# Fusion of Gaze and Scene Information for Driving Behaviour Recognition
## A Graph-Neural-Network- Based Framework

Yi, Yangtian; Lu, Chao; Wang, Boyang; Cheng, Long; Li, Zirui; Gong, Jianwei

**DOI**
10.1109/TITS.2023.3263875

**Publication date**
2023

**Document Version**
Final published version

**Published in**
IEEE Transactions on Intelligent Transportation Systems

# Fusion of Gaze and Scene Information for Driving Behaviour Recognition: A Graph-Neural-Network-Based Framework

Yangtian Yi, Chao Lu, Boyang Wang, Long Cheng, Zirui Li, and Jianwei Gong, *Member, IEEE*

***Abstract*— Accurate recognition of driver behaviours is the basis for a reliable driver assistance system. This paper proposes a novel fusion framework for driver behaviour recognition that utilises the traffic scene and driver gaze information. The proposed framework is based on the graph neural network (GNN) and contains three modules, namely, the gaze analysing (GA) module, scene understanding (SU) module and the information fusion (IF) module. The GA module is used to obtain gaze images of drivers, and extract the gaze features from the images. The SU module provides trajectory predictions for surrounding vehicles, motorcycles, bicycles and other traffic participants. The GA and SU modules are parallel and the outputs of both modules are sent to the IF module that fuses the gaze and scene information using the attention mechanism and recognises the driving behaviours through a combined classifier. The proposed framework is verified on a naturalistic driving dataset. The comparative experiments with the state-of-the-art methods demonstrate that the proposed framework has superior performance for driving behaviour recognition in various situations.**

***Index Terms*— Driving behaviours, graph neural network, gaze information, scene information, data fusion.**

## I. Introduction

ACCORDING to the forecast of the International Data Corporation (IDC), the global annual shipment of intelligent vehicles equipped with advanced driving assistant systems (ADASs) will reach about 76.2 million by 2024. With the rapid development of intelligent vehicles, the reliability of ADASs, such as the brake assist system (BAS), lane keeping system (LKS) and adaptive cruise control (ACC), has been one important concern of drivers and other road users.

Manuscript received 16 March 2022; revised 13 October 2022, 9 January 2023, and 4 March 2023; accepted 28 March 2023. Date of publication 12 April 2023; date of current version 2 August 2023. This work was supported in part by the Key Research and Development Program of Shandong Province, China, under Grant 2020CXGC010118; and in part by the National Natural Science Foundation of China under Grant 61703041 and Grant U19A2083. The Associate Editor for this article was L. M. Bergasa. *(Corresponding authors: Chao Lu; Boyang Wang.)*

Yangtian Yi, Chao Lu, Boyang Wang, and Jianwei Gong are with the School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China (e-mail: 3120200422@bit.edu.cn; chaolu@bit.edu.cn; boyang_wang@pku.edu.cn; gongjianwei@bit.edu.cn).

Long Cheng is with the Jiangsu Key Laboratory of Urban ITS, School of Transportation, Southeast University, Nanjing 211189, China (e-mail: longcheng@seu.edu.cn).

Zirui Li is with the School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China, and also with the Department of Transport and Planning, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: z.li@bit.edu.cn).

Digital Object Identifier 10.1109/TITS.2023.3263875

Accurate recognition of driver behaviours is the basis for a reliable ADAS. However, in urban areas, the recognition of driver behaviours is not an easy task due to the complexity of the traffic environment. The behaviours of drivers are affected by various dynamic factors, such as surrounding vehicles, bicycles and pedestrians, and static factors, such as road geometry, traffic signs and facilities, making it a challenge for intelligent vehicles to work reliably. Therefore, numerous studies have been conducted to improve the recognition accuracy of driving behaviours in various scenes.

For instance, in car-following scenes, a consolidated fuzzy clustering algorithm was proposed in [1] to classify different car-following patterns. In a similar work found in [2], the recognition results of driving patterns were based on vehicle speed, acceleration, range and other factors. The braking action in the car-following scene was regarded as a driving behaviour in [3], which was recognized by a model combining the Gaussian mixture model (GMM) and hidden Markov model (HMM).

Except for the longitudinal behaviour of drivers, such as breaking and accelerations, another group of studies focused on the lateral behaviour of drivers in lane-changing or turning scenes. The lane-changing behaviour of the ego vehicle was recognized in some studies based on the driver operation data and inertial measurement unit data. In [4], driving behaviour was divided into lane-keeping and lane-changing, while in [5], lane-changing behaviour is divided into 3 types: cautious, normal and aggressive. LiDAR can be used to obtain the position of surrounding vehicles in the scene. Using the scene information from LiDAR, the lane-changing and turning behaviour of surrounding vehicles can be recognised [6], [7], [8]. When the realistic data was insufficient, the data collected in driving simulator could be used to recognise driving behaviours based on transfer learning [9], [10].

