lanes, leading to potentially inaccurate or inadequate labels for such complex scenes (Zhang et al., 2022), thus misleading the model. To address this limitation, there is a pressing need to develop an integrated high-quality dataset specifically tailored to encompass such challenging driving scenes. Such an integrated dataset would serve as a valuable resource for enhancing the robustness and generalisability of lane detection models across diverse and complex environments. Moreover, the adoption of advanced methodologies like few-shot

learning[5] (Majee et al., 2021; Su et al., 2022) and contrastive learning[6] (J. Li et al., 2022; Radford et al., 2021; Z. Zhou et al., 2023) holds significant potential in advancing the field of vision-based sensing and perception to address the aforementioned problem. Few-shot learning enables models to adapt and generalise effectively to novel scenarios using limited annotated data samples, reducing the reliance on extensive labelled datasets. Contrastive learning, exemplified by ZegCLIP (Z. Zhou et al., 2023), pushes the boundaries further. ZegCLIP excels in zero-shot segmentation by leveraging large-scale, unannotated data and aligning images with textual descriptions to achieve accurate scene understanding without the need for extensive manual labelling. These methods mitigate the reliance on large-scale labelled datasets and facilitate more efficient model training and deployment.

Furthermore, an intriguing direction for future research lies in investigating the domain adaptation (Hu et al., 2022; C. Li et al., 2022) capabilities of lane detection models. This involves training the models on one dataset and subsequently evaluating their performance on a disparate dataset, particularly one sourced from a different geographical region or driving context. By assessing the transferability and adaptability of the models across diverse datasets, insights can be gathered into their robustness and suitability for real-world deployment in varied driving environments.

For practical recommendations, it is noteworthy to highlight the geographical disparity in lane detection dataset availability, with a predominant concentration of datasets originating from North America and Asia. This discrepancy underscores the need for concerted efforts to address the dataset gap, particularly in European countries. Establishing comprehensive and regionally diverse datasets is essential for fostering inclusive and globally applicable research in autonomous driving technologies.

Finally, although the models and methods presented in the sensing and perception pillar were developed specifically for the lane detection task, they can be customised and adapted to other vision-based sensing and perception tasks (e.g., object detection and tracking) as well.

[5] *Few-shot learning* is an example of meta-learning, where a learner undergoes training across various related tasks during the meta-training phase, which enables it to generalise proficiently to unseen, yet related tasks with minimal examples during the testing phase. An effective strategy for tackling the "few-shot learning" challenge involves acquiring a common representation for diverse tasks and subsequently training task-specific classifiers based on this representation. Adapted from https://paperswithcode.com/task/few-shot-learning.

[6] *Contrastive learning* is a deep learning technique for unsupervised representation learning that aims to map similar data instances close together and dissimilar ones far apart in the representation space. It has proven effective and powerful in various computer vision and natural language processing tasks like image retrieval, zero-shot learning, and cross-modal retrieval, where the learned representations serve as features for downstream tasks like classification, segmentation, and clustering. Adapted from https://paperswithcode.com/task/contrastive-learning.

10.2 Anomaly detection

Anomaly detection module

RQ2: *How to develop effective semi-supervised/unsupervised machine learning methods for anomaly detection leveraging unlabelled data?*

Sub research questions:
RQ 2-1: What are the key features for anomaly detection, and how can they be identified?
RQ 2-2: How to develop pipeline and method to make efficient use of unlabelled data for enhancing anomaly detection?

10.2.1 Key findings and summary

To address anomaly detection in automated driving, two case studies were carried out in Chapters 5 and 6, respectively.

Chapter 5 introduces a novel approach leveraging Transformer-based models (Dosovitskiy et al., 2021; Liu et al., 2021) with self-supervised pretraining and customised fine-tuning for intelligent anomaly detection in lane rendering images of digital map applications. There are seven types of anomalies, including *(a): the road centre line extends out of the junction; (b) the stop line is in the middle of a road; (c) the navigation route does not match actual roads; (d) the road shoulder is bumpy; (e) a part of the road is missing; (f) the road marking arrows overlap; (g) the lane lines overlap*. This chapter firstly transforms lane rendering image anomaly detection into a classification problem and then proposes a four-phase pipeline encompassing data pre-processing, self-supervised pre-training using masked image modelling (MiM) (He et al., 2022a; Xie et al., 2022), customised fine-tuning with cross-entropy loss and label smoothing, and post-processing. Experimental results demonstrate the pipeline's effectiveness, with significant improvements in detection accuracy and reduced training time achieved through self-supervised pre-training with MiM. For instance, employing Swin Transformer (Liu et al., 2021) with Uniform Masking (UM) as self-supervised pre-training (Swin-Trans-UM) yielded an accuracy of 94.77% and an Area Under The Curve (AUC) of 0.9743, compared to 94.01% accuracy and an AUC of 0.9498 without pre-training (Swin-Trans), while reducing the fine-tuning epochs from 280 to 41. Ablation studies further validate the pipeline's performance enhancements, particularly in addressing data imbalance between normal and abnormal instances. This approach not only enhances anomaly detection accuracy but also contributes to reducing labour costs associated with manual labelling and manual anomaly detection efforts, thereby offering significant societal benefits.

Additionally, Chapter 6 explores the crucial task of detecting abnormal driving behaviour. While many existing machine learning (ML) models rely on fully supervised methods, requiring substantial labelled data, this thesis addresses the need for more feasible and efficient approaches by exploring semi-supervised methods. Leveraging large-scale real-world driving data in the CitySim dataset (Zheng et al., 2023), the study identifies various abnormal driving behaviours and develops a semi-supervised ML method based on the Hierarchical Extreme Learning Machine (HELM). This novel approach utilises partly labelled data for accurate detection and introduces Surrogate Measures of Safety (SMoS) as input features to enhance performance. Results from extensive experiments demonstrate the effectiveness of the proposed

semi-supervised ML model, showcasing superior performance compared to other semi-supervised baseline methods. The integration of SMoS, particularly the event-based safety indicators of the Two-Dimensional Time-To-Collision (2D-TTC), significantly improves detection accuracy, highlighting the pivotal role of SMoS in enhancing model performance. By leveraging unlabelled data for training and only a small sample of labelled data for fine-tuning, the proposed semi-supervised approach achieves competitive performance while reducing dependency on fully labelled datasets, making it suitable for real-world applications with limited labelled data. The findings also underscore the critical value of event-based safety indicators in effectively detecting abnormal driving behaviours, with significant implications for safety-oriented research and evaluations.

To sum up, the exploration of semi-supervised and self-supervised machine learning methods in anomaly detection represents promising avenues for addressing the inherent limitations of fully supervised approaches, which heavily rely on extensive accurately labelled data for training. The pioneering research presented in this thesis represents a significant stride towards enhancing safety in driving environments through the utilisation of data-driven ML-based anomaly detection methodologies.

10.2.2 Discussion of limitations and recommendations

While the proposed semi-supervised ML-based anomaly detection approaches showcased promising results, certain baseline semi-supervised ML methods demonstrated inefficacy in specific use cases, potentially attributed to the unique characteristics of the use case and the employed datasets. Further exploration into these discrepancies is warranted in future studies to better understand the underlying factors influencing semi-supervised ML model performance.

Additionally, it is important to acknowledge that semi-supervised ML methods still necessitate portions of the data to be labelled with ground truth, thereby imposing constraints on scalability and resource efficiency. To address this limitation, future research efforts should focus on the continuous refinement and development of more efficient unsupervised ML techniques (Usmani et al., 2022), aiming to mitigate the reliance on (large-scale) labelled datasets. Furthermore, future investigations can delve into few-shot learning approaches (Wang et al., 2022; X. Zhou et al., 2021), which enable models to generalise effectively from a small number of labelled examples and hold significant promise for extending the applicability of anomaly detection models to new and unseen scenarios. By leveraging the inherent structure and relationships within the data, few-shot learning techniques have the potential to mitigate the dependence on extensive labelled datasets while ensuring accurate predictions with limited supervision.

While this thesis primarily concentrates on anomaly detection within automated driving applications, future research endeavours could venture into predictive modelling for identifying abnormal behaviours and situations before they occur or at an early stage. Furthermore, additional research is needed to develop techniques that extract robust spatial-temporal patterns as inputs for anomaly detection models. Integrating a broader range of anomalies, such as more diverse abnormal driving behaviours, and incorporating more pertinent features (e.g., other advanced safety indicators), could enhance our understanding and identification capabilities. Consequently, these advancements would significantly contribute to the advancement of data-

driven monitoring of abnormal behaviours and situations, thereby bolstering road traffic and transportation safety.

10.3 Planning and control

Planning and control module

RQ3: *How to develop and optimise automated vehicles' driving strategies and styles to ensure safety, efficiency, and, particularly, social compliance in mixed-traffic environments?*

Sub research questions:
RQ 3-1: How can social norms and driving-related benefits for human-driven vehicles be effectively integrated into the development of automated driving strategies?
RQ 3-2: How do different deep reinforcement learning algorithms perform across different driving manoeuvres?
RQ 3-3: How can model performance be comprehensively evaluated and compared, particularly in terms of their adaptability to handle scenario shifts?

10.3.1 Key findings and summary

To address the above research questions concerning AVs' planning and control in the mixed-traffic context, an integrated conceptual framework for socially compliant automated driving (Schwarting et al., 2019) is developed based on a comprehensive literature review, as outlined in Chapter 7. The framework incorporates various social components, including cultural differences, norms, and cues, alongside different driving styles (e.g., aggressive, cautious, pro-social). A novel concept of bidirectional behavioural adaptation is introduced within this framework, emphasising the dynamic interactions and adaptations between AVs and human drivers, i.e., human drivers will have already adapted their driving behaviour when interacting with AVs, and AVs need to adapt to human drivers' behavioural adaptation. Moreover, the proposed framework underscores the importance of balancing the benefits of AVs with the needs and expectations of other road users, particularly in terms of safety, comfort, and efficiency, highlighting the necessity for a nuanced trade-off strategy on a case-by-case basis. Additionally, the framework proposes the implementation of a spatial-temporal memory module to facilitate long-term and short-term knowledge and rule upgrading. This module enables the regular refinement of driving strategies that consider bidirectional behavioural adaptation. Furthermore, an online questionnaire-based survey is conducted to gather expert insights and feedback on the proposed conceptual framework, assessing its validity and effectiveness. The results provided valuable validation and refinement of the framework's components, along with insightful suggestions for improvement. Overall, Chapter 7 lays the groundwork for developing socially compliant automated vehicles by offering a structured conceptual framework. This framework serves as a guiding tool for the implementation of learning-based [Chapter 8] and model-based [Chapter 9] approaches in this thesis and holds the potential for informing future research endeavours in this domain.

In the learning-based approach, Chapter 8 explores the application of Deep Reinforcement Learning (DRL) in automated driving, with a focus on integrating considerations of safety, efficiency, comfort, and energy consumption into the learning framework. Multiple DRL

algorithms, including Deep Q-Network (DQN), Deep Deterministic Policy Gradient (DDPG), Proximal Policy Optimisation (PPO), and Trust Region Policy Optimisation (TRPO), are employed and evaluated for their effectiveness in guiding automated vehicles through various scenarios. The thesis emphasises the importance of considering real-world requirements in the reward function design and simulation-based training and verification to ensure the safety and efficacy of the learning-based approach. Evaluation and comparison of DRL algorithms, such as DQN, DDPG, PPO, and TRPO, are conducted across various driving manoeuvres, including highway merging and unsignalised intersections and particularly roundabout driving. Results indicate that TRPO outperforms other algorithms in terms of safety and efficiency, while PPO excels in comfort level for roundabout driving. Furthermore, to train a uniform driving model that can tackle various driving manoeuvres, this thesis expands the *highway-env* (Leurent, 2018) and develops an extra customised training environment, namely, "*ComplexRoads*", integrating various driving manoeuvres and multiple road scenarios together. Models trained on the designed *ComplexRoads* environment show promising adaptability to other driving scenarios with overall good performance. As a preliminary step, this thesis represents a pioneering effort in conducting a comprehensive DRL model performance evaluation, particularly considering scenario shifting (Hauer et al., 2019), i.e., models trained on one scenario but evaluated on other scenarios. Collectively, the findings highlight the potential of DRL-based automated driving in addressing complex traffic scenarios, offering meaningful insights for future research towards developing automated driving with DRL and simulation.

Regarding the model-based approach, previous research has predominantly concentrated on trajectory planning for AVs using Model Predictive Control. However, there has been a significant gap in integrated planning and control methods, particularly in socially aware driving scenarios. This thesis aims to address these research gaps by developing an algorithm that enhances the understandability and predictability of AVs to human drivers, especially during AVs' interactions with human-driven vehicles (HDVs). Chapter 9 presents an integrated social-aware planning and control algorithm, termed DRF-SVO-MPCC, which incorporates three interdisciplinary concepts: perceived Driving Risk Field (DRF), Social Value Orientation (SVO), and Model Predictive Contouring Control (MPCC). This integration enables AVs to consider the welfare of HDVs on the road, balancing their own benefits with those of HDVs, particularly during challenging manoeuvres such as driving through roundabouts. The designed DRF-SVO-MPCC algorithm undergoes testing and verification through simulation experiments on various driving scenarios using the open-sourced *highway-env* platform (Leurent, 2018). Initially, the thesis compares the control accuracy and performance of three controllers, i.e., Pure Pursuit (PP) controller, Nonlinear Model Predictive Control (NMPC), and the proposed DRF-SVO-MPCC, on different types of roundabouts (e.g, single-lane, two-lane) with and without surrounding HDVs. The results demonstrate that DRF-SVO-MPCC outperforms the other controllers, achieving smoother trajectory tracking and better handling of challenging driving conditions. Furthermore, this thesis investigates the effects of different driving styles, such as prosocial and egoistic behaviours, on AV planning. The findings indicate that the DRF-SVO-MPCC algorithm can generate different driving styles while maintaining safety, potentially enabling AVs to adapt their behaviour based on the prevailing social context. Extensive testing results demonstrate the proposed model's robustness and its superior performance compared to baseline methods, particularly the AVs using DRF-SVO-MPCC can dynamically adjust their behaviour, prioritising safety and social considerations while optimising their own benefits. Overall, Chapter 9 highlights the effectiveness of the model-

based DRF-SVO-MPCC algorithm in enabling AVs to navigate mixed-traffic environments both safely and in a socially responsible manner, paving the way for further advancements in social-aware automated driving systems.