In the traffic scene, the longitudinal and lateral behaviours of vehicles are not isolated from each other. Therefore, some studies considered the two together. Overtaking is a traffic scene combining longitudinal and lateral behaviours. A combined learning framework based on the natural actor-critic learning and general regression neural network was developed in [11], to learn the driver-specific behaviour for overtaking. However, in some studies, the classification of driving behaviours was not clear enough. For instance, all lateral behaviours were classified as turning in [12], and only

acceleration and normal driving were considered in longitudinal behaviour in [13].

Although the above-mentioned studies have improved the accuracy and efficiency in the task of recognising driving behaviours, they are mainly concerned about the impact of scene information, such as the velocity of the ego vehicle and the trajectories of surrounding traffic participants. The gaze information was ignored by the above-mentioned studies. The literature shows that gaze information is highly related to human behaviours and in many situations, even guides human actions [14]. Consequently, several studies have paid attention to gaze information and tried to correlate this kind of information with driving behaviour. In [15], the eye movement and gaze information were captured by eye tracking glasses, which are portable and do not affect the normal driving of the drivers. Similarly, methods using camera images and semantic segmentation images to predict gaze images of the driver have been proposed in [16] and [17]. To explore the relationship between gaze and driving behaviour, the gaze of the driver at the traffic signal was analysed in [18], to predict the behaviour of the driver 3-4 seconds before reaching the intersection. Martin and Trivedi judged the attention area of the driver by the camera arranged in the vehicle, and studied lane-changing driver behaviour [19]. Further, they proposed a machine-vision-based framework for predicting the behaviours of drivers [20]. In addition to recognizing driving behaviour at intersections, in [21], the driver gaze location was also used to predict the braking behaviour and estimate the driver's steering angle in various urban traffic scenes.

However, the above-mentioned studies containing driver gaze information did not consider scene information (dynamic and static factors). Hence, there is a need for a framework that can combine traffic scene and driver gaze information for driving behaviour recognition. Therefore, this paper proposes a framework based on a graph neural network (GNN) for fusing traffic scene and driver gaze information to recognize the driving behaviours in different situations. The proposed framework contains a gaze analysing (GA) module, a scene understanding (SU) module and an information fusion (IF) module. The GA module extracts the gaze features of the driver by processing the camera, optical flow and gaze images in time series. The SU module, a GNN, is responsible for capturing the trajectories of traffic participants as input and comprehending the traffic scene. (In this paper, the GNN is fed with trajectories of traffic participants, and outputs the vectors containing scene features. This process is called scene comprehending.) The IF module adopts the attention mechanism to fuse gaze and scene features in the proposed framework, to recognize up to 8 kinds of driving behaviours.

The rest of this paper is arranged as follows. Section II describes the GNN-based framework and defines the problem. The methodology is presented in Section III. Experimental settings and comparative results are presented in Section IV. Finally, Section V concludes the paper and provides ideas for future work.

## II. Description of the Graph-Neural-Network-Based Framework

This section describes the proposed framework. Firstly, the problem is defined and the inputs and outputs of the proposed framework are determined. Then, three main modules of the proposed framework are introduced.

### A. Problem Definition

The proposed framework can recognise the behaviour of drivers in urban traffic. Assuming that the images from an onboard camera with the corresponding driver gaze images and spatial coordinates of surrounding traffic participants are available. At each time $t$, the input of the model is defined as:

$$\mathbf{F}^t = \{\mathbf{I}^t, \mathbf{S}^t, \mathbf{O}^t, \mathbf{C}^t\} \tag{1}$$

where $\mathbf{I}^t$, $\mathbf{S}^t$ and $\mathbf{O}^t$ are the pixel matrices for the original image collected from the on-board camera (camera image), the processed image with gaze information (gaze image), and the optical flow image at time $t$, respectively. $\mathbf{C}^t$ is an aggregation of characteristics of surrounding traffic participants, $\mathbf{C}^t = \{\mathbf{f}_1^t, \mathbf{f}_2^t, \ldots, \mathbf{f}_i^t\}$. $\mathbf{f}_i^t$ is the feature vector of the $i$th surrounding traffic participant at time $t$, $\mathbf{f}_i^t = \{x_i^t, y_i^t, c_i\}$, in which $\{x_i^t, y_i^t\}$ are the relative position of traffic participants and ego vehicle in the vehicle body coordinate system. X and Y represent lateral and longitudinal directions respectively, and $c_i$ is the category label, $c_i \in \{1, 2, 3\}$, where 1, 2 and 3 stand for pedestrians, vehicles and riders, respectively. Since the number of traffic participants around the ego vehicle changes with time, the number of elements in $\mathbf{C}^t$ is time-varying. Taking information $[\mathbf{F}^1, \mathbf{F}^2, \ldots, \mathbf{F}^{T_{\text{obs}}}]$ of time interval $[1 : T_{\text{obs}}]$ as input, the proposed framework can recognize 8 kinds of driving behaviours during this period of time.