10.3.2 Discussion of limitations and recommendations

While the methods developed in this thesis primarily focus on controlling a single AV, they can be customised and upgraded for the planning and control of multiple AVs simultaneously. This scalability is crucial for future urban mobility scenarios where fleets of AVs will operate in tandem to optimise traffic flow and ensure safety.

Furthermore, the current model-based approach only focuses on the interaction between one AV and one HDV, it is imperative to recognise the intricate nature of real-world traffic situations, particularly in densely congested urban environments. In such dynamic settings, multiple interactions between AVs and HDVs must be comprehensively considered to ensure smooth and safe navigation. For real-world urban applications, the inclusion of other road participants, such as cyclists and pedestrians, is indispensable. Integrating these diverse elements, especially vulnerable road users, into automated driving systems is essential for creating holistic solutions that cater to the complexities of urban traffic scenarios, prioritising safety, efficiency, and benefits for all road users.

In addition to the current separate implementation of model-based and learning-based approaches in this thesis, future endeavours are advised to integrate these methodologies simultaneously into a unified framework. By combining model-based and learning-based techniques, a synergistic approach can be achieved, leading to the development of more robust and adaptive automated driving systems. The integration of these approaches holds the potential to capitalise on the strengths of both methodologies: Model-based techniques offer a structured and rule-based framework for decision-making and control, leveraging explicit models of the environment and vehicle dynamics; On the other hand, learning-based methods, particularly deep reinforcement learning approaches, excel in capturing complex patterns and behaviours from large datasets and intensive simulation, enabling adaptation to diverse, dynamic, and challenging environments. By merging these approaches, AVs can benefit from the precision and reliability of model-based planning and control, while also harnessing the adaptability and scalability of learning-based methods. This integration would facilitate the development of comprehensive systems capable of handling a wide range of traffic scenarios efficiently and safely, ultimately advancing the realisation of AVs in real-world settings.

Lastly, for future research, there is a pressing need to transition towards the development of a unified driving model, as opposed to addressing individual driving manoeuvres on a case-by-case basis. This shift in approach would entail the creation of a holistic framework that encapsulates diverse driving scenarios and behaviours within a singular, comprehensive model. A unified driving model will promote consistency and coherence in the behaviour of AVs across various situations, thereby enhancing predictability and reliability. Furthermore, transitioning to a uniform driving model enables more efficient utilisation of resources and expertise in the development and testing phases. Rather than developing specialised algorithms for each specific manoeuvre or scenario, researchers and engineers can focus on refining a single, overarching model that encompasses the entire spectrum of driving tasks. Ultimately, by establishing a standardised framework that accommodates diverse real-world scenarios, the

adoption of a uniform driving model lays the foundation for safer, more efficient, as well as socially responsible and compliant automated driving solutions.

10.4 Overall conclusions

This thesis has made significant contributions to the field of automated driving, focusing on three key pillars: sensing and perception, anomaly detection, as well as planning and control. Through a combination of theoretical frameworks, methodological innovations, and data-driven empirical evaluations, the research presented herein has advanced our understanding and capabilities in developing safe, efficient, and socially compliant automated driving systems.

To enhance the ***sensing and perception*** capabilities of AVs, this thesis introduces two novel hybrid spatial-temporal DNN architectures for vision-based lane detection, emphasising the integration of single-image feature extraction with spatial-temporal correlations among continuous images. These architectures, coupled with customised spatial-temporal attention mechanisms, self-supervised pretraining techniques, and tailored loss function, demonstrate remarkable performance improvements in lane detection accuracy and robustness across diverse driving scenarios. By leveraging the proposed methodologies, the thesis underscores the importance of enhancing spatial relation abstraction and spatial-temporal correlations simultaneously to bolster vision-based sensing and perception capabilities in automated vehicles.

Furthermore, the thesis addresses the challenge of ***anomaly detection*** in automated driving applications through the extensive exploration of supervised, semi-supervised, and self-supervised machine learning methods. From Transformer-based models for intelligent anomaly detection in lane rendering images to novel semi-supervised ML approaches for abnormal driving behaviour detection, the research offers valuable insights into enhancing the safety of driving. The findings highlight the efficacy of leveraging semi-supervised and self-supervised learning techniques to mitigate the reliance on extensive labelled datasets, paving the way for scalable and efficient anomaly detection solutions.

Finally, to advance AVs' ***planning and control***, the thesis presents an integrated conceptual framework for socially compliant automated driving, emphasising bidirectional behavioural adaptation and the balance between AVs' benefits and the needs of other road users. Model-based and learning-based approaches for implementations on simulated environments further verify and extend these concepts, demonstrating the effectiveness of social-aware planning and control algorithms in navigating mixed-traffic environments safely, efficiently, and responsibly. Through comprehensive evaluations of deep reinforcement learning algorithms and the preliminary exploration of a unified driving model, the research showcases the potential of simulation-based learning approaches in optimising AV behaviours across diverse driving scenarios. Furthermore, the incorporation of social psychological factors, i.e., Social Value Orientation, enables a model-based implementation that accommodates different driving styles, e.g., prosocial and egoistic. The prosocial driving style allows AVs to navigate complex traffic scenarios while balancing their own safety and efficiency with the benefits of surrounding HDVs. This research represents an early but significant step toward implementing socially compliant automated driving, highlighting the importance of integrating social factors into AV decision-making to enhance acceptance and safety in real-world mixed-traffic environments.

Limitations and Recommendations:

While the research presented in this thesis represents significant advancements in the field, several limitations and opportunities for future research have been identified. These include the need for more extensive datasets encompassing diverse driving scenarios, the exploration of domain adaptation techniques, and the integration of model-based and learning-based methodologies. Moreover, transitioning towards the development of a unified driving model and addressing the challenges of multi-agent interactions in complex urban environments remain critical areas for future investigation.

Overall, this thesis makes substantial contributions to the advancement of automated driving technologies, spanning vision-based sensing and perception, anomaly detection, and planning and control. By integrating innovative methodologies with rigorous empirical evaluations, the research presented herein lays the groundwork for developing safer, more efficient, and socially responsible automated driving systems. Moving forward, continued interdisciplinary research efforts are essential to address the remaining challenges and realise the full potential of automated vehicles in transforming the future of transportation.

10.5 Implementations and recommendations

The methods, findings, and contributions outlined in this thesis offer valuable insights and recommendations for various stakeholders involved in the development, implementation, and regulation of AVs in mixed-traffic environments. To be specific, the following implementation strategies and recommendations are proposed:

(1) ***Integration of Advanced Deep Learning Based Sensing and Perception Technologies:*** Original Equipment Manufacturers (OEMs) and automotive technology developers are encouraged to integrate advanced deep learning based sensing and perception technologies, such as the hybrid spatial-temporal DNN models with spatial-temporal attention mechanism developed in this thesis, into their vehicles. These models, along with the proposed self-supervised pretraining method and customised loss function, can significantly enhance vision-based sensing and perception capabilities, improving accuracy while reducing model complexity. By incorporating these technologies, OEMs can enhance the perception capabilities and pave the way for safer and more reliable AVs.

(2) ***Enhanced Road Maintenance Practices:*** Road maintenance operators are encouraged to adopt the developed lane detection methods to streamline lane marking inspection and maintenance processes. Automation or semi-automation of lane detection tasks can lead to cost savings, improved productivity, and enhanced road safety by ensuring clear and well-maintained lane markings.

(3) ***Adoption of Anomaly Detection Systems:*** OEMs and automotive safety stakeholders are advised to adopt anomaly detection systems for early detection and prediction of abnormal situations in vehicle systems and driving behaviour patterns. The semi-supervised, self-supervised, and fully-supervised machine learning techniques for anomaly detection offer opportunities for predictive maintenance strategies and proactive interventions to mitigate potential risks. By implementing these systems, stakeholders can enhance overall system reliability and contribute to safer traffic flow and transportation systems.

(4) ***Implementation of Driving Monitoring Systems:*** Based on the developed abnormal driving behaviour detection methods, authorities and insurance companies should consider implementing monitoring systems in a privacy-protective way, e.g., under the General Data Protection Regulation (GDPR). These systems can support driver training initiatives, insurance pricing strategies, and accident prevention efforts.

(5) ***Development of Socially Compliant AVs:*** Car manufacturers and AV developers should leverage the proposed conceptual framework for socially-compliant automated driving to design AVs that interact seamlessly with other road users. By systematically integrating social components (such as culture, social norms, and cues), driving styles (e.g., aggressive and defensive, selfish and prosocial), and bidirectional behavioural adaptation into AV design, manufacturers can ensure that AVs prioritise safety, efficiency, and social responsibility in their interactions with surrounding vehicles and other road users.

(6) ***Considerations of Multi-Vehicle Interactions***: This thesis primarily focused on one-and-one interaction, i.e., one AV interacting with one HDV. In future studies, AV developers should prioritise the modelling and simulation of multi-vehicle interactions to address the complexities of mixed-traffic environments. Multi-agent reinforcement learning (MARL) techniques could be explored to enhance AVs' decision-making capabilities in scenarios involving numerous interacting vehicles, such as roundabouts and urban intersections. This would ensure AVs can navigate cooperatively while minimising disruptions and risks. Additionally, the design of AVs should include adaptive algorithms capable of dynamically responding to the intentions and behaviours of multiple road users in real time.

(7) ***Consideration of Vulnerable Road Users in the Design and Development:*** Ensuring the safety of vulnerable road users (VRUs), such as pedestrians and cyclists, is also paramount in the development of AVs. Advanced sensing technologies, such as LiDAR and thermal imaging, combined with predictive models of VRU behaviour, can enhance the ability of AVs to detect and respond to VRUs effectively. Incorporating ethical decision-making frameworks that prioritise the safety of VRUs is also essential. Collaboration with urban planners and traffic engineers is recommended to design infrastructure that accommodates VRUs alongside AVs, fostering safer and more inclusive transportation systems.

(8) ***Unified End-to-end Research Framework:*** This thesis addresses various tasks utilising distinct datasets and diverse use cases. For future research endeavours, it is recommended to concentrate on an integrated setting, employing a consistent inclusive dataset, and progressing seamlessly from sensing and perception, anomaly detection, to planning and control, covering the entire spectrum of automated driving functionality. This approach can provide a comprehensive understanding and evaluation of AV systems within a unified framework, facilitating deeper insights and advancements in automated driving technology. To accomplish this, close cooperation with dataset providers such as OEMs and automated driving companies, e.g., Waymo, is imperative. Such collaboration ensures access to high-quality, integrated datasets necessary for comprehensive research and development.

(9) ***Multidisciplinary Research Cooperation:*** Collaboration across disciplines is crucial to advancing socially compliant automated driving. Multidisciplinary research cooperation among, e.g., engineering, human factors, social psychology, ethics, and urban planning experts, can provide valuable insights and solutions to complex challenges in AV development and deployment.

(10) ***Policy Formulation for AV Deployment:*** Policymakers are urged to develop comprehensive strategies and regulatory frameworks for the deployment of socially compliant AVs. By considering social factors such as culture, norms, and stakeholder engagement, policymakers can promote the equitable integration of AV technology into existing transportation systems while ensuring adherence to socially acceptable norms and behaviours.

(11) ***Public Awareness and Engagement:*** Efforts should be made to increase public awareness and understanding of AV technology as well as its benefits and implications for transportation systems. By fostering informed dialogue and engagement among stakeholders, including human drivers, OEMs, and the general public, misconceptions and concerns surrounding AV technology can be addressed, paving the way for a smoother transition and widespread adoption and acceptance of AV technology in real-world environments.

Overall, it is suggested that relevant stakeholders make reasonable plans for prioritising the implementation of these recommendations through collaborative efforts and strategic planning to foster the development of technologically advanced and socially responsible AVs. By strategically allocating resources and efforts to integrate these recommendations into existing frameworks, stakeholders can lay the groundwork for a future where automated vehicles play a pivotal role in shaping a safer, more efficient, more sustainable and inclusive transportation system.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021 - 9th International Conference on Learning Representations.

Hauer, F., Schmidt, T., Holzmuller, B., & Pretschner, A. (2019). Did we test all scenarios for automated and autonomous driving systems? 2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019. https://doi.org/10.1109/ITSC.2019.8917326

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR52688.2022.01553

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, C., Hudson, S., Ethier, M., Al-Sharman, M., Rayside, D., & Melek, W. (2022). Sim-to-real domain adaptation for lane detection and classification in autonomous driving. In 2022 IEEE Intelligent Vehicles Symposium (IV) (pp. 457-463). IEEE. https://doi.org/10.1109/IV51971.2022.9827450

Leurent, E. (2018). An environment for autonomous driving decision-making. Accessed 2024-05-09 from https://github.com/eleurent/highway-env

Li, C., Zhang, B., Shi, J., & Cheng, G. (2022). Multi-level domain adaptation for lane detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. https://doi.org/10.1109/CVPRW56347.2022.00484

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning (pp. 12888-12900). PMLR.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the IEEE International Conference on Computer Vision. https://doi.org/10.1109/ICCV48922.2021.00986

Majee, A., Agrawal, K., & Subramanian, A. (2021). Few-shot learning for road object detection. Proceedings of Machine Learning Research, 140, 115–126.

Pan, X., Shi, J., Luo, P., Wang, X., & Tang, X. (2018). Spatial as deep: Spatial CNN for traffic scene understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of Machine Learning Research.

Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., & Rus, D. (2019). Social behavior for autonomous vehicles. Proceedings of the National Academy of Sciences of the United States of America, 116(50), 2492–24978. https://doi.org/10.1073/pnas.1820676116

Su, B., Zhang, H., Wu, Z., & Zhou, Z. (2022). FSRDD: An efficient few-shot detector for rare city road damage detection. IEEE Transactions on Intelligent Transportation Systems, 23(12), 24379–24388. https://doi.org/10.1109/TITS.2022.3208188

Usmani, U. A., Happonen, A., & Watada, J. (2022). A review of unsupervised machine learning frameworks for anomaly detection in industrial applications. Lecture Notes in Networks and Systems. https://doi.org/10.1007/978-3-031-10464-0_11

Wang, Z., Zhou, Y., Wang, R., Lin, T. Y., Shah, A., & Lim, S. N. (2022). Few-shot fast-adaptive anomaly detection. Advances in Neural Information Processing Systems, 35, 4957-4970.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., & Hu, H. (2022). SimMIM: A simple framework for masked image modeling. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR52688.2022.00943

Zhang, J., Deng, T., Yan, F., & Liu, W. (2022). Lane detection model based on spatio-temporal network with double convolutional gated recurrent units. IEEE Transactions on Intelligent Transportation Systems, 23(7), 6666–6678. https://doi.org/10.1109/TITS.2021.3060258

Zheng, O., Abdel-Aty, M., Yue, L., Abdelraouf, A., Wang, Z., & Mahmoud, N. (2023). CitySim: A drone-based vehicle trajectory dataset for safety-oriented research and digital twins. Transportation Research Record. https://doi.org/10.1177/03611981231185768

Zhou, X., Liang, W., Shimizu, S., Ma, J., & Jin, Q. (2021). Siamese neural network based few-shot learning for anomaly detection in industrial cyber-physical systems. IEEE Transactions on Industrial Informatics. https://doi.org/10.1109/TII.2020.3047675

Zhou, Z., Lei, Y., Zhang, B., Liu, L., & Liu, Y. (2023). ZegCLIP: Towards adapting CLIP for zero-shot semantic segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR52729.2023.01075

Zou, Q., Jiang, H., Dai, Q., Yue, Y., Chen, L., & Wang, Q. (2020). Robust lane detection from continuous driving scenes using deep neural networks. IEEE Transactions on Vehicular Technology, 69(1), 41–54. https://doi.org/10.1109/TVT.2019.2949603

About the author

Yongqi Dong was born in Dongming, Shandong Province, China, in 1991. He obtained his Bachelor's degree in Telecommunication Engineering with distinction (outstanding thesis) from Beijing Jiaotong University in 2014, where he actively participated in various academic competitions and garnered several provincial, national, and international awards. In 2017, he obtained his Master's degree in Control Science and Engineering from Tsinghua University, where he minored in Big Data. His research during this period focused on data-driven shared and smart mobility. Additionally, he completed a one-month research internship at the Singapore-MIT Alliance for Research and Technology (SMART) from August to September 2016.

In 2020, Yongqi joined the Department of Transport and Planning at Delft University of Technology (TU Delft) to conduct his Ph.D. research on *Safe, Efficient, and Socially Compliant Automated Driving in Mixed Traffic*, as part of the NWO project titled "*Safe and efficient operation of AutoMated and human drivEN vehicles in mixed traffic*". In the following four and a half years, Yongqi presented his research findings at numerous international and national conferences, as well as through high-quality journal publications. He supervised seven Master's theses, two Bachelor's theses, and several graduate and undergraduate group research projects.

From May 2023 to October 2023, Yongqi conducted a research visit at the Mechanical Systems Control Lab at the University of California, Berkeley, under the supervision of Prof. Masayoshi Tomizuka.

Throughout his Ph.D. studies, Yongqi secured several research grants and travel fellowships. He took the initiative to organise three international workshops, and he established the interdisciplinary research and technical community of "Automated Mobility in Mixed Traffic".

Yongqi's academic homepage is available at https://yongqidong.github.io/.

List of publications

Journal publications

- **Dong, Y.**, Zhang, L., Farah, H., Zgonnikov, A., & Van Arem, B. (2025). Data-Driven Semi-supervised Machine Learning with Safety Indicators for Abnormal Driving Behavior Detection. Transportation Research Record: Journal of the Transportation Research Board, 1-16. https://doi.org/10.1177/03611981241306752

- **Dong, Y.**, Farah, H., & Van Arem, B. (2025). Towards Developing Socially Compliant Automated Vehicles: Advances, Expert Insights, and A Conceptual Framework. Accepted by Communications in Transportation Research (currently under publication process). Preprint available at https://doi.org/10.48550/arXiv.2501.06089

- **Dong, Y.**, Lu, X., Li, R., Song, W., Van Arem, B., & Farah, H. (2025). Intelligent Anomaly Detection for Lane Rendering Using Transformer with Self-Supervised Pre-Training and Customized Fine-Tuning. Transportation Research Record: Journal of the Transportation Research Board. https://doi.org/10.1177/03611981251333341

- Berge, S. H., De Winter, J., Dodou, D., Afghari, A. P., Papadimitriou, E., Reddy, N., **Dong, Y.**, Raju, N., & Farah, H. (2025). Understanding Cyclists' Perception of Driverless Vehicles through Eye-Tracking and Interviews. Transportation Research Part F: Traffic Psychology and Behaviour, 109, 399-420. https://doi.org/10.1016/j.trf.2024.11.015

- Liu, W., Song, L., **Dong, Y.**, Zhang, X., & Xu, L. (2025). Unified Model Predictive Control Method of Automated Vehicles for Lane Changing and Lane Keeping Maneuvers. Journal of Intelligent Transportation Systems, 1-21. https://doi.org/10.1080/15472450.2025.2479235

- Lingam, S. N., De Winter, J., **Dong, Y.**, Tsapi, A., Van Arem, B., & Farah, H. (2024). eHMI on the Vehicle or on the Infrastructure? A Driving Simulator Study. European Journal of

Transport and Infrastructure Research, 24(2), 1–24. https://doi.org/10.59490/ejtir.2024.24.2.7273

- **Dong, Y.**, Patil, S., Van Arem, B., & Farah, H. (2023). A Hybrid Spatial-temporal Deep Learning Architecture for Lane Detection. Computer-Aided Civil and Infrastructure Engineering, 38(1), 67-86. https://doi.org/10.1111/mice.12829
- Li, R.[#], & **Dong, Y.**[#,*] (2023). Robust Lane Detection Through Self Pre-Training With Masked Sequential Autoencoders and Fine-Tuning With Customized PolyLoss. IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 12, pp. 14121-14132. https://doi.org/10.1109/TITS.2023.3305015
- Farah, H., Postigo, I., Reddy, N., **Dong, Y**., Rydergren, C., Raju, N., & Olstam, J. (2022). Modeling Automated Driving in Microscopic Traffic Simulations for Traffic Performance Evaluations: Aspects to Consider and State of the Practice. IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 6, pp. 6558-6574. https://doi.org/10.1109/TITS.2022.3200176
- Raju, N., Schakel, W., Reddy, N., **Dong, Y.**, Farah, H. (2022). Car-Following Properties of a Commercial Adaptive Cruise Control System: A Pilot Field Test. Transportation Research Record: Journal of the Transportation Research Board, 2676(7), 128-143. https://doi.org/10.1177/03611981221077085
- **Dong, Y.**, Wang, S., Li, L., Zhang, Z. (2018). An Empirical Study on Travel Patterns of Internet Based Ride-Sharing. Transportation Research Part C: Emerging Technologies, 86, 1-22. https://doi.org/10.1016/j.trc.2017.10.022