### B. Framework Structure

There are three main modules in the proposed framework. Two parallel modules are the GA and SU modules. The third is the IF module. Fig.1 shows the overall structure of the proposed framework. The matrix sequences of camera images $[\mathbf{I}^1, \mathbf{I}^2, \ldots, \mathbf{I}^{T_{\text{obs}}}]$, optical flow images $[\mathbf{O}^1, \mathbf{O}^2, \ldots, \mathbf{O}^{T_{\text{obs}}}]$ and gaze images $[\mathbf{S}^1, \mathbf{S}^2, \ldots, \mathbf{S}^{T_{\text{obs}}}]$ of time interval $[1{:}T_{\text{obs}}]$ are sent to the GA module. The sequence of features $[\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^{T_{\text{obs}}}]$ of surrounding traffic participants is sent to the SU module. Then, the IF module combines the outputs of the GA and SU modules through the attention mechanism. The output of the IF module is connected with a classifier (such as MLP or SVM) to classify the driving behaviours into 8 categories (shown in Fig.5).

In an urban traffic environment, human drivers cannot observe all the details of the environment at any time. The gaze of the driver is always focused on the local area. For example, when the driver is ready to overtake, he or she will pay special attention to the traffic on the overtaking side; Another example is that when the driver is following a vehicle, he or she will look at the vehicle in front and adjust the velocity

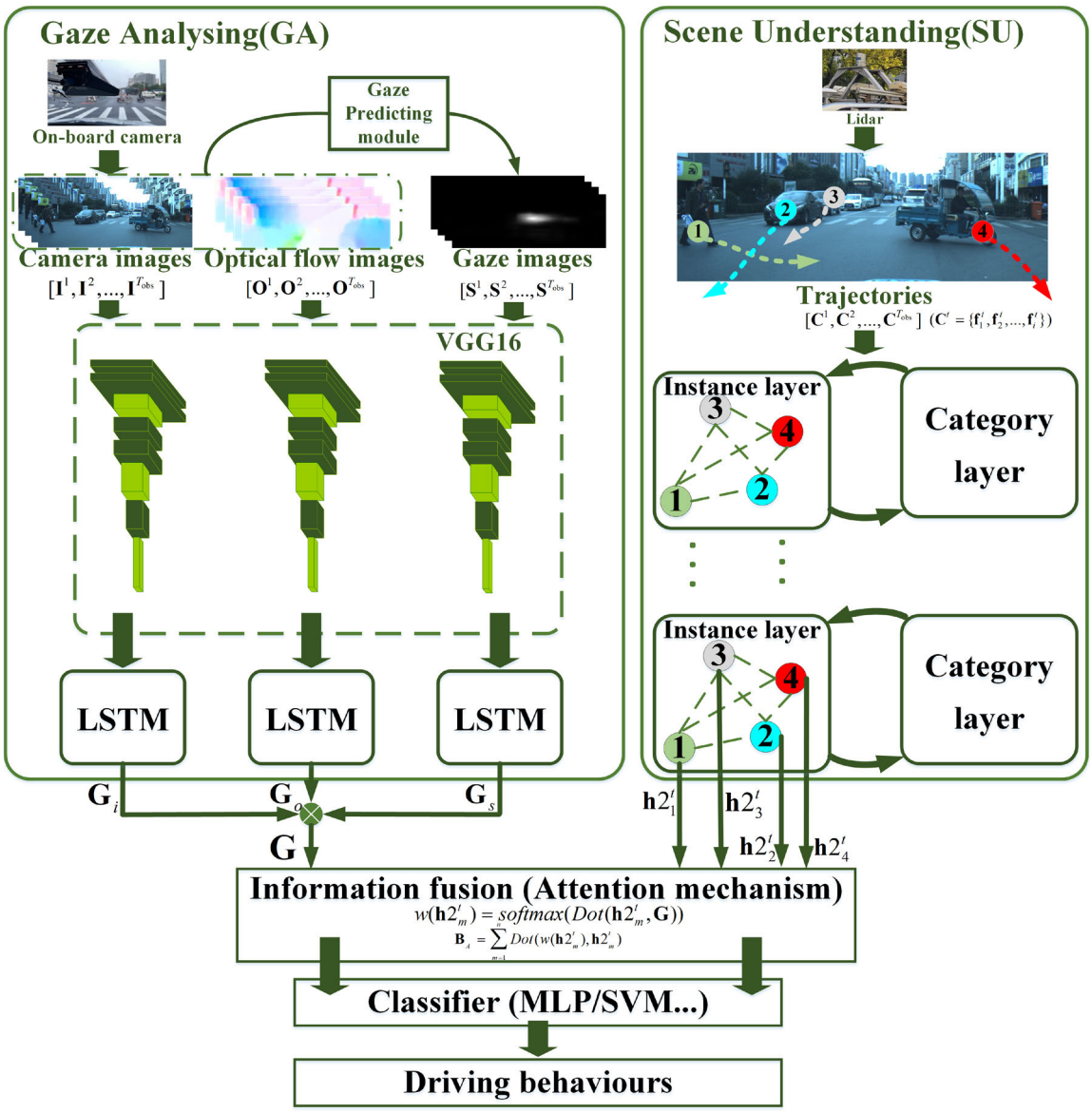


Fig. 1. The proposed framework.