Journal articles under review

- Patil, S.[#], **Dong, Y.**[#,*], Farah, H, & Hellendoorn, J. (2025). Efficient Sequential Neural Network Based on Spatial-Temporal Attention and Linear LSTM for Robust Lane Detection Using Multi-Frame Images (Under Review). Preprint available at https://doi.org/10.36227/techrxiv.174195585.50092304/v1
- Zhang, Y., **Dong, Y.**[*] (2025). Optimization of Coordinated Flow Restriction and Skip-Stopping Schemes for Urban Rail Stations Considering Platform Carrying Capacity (Under review). Preprint available at https://doi.org/10.36227/techrxiv.21779894.v1

Conference proceedings and presentations

- **Dong, Y.**, Lu, X., Li, R., Song, W., Van Arem, B., & Farah, H. (2024). Intelligent Anomaly Detection for Lane Rendering Using Transformer with Self-Supervised Pre-Training and Customized Fine-Tuning. Presented at the 2024 Transportation Research Board (TRB) 103rd annual meeting. Poster available at http://resolver.tudelft.nl/uuid:00b74fa5-8fef-4514-a21f-877159d58c88
- Huang, Y., **Dong, Y.** [*], Tang, Y. [*], & Li, L. (2024). Leverage Multi-source Traffic Demand Data Fusion with Transformer Model for Urban Parking Prediction. Presented at the 2024

29th International Conference of Hong Kong Society for Transportation Studies (HKSTS). Preprint available at https://doi.org/10.48550/arXiv.2405.01055

- **Dong, Y.**, Liu, C., Wang, Y., & Fu, Z. (2024). Towards Understanding Worldwide Cross-cultural Differences in Implicit Driving Cues: Review, Comparative Analysis, and Research Roadmap. 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), Edmonton, AB, Canada, 2024, pp. 1569-1575. http://dx.doi.org/10.1109/ITSC58415.2024.10919561

- **Dong, Y.**, Detema, T., Wassenaar, V., Van de Weg, J., Kopar, T., & Suleman, H. (2023). Comprehensive Comparison of Deep Reinforcement Learning for Automated Driving on Various Driving Maneuvers with Simulation. 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 2023, pp. 6165-6170. http://dx.doi.org/10.1109/ITSC57777.2023.10422159

- Yuan, H., Li, P., Van Arem, B., Kang, L., Farah, H., & **Dong, Y.**[*] (2023). Safe, Efficient, Comfort, and Energy-saving Automated Driving through Roundabout Based on Deep Reinforcement Learning. 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 2023, pp. 6074-6079. http://dx.doi.org/10.1109/ITSC57777.2023.10422488

- Zhang, L.[#], **Dong, Y.** [#,*]. Farah, H., & Van Arem, B. (2023). Social-aware Planning and Control for Automated Vehicles based on Driving Risk Field and Model Predictive Contouring Control: Driving through Roundabouts as a Case Study. 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Honolulu, Oahu, HI, USA, 2023, pp. 3297-3304. http://dx.doi.org/10.1109/SMC53992.2023.10394462

- **Dong, Y.**[#,*], Patil, S.[#], Farah, H, & Hellendoorn, J. (2023). Sequential Neural Network Model with Spatial-Temporal Attention Mechanism for Robust Lane Detection Using Multi Continuous Image Frames. Presented at the 2023 Transportation Research Board (TRB) 102nd annual meeting. Poster available at http://resolver.tudelft.nl/uuid:01d3bb14-9793-447c-962b-49a70c2b0883

- **Dong, Y.**[#,*], Li, R.[#], Farah, H. (2023). Robust Lane Detection through Self Pre-training with Masked Sequential Autoencoders and Fine-tuning with Customized PolyLoss. Presented at the 2023 Transportation Research Board (TRB) 102nd annual meeting. Poster available at http://resolver.tudelft.nl/uuid:62690e30-572d-44c2-aa8f-f0b1cb835f29

- Zhang, L.[#], **Dong, Y.**[#,*], Farah, H., Zgonnikov, A., & Van Arem, B. (2023). Data-driven Semi-supervised Machine Learning with Surrogate Safety Measures for Abnormal Driving Behavior Detection. Presented at the 35th annual meeting of International Co-operation on Theories and Concepts in Traffic Safety. Poster available at http://resolver.tudelft.nl/uuid:a8e31a16-6609-4c30-8d44-2b4052f0ec42

- Xue, C.[#], **Dong, Y.**[#], Liu, J.[*] , Liao, Y., & Li, L. (2023). Design of the Reverse Logistics System for Medical Waste Recycling Part I: System Architecture and Disposal Site Selection Algorithm. 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 2023, pp. 1741-1746. http://dx.doi.org/10.1109/ITSC57777.2023.10422624

- Xue, C.[#], **Dong, Y.**[#], Liu, J.[*] , Liao, Y., & Li, L. (2023). Design of the Reverse Logistics System for Medical Waste Recycling Part II: Route Optimization with Case Study under COVID-19 Pandemic. 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 2023, pp. 4011-4017. http://dx.doi.org/10.1109/ITSC57777.2023.10422236

- Zhang, Y., **Dong, Y.**[*] (2023). Optimization of Coordinated Flow Restriction and Skip-Stopping Schemes for Urban Rail Stations Considering Platform Carrying Capacity. Presented at the 2023 Transportation Research Board (TRB) 102nd annual meeting. Poster available at http://resolver.tudelft.nl/uuid:fd0562a1-b5ba-4342-a9b7-587308b139c5

- Berge, S. H., De Winter, J., Dimitra, D., Afghari, A. P., Papadimitriou, E., Reddy, N., **Dong, Y.**, Raju, N., & Farah, H. (2023). Cyclists' Gaze Patterns and Driver Detection when Encountering Manual and Driverless Vehicles: A Field Study. Presentation at the 11th International Cycling Safety Conference 2023 (pp. 150-153). https://swov.nl/sites/default/files/bestanden/downloads/ICSC2023_Book_of_abstracts.pdf

- **Dong, Y.**, Chen, K., Peng, Y., & Ma, Z. (2022). Comparative Study on Supervised versus Semi-supervised Machine Learning for Anomaly Detection of In-vehicle CAN Network. 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), 2022, pp. 2914-2919. https://doi.org/10.1109/ITSC55140.2022.9922235

- **Dong, Y.**, Farah, H., & Van Arem, B. (2022). Towards Developing Socially-Compliant Automated Vehicles: State of the Practice, Experts Expectations, and a Conceptual Framework. Presented at the 4th Symposium on Management of Future Motorway and Urban Traffic Systems 2022 (MFTS2022). Preprint available at https://doi.org/10.48550/arXiv.2501.06089

- **Dong, Y.**, Yang, Z., Yue, Y., Pei, X., & Zhang, Z. (2018). Revealing Travel Patterns of Sharing-bikes in a Spatial-temporal Manner Using Non-negative Matrix Factorization Method. In CICTP 2018: Intelligence, Connectivity, and Mobility (pp. 1665-1674). Reston, VA: American Society of Civil Engineers. https://doi.org/10.1061/9780784481523.165

- Yue, Y., Pei, X., Yang, Z., **Dong, Y.**, & Yao, D. (2018). A Trip Building and Chaining Methodology Using Traffic Surveillance Data. In CICTP 2018: Intelligence, Connectivity, and Mobility (pp. 2254-2262). Reston, VA: American Society of Civil Engineers. https://doi.org/10.1061/9780784481523.224

- **Dong, Y.**, Zhang, Z., Fu, R., Xie, N. (2016). Revealing New York Taxi Drivers' Operation Patterns Focusing on the Revenue Aspect. In 12th World Congress on Intelligent Control and Automation (WCICA), (pp. 1052-1057). IEEE. https://doi.org/10.1109/WCICA.2016.7578771

Conference papers accepted (pending presentation)

- Ji, J., Lu, R., Belkessa, L., **Dong, Y.**, Wang, L., Madadi, B., Varotto, S., Saunier, N., MacFarlane, G., & Wu, C., (2025). Exploring Artifacts Availability in Transportation Research Using Large Language Models. Accepted by the 2025 Transportation Research Symposium (TRS), the 2025 International Symposium on Transportation Data &

Modelling (ISTDM), and the 2025 Modelling Mobility Conference (MoMo) for presentation.

Patents

Dutch patents

- **Dong, Y.**, & Li, R. (2024). Automated Lane Detection (Dutch Patent No. NL2033551). Netherlands: Netherlands Patent Office.
- **Dong, Y.**, Zhang, L., Farah, H., & Van Arem, B. (2025). Socially-compliant Automated Driving in Mixed Traffic (Dutch Patent No. NL2035943, submitted & filed, OCT-23-056).

Chinese invention patent

- Ruan, H., **Dong, Y.**, Wang, W., & Wang, F. (2016). Intelligent Demonstration Instrument of Simple Harmonic Oscillation Composition and Five Polarization States of Light (Chinese Patent No. CN103236211B). China: China National Intellectual Property Administration.

Software copyright

- Spatial-Temporal Attention Integrated Sequential Neural Network for Vision-based Lane Detection (i-DEPOT 142731, submitted & filed).

Open resource repository

- Datasets, Simulation Platforms, and Relevant Publications on Emerging Mixed Traffic of Automated Vehicles and Human-driven Vehicles, Online available at https://qiqiqi.gitbook.io/mixed-traffic

*The superscript # indicates equal contribution, and * indicates the corresponding author.*

List of codes and datasets

- **Dong, Y.**, van Arem, B., & Haneen Farah. (2025). *Code and Data Underlying the Publication: Towards Developing Socially Compliant Automated Vehicles: Advances, Expert Insights, and A Conceptual Framework* [Data set]. 4TU.ResearchData. https://doi.org/10.4121/3A46E61C-F5F0-4399-A4B8-4D146B62A4F7
- **Dong, Y.**, Patil, S., Van Arem, B., & Farah, H. (2025). *Code Underlying the Publication: A Hybrid Spatial-temporal Deep Learning Architecture for Lane Detection.* [Data set]. 4TU.ResearchData. https://doi.org/10.4121/ba5805cb-a909-4185-97b0-296739df7def
- **Dong, Y.**, Li, R., & Farah, H. (2025). *Code Underlying the Publication: Robust Lane Detection Through Self Pre-Training with Masked Sequential Autoencoders and Fine-Tuning With Customized PolyLoss*. [Data set]. 4TU.ResearchData. https://doi.org/10.4121/08277f5d-c904-4274-992e-085b3edeb19f
- **Dong, Y.**, Patil, S., Farah, H, & Hellendoorn, J. (2025). *Code Underlying the Publication: Efficient Sequential Neural Network Based on Spatial-Temporal Attention and Linear LSTM for Robust Lane Detection Using Multi-Frame Images*. [Data set]. 4TU.ResearchData. [Data set]. https://doi.org/10.4121/4619cab6-ae4a-40d5-af77-582a77f3d821
- **Dong, Y.**, Zhang, L., Farah, H., & Van Arem, B. (2025). *Code and Data Underlying the Publication: Social-aware Planning and Control for Automated Vehicles Based on Driving Risk Field and Model Predictive Contouring Control: Driving through Roundabouts as a Case Study.* [Data set]. 4TU.ResearchData. https://doi.org/10.4121/70e29cf5-8502-4e8d-bf32-2953431a83ff
- **Dong, Y.**, Zhang, L., Farah, H., Zgonnikov, A., & Van Arem, B. (2025). *Code and Data Underlying the Publication: Data-driven Semi-supervised Machine Learning with Safety*

Indicators for Abnormal Driving Behavior Detection. [Data set]. 4TU.ResearchData. https://doi.org/10.4121/b60dfda0-055a-4046-a615-e0166a356c95

- Yuan, H., **Dong, Y.**, Li, P., Van Arem, B., Kang, L., & Farah, H. (2025). *Code Underlying the Publication: Safe, Efficient, Comfort, and Energy-Saving Automated Driving Through Roundabout Based on Deep Reinforcement Learning.* [Data set]. 4TU.ResearchData. https://doi.org/10.4121/c1020a3f-0053-491f-8ead-35d18819d37e

- **Dong, Y.**, Datema, T., Van de Weg, J., Kopar, C.T., & Suleman, H. (2025). *Code Underlying the Publication: Comprehensive Training and Evaluation on Deep Reinforcement Learning for Automated Driving in Various Simulated Driving Maneuvers.* [Data set]. 4TU.ResearchData. https://doi.org/10.4121/26e8f131-53f8-44b9-8ecf-249bfedb0154

- Berge, S. H., De Winter, J., Dodou, D., Afghari, A. P., Papadimitriou, E., Reddy, N., **Dong, Y.**, Raju, N., & Farah, H. (2024): *Supplementary Data for the Paper 'Understanding Cyclists' Perception of Driverless Vehicles through Eye-Tracking and Interviews'*. [Data set]. 4TU.ResearchData. https://doi.org/10.4121/ee1abac8-bfc9-4a1b-9a45-29f1bc461eb9

- Lingam, S. N., De Winter, J., **Dong, Y.**, Tsapi, A., Van Arem, B., & Farah, H. (2024). *Supplementary Data for the Paper 'eHMI on the Vehicle or on the Infrastructure? A Driving Simulator Study'*. [Data set]. 4TU.ResearchData. https://doi.org/10.4121/8e1c6604-53c7-4905-8ef9-4a6b2acc4e7a

Acknowledgements

Funding

All of the research outlined in this thesis is generously funded by the Applied and Technical Sciences (TTW), a subdomain of the Dutch Institute for Scientific Research (NWO) through the Project *Safe and Efficient Operation of Automated and Human-Driven Vehicles in Mixed Traffic* (SAMEN) under Contract 17187.

The author also acknowledges the generous financial support provided by the Transport & Mobility Institute at Delft University of Technology (TU Delft) for supporting interdisciplinary research and a five-month international research visit at the University of California, Berkeley. Additionally, the author expresses gratitude for the Erasmus+ mobility Grants, as well as the young professional fellowships and initial funding from the IEEE Intelligent Transportation Systems Society (ITSS) for attending international training, conferences, and workshops.