according to the behaviour of the vehicle in front. Therefore, the gaze information of the driver can reflect the driving intention to a certain extent, which should be considered in the recognition of driving behaviour. The gaze images in Fig. 1 are obtained by a gaze predicting method shown in Fig. 2, which is simplified from [16]. The gaze predicting method has two identical branches each containing an encoder-decoder composed of convolutional neural networks (CNNs) in series. One branch takes the camera images as input, and the other takes the optical flow images as input. An optical flow image can represent the instantaneous velocity of the pixel motion of a moving object on the observation imaging plane. The corresponding relationship between the previous frame and the current frame is determined using the displacement of pixels in the time domain, calculating the moving information of objects between adjacent frames [22]. Optical flow images have been applied in some image processing fields [23], [24]. The calculation method of optical flow image applied by this work is described in Section III. The change of the size of images during the data processing is shown in Fig 2. The calculation results of two submodules are two matrices with the size of $1 \times 112 \times 112$. They are added and normalized to obtain the final prediction result. The result of gaze predicting is a probability map.

In the field of image processing, many models based on convolutional neural networks (CNN) including visual geometry group (VGG), LeNet and AlexNet have achieved remarkable results [25], [26], [27]. However, these models are composed of 2D convolution layers and do not possess the ability to process images in time series. The long short-term memory (LSTM) has been widely applied in trajectory prediction and behaviour recognition due to its good performance in sequence data processing. Therefore, the advantages of VGG and LSTM are combined in the GA module. The strong feature extraction ability of the VGG is adopted to extract the features of each frame in sequence input, and the LSTM deals with

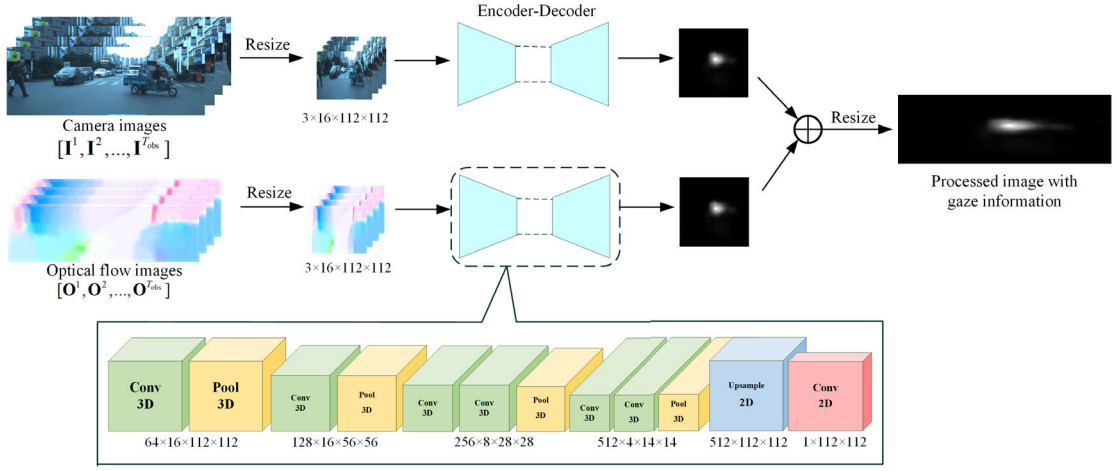


Fig. 2. The flowchart for gaze predicting.

the sequence data after feature extraction. The GA module consists of three parallel submodules as shown in Fig. 1. Each submodule consists of a VGG and an LSTM. Each image is resized to 3 × 224 × 224. Through 5 convolution layers, the image is processed into a vector of 1 × 25088. Then, the vectors in a time series are sent to LSTM to obtain the output result of the corresponding submodule.

Following [28], a GNN is built here to form the basis of the SU module. As mentioned earlier, the number of elements in $\mathbf{C}^t = \{\mathbf{f}_1^t, \mathbf{f}_2^t, \ldots, \mathbf{f}_i^t\}$ is not constant and depends on the number of traffic participants around the ego vehicle. Therefore, the SU module can be described as a GNN. The effectiveness of GNN has been validated in several studies on trajectory prediction and behaviour recognition [28], [29], [30], [31]. The SU module has a hierarchical structure, which is divided into an instance layer and a category layer.

In the instance layer, at any time $t$, each participant is modelled as an instance node $A_i^t$ with a feature defined as $\mathbf{f}_i^t = \{x_i^t, y_i^t, c_i\}$. The edge between $A_i^t$ and its surrounding nodes is called spatial edge $(A_i^t, A_j^t)$, which is bidirectional. Considering $(A_i^t, A_j^t)$, it represents the influence of node $A_j^t$ on $A_i^t$ in space and is characterized by $\mathbf{f}_{ij}^t = (x_{ij}^t, y_{ij}^t, c_{ij}^t)$, where $x_{ij}^t = x_j^t - x_i^t$ and $y_{ij}^t = y_j^t - y_i^t$ are the coordinates of $A_j^t$ relative to $A_i^t$ in the X and Y direction, respectively, and $c_{ij}^t$ is obtained by encoding the types of two nodes. Concerning $(A_j^t, A_i^t)$, the feature is computed as $\mathbf{f}_{ji}^t = (x_{ji}^t, y_{ji}^t, c_{ji}^t)$. Depending on the spatial edge, the SU module can consider not only the location of each traffic participant, but also the spatial relationship between the participants. When the same node $A_i$ appears in two adjacent frames, a temporary edge $(A_i^t, A_i^{t+1})$ is defined between them. In contrast to spatial edge, the temporary edge is unidirectional and represents the transmission of information in the time stream. The feature of the temporal edge $(A_i^t, A_i^{t+1})$ is defined as $\mathbf{f}_{ii}^t = (x_{ii}^t, y_{ii}^t, c_{ii}^t)$, where $x_{ii}^t = x_i^{t+1} - x_i^t$ and $y_{ii}^t = y_i^{t+1} - y_i^t$. $c_{ii}^t$ is the only encoding for the traffic participants of its type.

To extract behaviour characteristics of traffic participants with different types, a category layer is built containing several super nodes $C_c^t (c \in \{1, 2, 3\})$ for different categories of traffic participants at each time step. The information of participants of the same type is gathered to the super node through an edge. The super node extracts the behaviour characteristics of the same kind of traffic participants, and sends the information back to the instance node through the edge to obtain the driver behaviour recognition result $\mathbf{B}_A$.

In this paper, the attention mechanism is used to fuse the output of the GA and SU modules [32]. The gaze images represent the area that the driver pays special attention to. The impact of traffic participants in this area on the driver's behaviour should be different from the impact of participants outside this area. Therefore, different weights should be given to the outputs of the hidden states by all nodes in the SU module. The attention mechanism distributes the different weights to different traffic participants, which represent the impact of different traffic participants on the driving behaviour of ego vehicles. After weighted summation, gaze and scene information is fused.

## III. Methodology

This section describes the methodology and the flow of data in the proposed framework.

### A. Gaze Analysing

According to the assumption of brightness constancy, small motion and spatial coherence [22], the optical flow can be calculated as:

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0 \tag{2}$$

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \tag{3}$$

where $I$ is the grayscale image, $x$ and $y$ are pixel coordinates, $t$ represents the time index, $V_x$ and $V_y$ are the velocities of the pixel (x, y) in the X and Y direction, respectively. Depending on the optical flow image, the fast-moving targets in the driver's field of vision can be analysed, which are often the objects of special concern to the driver. Fig. 3 shows

Fig. 3. Calculated optical flow image (converted to RGB format).

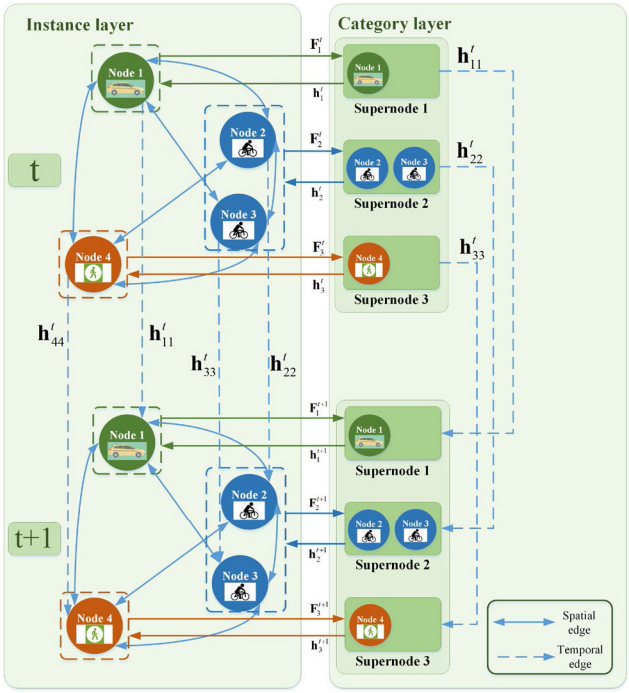


Fig. 4. The hierarchical structure of the SU module.

the optical flow image converted to RGB format using the proposed method.