High-performance computing platforms

Portions of the research conducted in this thesis were facilitated by the advanced computing resources provided by *Snellius: the National Supercomputer* and *DelftBlue: the TU Delft supercomputer*. The availability of high-performance computing infrastructure significantly contributed to the efficiency of the computational tasks involved in this thesis.

Digital resources

Various digital resources have been instrumental in the creation of this thesis, including an AI-driven generative conversational agent developed on the GPT-3.5 platform. This agent has served as a dynamic tool, facilitating an iterative process of summarising written content, proofreading and polishing English writing, as well as translating English summaries into Dutch. It is essential to clarify that while this conversational agent assisted in generating text in response to specific queries and prompts designed by the author, all final interpretations, articulations, findings, and conclusions are solely the author's own.

TRAIL Thesis Series

The following list contains the most recent dissertations in the TRAIL Thesis Series. For a complete overview of more than 400 titles, see the TRAIL website: www.rsTRAIL.nl.

The TRAIL Thesis Series is a series of the Netherlands TRAIL Research School on transport, infrastructure and logistics.

Dong, Y., Safe, *Efficient, and Socially Compliant Automated Driving in Mixed Traffic: Sensing, Anomaly Detection, Planning and Control,* T2025/6, May 2025, TRAIL Thesis Series, the Netherlands

Droffelaar, I.S. van, *Simulation-optimization for Fugitive Interception,* T2025/5, May 2025, TRAIL Thesis Series, the Netherlands

Fan, Q., *Fleet Management Optimisation for Ride-hailing Services: from mixed traffic to fully automated environments,* T2025/4, April 2025, TRAIL Thesis Series, the Netherlands

Hagen, L. van der, *Machine Learning for Time Slot Management in Grocery Delivery,* T2025/3, March 2025, TRAIL Thesis Series, the Netherlands

Schilt, I.M. van, *Reconstructing Illicit Supply Chains with Sparse Data: a simulation approach,* T2025/2, January 2025, TRAIL Thesis Series, the Netherlands

Ruijter, A.J.F. de, *Two-Sided Dynamics in Ridesourcing Markets,* T2025/1, January 2025, TRAIL Thesis Series, the Netherlands

Fang, P., *Development of an Effective Modelling Method for the Local Mechanical Analysis of Submarine Power Cables,* T2024/17, December 2024, TRAIL Thesis Series, the Netherlands

Zattoni Scroccaro, P., *Inverse Optimization Theory and Applications to Routing Problems,* T2024/16, October 2024, TRAIL Thesis Series, the Netherlands

Kapousizis, G., *Smart Connected Bicycles: User acceptance and experience, willingness to pay and road safety implications,* T2024/15, November 2024, TRAIL Thesis Series, the Netherlands

Lyu, X., *Collaboration for Resilient and Decarbonized Maritime and Port Operations,* T2024/14, November 2024, TRAIL Thesis Series, the Netherlands

Nicolet, A.*, Choice-Driven Methods for Decision-Making in Intermodal Transport: Behavioral heterogeneity and supply-demand interactions,* T2024/13, November 2024, TRAIL Thesis Series, the Netherlands

Kougiatsos, N., *Safe and Resilient Control for Marine Power and Propulsion Plants,* T2024/12, November 2024, TRAIL Thesis Series, the Netherlands

Uijtdewilligen, T., *Road Safey of Cyclists in Dutch Cities,* T2024/11, November 2024, TRAIL Thesis Series, the Netherlands

Liu, X., *Distributed and Learning-based Model Predictive Control for Urban Rail Transit Networks*, T2024/10, October 2024, TRAIL Thesis Series, the Netherlands

Clercq, G. K. de, *On the Mobility Effects of Future Transport Modes,* T2024/9, October 2024, TRAIL Thesis Series, the Netherlands

Dreischerf, A.J., *From Caveats to Catalyst: Accelerating urban freight transport sustainability through public initiatives,* T2024/8, September 2024, TRAIL Thesis Series, the Netherlands

Zohoori, B., *Model-based Risk Analysis of Supply Chains for Supporting Resilience,* T2024/7, October 2024, TRAIL Thesis Series, the Netherlands

Poelman, M.C., *Predictive Traffic Signal Control under Uncertainty: Analyzing and Reducing the Impact of Prediction Errors,* T2024/6, October 2024, TRAIL Thesis Series, the Netherlands

Berge, S.H.*, Cycling in the Age of Automation: Enhancing cyclist interaction with automated vehicles through human-machine interfaces,* T2024/5, September 2024, TRAIL Thesis Series, the Netherlands

Wu, K., *Decision-Making and Coordination in Green Supply Chains with Asymmetric Information,* T2024/4, July 2024, TRAIL Thesis Series, the Netherlands

Wijnen, W., *Road Safety and Welfare,* T2024/3, May 2024, TRAIL Thesis Series, the Netherlands

Caiati, V., *Understanding and Modelling Individual Preferences for Mobility as a Service,* T2024/2, March 2024, TRAIL Thesis Series, the Netherlands

Vos, J., *Drivers' Behaviour on Freeway Curve Approach,* T2024/1, February 2024, TRAIL Thesis Series, the Netherlands

Geržinič, N., *The Impact of Public Transport Disruptors on Travel Behaviour,* T2023/20, December 2023, TRAIL Thesis Series, the Netherlands

Dubey, S., *A Flexible Behavioral Framework to Model Mobility on-Demand Service Choice Preference,* T2023/19, November 2023, TRAIL Thesis Series, the Netherlands

Sharma, S., *On-trip Behavior of Truck Drivers on Freeways: New mathematical models and control methods,* T2023/18, October 2023, TRAIL Thesis Series, the Netherlands

Ashkrof, P., *Supply-side Behavioural Dynamics and Operations of Ride-sourcing Platforms,* T2023/17, October 2023, TRAIL Thesis Series, the Netherlands

Sun, D., *Multi-level and Learning-based Model Predictive Control for Traffic Management,* T2023/16, October 2023, TRAIL Thesis Series, the Netherlands

TRAIL THESIS SERIES T2025/6

Yongqi Dong

Safe, Efficient, and Socially Compliant Automated Driving in Mixed Traffic

Summary

As automated vehicles (AVs) gradually integrate into mixed traffic with human-driven vehicles, this thesis addresses critical challenges during the transition era. It enhances AV capabilities in sensing and perception, anomaly detection, as well as planning and control. Employing spatial-temporal deep learning models, self-supervised pretraining methods with masked sequential autoencoders, and innovative social-aware decision-making strategies, this thesis aims to facilitate safe, efficient, and socially compliant automated driving, thereby advancing future transportation systems.

About the Author

Yongqi Dong is a researcher specialising in automated driving systems and artificial intelligence. He conducted his PhD research at TU Delft, focusing on enhancing automated vehicles' capabilities in mixed-traffic environments. He holds degrees in Control Science and Engineering and Telecommunication Engineering.

TRAIL Research School ISBN 978-90-5584-361-9

Safe, Efficient, and Socially Compliant Automated Driving in Mixed Traffic

Sensing, Anomaly Detection, Planning and Control

Yongqi Dong