Therefore, the GA module consists of three parallel modules with matrices of camera images $\mathbf{I}$, optical flow images $\mathbf{O}$ and driver gaze images $\mathbf{S}$ as inputs:

$$\mathbf{R}_i = VGG([\mathbf{I}^1, \mathbf{I}^2, \ldots, \mathbf{I}^{\mathrm{T_{obs}}}]) \tag{4}$$
$$\mathbf{R}_o = VGG([\mathbf{O}^1, \mathbf{O}^2, \ldots, \mathbf{O}^{\mathrm{T_{obs}}}]) \tag{5}$$
$$\mathbf{R}_s = VGG([\mathbf{S}^1, \mathbf{S}^2, \ldots, \mathbf{S}^{\mathrm{T_{obs}}}]) \tag{6}$$
$$\mathbf{G}_i = LSTM(\mathbf{R}_i; \mathbf{W}_i) \tag{7}$$
$$\mathbf{G}_\mathrm{o} = LSTM(\mathbf{R}_o; \mathbf{W}_o) \tag{8}$$
$$\mathbf{G}_s = LSTM(\mathbf{R}_s; \mathbf{W}_s) \tag{9}$$

where $\mathbf{W}_i$, $\mathbf{W}_o$ and $\mathbf{W}_s$ are the weight matrices of LSTM cells. The sizes of the camera and gaze images are pre-processed to 224 * 224 and are sent into the GA module to obtain the results. $\mathbf{G}_i$, $\mathbf{G}_\mathrm{o}$ and $\mathbf{G}_s$ are the vectors that are output by the LSTM. The output $\mathbf{G}$ of the GA module is the average vector of $\mathbf{G}_i$, $\mathbf{G}_\mathrm{o}$ and $\mathbf{G}_s$.

### B. Scene Understanding

The hierarchical structure of the SU module is shown in Fig. 4. The two layers are called the instance layer and the category layer.

*1) Instance Layer:* The instance layer learns the movement of each traffic participant, including the spatial relationship between the participants and the displacement of each traffic participant over time. An LSTM $L_i$ is set for each instance node $A_i$, which is used to collect the information from the edge connected to it. For the spatial and temporal edges $(A_i^t, A_j^t)$ and $(A_i^t, A_i^{t+1})$, LSTMs $L_{ij}$ and $L_{ii}$ are set, respectively. At any time, feature vectors $\mathbf{f}_{ij}^t$ and $\mathbf{f}_{ii}^t$ are embedded, and then fed into $L_{ij}$ and $L_{ii}$:

$$\mathbf{e}_{ij}^t = embed(\mathbf{f}_{ij}^t; \mathbf{W}_{\mathrm{spa}}^e) \tag{10}$$
$$\mathbf{e}_{ii}^t = embed(\mathbf{f}_{ii}^t; \mathbf{W}_{\mathrm{temp}}^e) \tag{11}$$
$$\mathbf{h}_{ij}^t = LSTM(\mathbf{h}_{ij}^{t-1}, \mathbf{e}_{ij}^t; \mathbf{W}_{\mathrm{spa}}^r) \tag{12}$$
$$\mathbf{h}_{ii}^t = LSTM(\mathbf{h}_{ii}^{t-1}, \mathbf{e}_{ii}^t; \mathbf{W}_{\mathrm{temp}}^r) \tag{13}$$

where $embed(\cdot)$ is an embedding function that uses a linear mapping layer to encode the input into a vector of a specific length, $\mathbf{h}_{ij}^t$ and $\mathbf{h}_{ii}^t$ are the output vectors of $L_{ij}$ and $L_{ii}$, respectively and are called hidden states of the corresponding LSTM, $\mathbf{W}_{\mathrm{spa}}^e$ and $\mathbf{W}_{\mathrm{temp}}^e$ are the embedding weight matrices, while $\mathbf{W}_{\mathrm{spa}}^r$ and $\mathbf{W}_{\mathrm{temp}}^r$ are the weight matrices of LSTM cells.

When there are multiple participants in traffic, a node is connected to multiple spatial edges. The GNN assigns weights to the output of each spatial edge using soft attention mechanism, considering their spatial impact on the node:

$$w(\mathbf{h}_{ij}^t) = softmax(\lambda \cdot Dot(\mathbf{W}_{ii}\mathbf{h}_{ii}^t, \mathbf{W}_{ij}\mathbf{h}_{ij}^t)) \tag{14}$$

where $w(\mathbf{h}_{ij}^t)$ is the weight corresponding to the spatial edge output $\mathbf{h}_{ij}^t$, $\mathbf{W}_{ii}$ and $\mathbf{W}_{ij}$ are the embedding weight matrices, $Dot(\cdot)$ is the dot product, and $\lambda$ is the scaling factor suggested by [32].

$\mathbf{H}_i^t$ is obtained by calculating the weighted average of $\mathbf{h}_{ij}^t$, which represents the spatial impact of the surrounding traffic participants on the agent $A_i$ at time $t$. $\mathbf{H}_i^t$ and $\mathbf{h}_{ii}^t$ contains all information of edges connected to the node $A_i$, which are concatenated and fed into an embedding layer to obtain the fixed length vector $\mathbf{a}_i^t$. Meanwhile, the instance node feature $\mathbf{f}_i^t$ is also embedded as $\mathbf{e}_i^t$:

$$\mathbf{e}_i^t = embed(\mathbf{f}_i^t; \mathbf{W}_{\mathrm{ins}}^e) \tag{15}$$
$$\mathbf{a}_i^t = embed(concat(\mathbf{h}_{ii}^t, \mathbf{H}_i^t); \mathbf{W}_{\mathrm{ins}}^a) \tag{16}$$

where $\mathbf{W}_{\mathrm{ins}}^e$ and $\mathbf{W}_{\mathrm{ins}}^a$ are the embedding weight matrices.

$\mathbf{e}_i^t$ and $\mathbf{a}_i^t$ are input into the instance LSTM to obtain the hidden state $\mathbf{h}1_i^t$, which is the first output of the instance node:

$$\mathbf{h}1_i^t = LSTM(\mathbf{h}2_i^{t-1}, concat(\mathbf{e}_i^t, \mathbf{a}_i^t); \mathbf{W}_{\mathrm{ins}}^r) \tag{17}$$

where $\mathbf{W}_{\mathrm{ins}}^r$ are the weight matrices of the LSTM.

$\mathbf{h}2_i^{t-1}$ is the final output of the instance node $A_i$ at time $t-1$.

*2) Category Layer:* In the multi-participant urban traffic environment, the number of traffic participants of the same category is often more than one. Therefore, the proposed model includes a category layer, in which the super node collects the hidden states $\mathbf{h}1_m^t$ and the cell states $\mathbf{c}_m^t$ from the same kind of instance LSTM, extracting the characteristics of the same kind of traffic participants and feedback it to

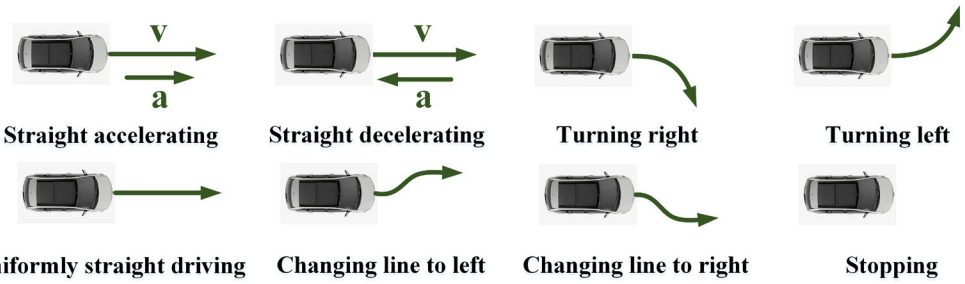


Fig. 5. Eight types of driving behaviours.

the instance node. The category layer enables the model to learn the motion characteristics of different categories of traffic participants and obtain a comprehensive understanding of the traffic scene.

For the $m$th instance node, the hidden state $\mathbf{h}1_m^t$ and the cell state $\mathbf{c}_m^t$ contain the historical trajectory information. Therefore, the proposed approach uses a self-attention mechanism to calculate the motion characteristics $\mathbf{d}_m^t$ of the $m$ th instance node, and computes the average vector of $\mathbf{d}_m^t$ as the features of the super node $C_c^t$ [32]:

$$\mathbf{d}_m^t = \mathbf{h}1_m^t \otimes softmax(\mathbf{c}_m^t) \tag{18}$$

$$\mathbf{F}_c^t = \frac{1}{n}\sum_{m=1}^{n} \mathbf{d}_m^t \tag{19}$$

where $n$ is the number of instance nodes corresponding to the super node $C_c^t$. $\mathbf{F}_c^t$ contains motion characteristics of a category of agents at time $t$ and is included as a part of the input of super node LSTM. There is also a temporal edge between the same super nodes of two consecutive frames. Similar to the instance layer, the temporal edge in the category layer takes $\mathbf{F}_{cc}^t = \mathbf{F}_c^t - \mathbf{F}_c^{t-1}$ as input and outputs the hidden state $\mathbf{h}_{cc}^t$ containing time series information:

$$\mathbf{e}_{cc}^t = embed(\mathbf{F}_{cc}^t; \mathbf{W}_{\text{st}}^e) \tag{20}$$

$$\mathbf{h}_{cc}^t = LSTM(\mathbf{h}_{cc}^{t-1}, \mathbf{e}_{cc}^t; \mathbf{W}_{\text{st}}^r) \tag{21}$$

where $\mathbf{W}_{\text{st}}^r$ and $\mathbf{W}_{\text{st}}^e$ are the embedding weight matrices of LSTM cells.

As another part of the input, $\mathbf{h}_{cc}^t$ is fed into the super node LSTM together with $\mathbf{F}_c^t$:

$$\mathbf{e}_c^t = embed(\mathbf{F}_c^t; \mathbf{W}_{\text{sup}}^e) \tag{22}$$

$$\mathbf{h}_c^t = LSTM(\mathbf{h}_c^{t-1}, concat(\mathbf{e}_c^t, \mathbf{h}_{cc}^t); \mathbf{W}_{\text{sup}}^k) \tag{23}$$

As feedback information, $\mathbf{h}_c^t$ is sent back to the instance layer along the edge from the super node to the instance node. The $m$th instance node corresponding to $C_c^t$ receives $\mathbf{h}_c^t$, and then concatenates $\mathbf{h}1_m^t$ with it. $\mathbf{h}1_m^t$ and $\mathbf{h}_c^t$ are input into an embedding layer to obtain the final output $\mathbf{h}2_m^t$:

$$\mathbf{h}2_m^t = embed(concat(\mathbf{h}1_m^t, \mathbf{h}_c^t); \mathbf{W}_{\text{s}}^r) \tag{24}$$

where $\mathbf{W}_{\text{s}}^r$ are the embedding weight matrices. $\mathbf{h}2_m^t$ collects the trajectory information of the $m$th agent, the influence of the surrounding agents on it, and its motion characteristics.

### C. Information Fusion

In the proposed framework, gaze and scene information generated by the GA and SU modules, respectively, is fused by the IF module. The weights of hidden states $\mathbf{h}2_m^t$ are obtained by processing $\mathbf{h}2_m^t$ and the features $\mathbf{G}$ by attention mechanism:

$$w(\mathbf{h}2_m^t) = softmax(Dot(\mathbf{h}2_m^t, \mathbf{G})) \tag{25}$$

$$\mathbf{B}_A = \sum_{m=1}^{n} Dot(w(\mathbf{h}2_m^t), \mathbf{h}2_m^t) \tag{26}$$

where $n$ is the total number of nodes in the SU module.

$\mathbf{B}_A$ is input into a classifier to obtain $\mathbf{B}$, which is the result of driver behaviour recognition.

The effect of the cross-entropy function has been verified in the field of classification. Therefore, the cross-entropy function is used as the loss function to evaluate the identification results:

$$loss = CrossEntropy(\mathbf{B}, \mathbf{B}_{\text{T}}) \tag{27}$$

where $\mathbf{B}_{\text{T}}$ is the ground truth of driving behaviour. During the training, after each sequencing sample was input into the network, the error was calculated, and all the weight parameters were updated by the back propagation algorithm.

## IV. Experiments and Validation

In this section, the proposed framework is verified in the urban environment with heterogeneous traffic participants. The BLVD dataset was used for validation. The BLVD is a large-scale 5D semantics benchmark collected in Changshu, Jiangsu province, China [33]. The proposed framework is compared with the state-of-the-art approaches using the BLVD dataset. The details of data pre-processing, implementation, evaluation metrics, baseline methods and experimental results are presented in the following subsections.

### A. Dataset and Data Pre-Processing

The BLVD dataset was selected to verify the effectiveness of the proposed method. The BLVD is a dataset containing 7 categories of driving scenarios with 654 calibrated videos, and the sampling frequency is 10Hz. For each frame, the categories (a total of three, pedestrian, rider and vehicle) and the coordinates of surrounding traffic participants are labelled, and each traffic participant is given a unique ID. At the same time, the categories of driver behaviour of the ego vehicle are also labelled. As shown in Fig. 5, driver behaviours are divided

TABLE I
THE CORRESPONDING RELATIONSHIP BETWEEN LABELS ANDDRIVER BEHAVIOURS

| Label | Behaviour of driver | Number of samples (Original / After sampling) |
|---|---|---|
| 1 | Straight accelerating | 2073 / 625 |
| 2 | Straight decelerating | 1415 / 612 |
| 3 | Turning right | 2695 / 519 |
| 4 | Turning left | 2304 / 665 |
| 5 | Uniformly straight driving | 8915 / 749 |
| 6 | Changing lane to left | 1380 / 477 |
| 7 | Changing lane to right | 630 / 622 |
| 8 | Stopping | 1042 / 382 |

into 8 categories and labelled with numbers 1-8. It should be noted here that each frame is labelled with one of the 8 driving behaviours, so that a driving manoeuvre may consist of some number of consecutive frames labelled the same. The labelling of driving behaviours is done manually. In this paper, the labels are used as the ground truth in training.

The proportion of samples of 8 behaviours in the BLVD dataset is not balanced. The behaviour with the largest proportion of quantity is uniformly straight driving, up to 43%. The behaviour with the smallest proportion of quantity is changing the lane to left, only 3%. In the process of training, if the quantity of data for one category is much larger than other categories, this category of data will remarkably dominate the classification result and lead to the problem of imbalanced data [34], [35], In this situation, the trained classification model cannot classify the data from categories with few training samples correctly. Therefore, in order to eliminate the imbalance between different categories of data, the dataset was randomly sampled to obtain a data-balanced training set and eliminate the imbalance between different categories of data. In the selected data, the number of tags in the eight categories was roughly balanced, as shown in Table I. The dataset was divided into training, verification and test sets according to the ratio of 8:1:1.

### B. Details of Implementation

In the experiment, the number of cells in the instance node LSTM was set to 64, and the number of cells in other LSTMs was set to 128. The number of cells in the embedding layer was also set to 128. The learning rate was set to 0.001 and the optimizer was Adam. The training equipment was a computer equipped with an i7-10870H CPU and RTX2060 graphics card.

### C. Evaluation Metrics and Baselines

*1) Evaluation Metrics:* The recall rate and the accuracy of recognition are used to evaluate the performance of the proposed model:

$$accuracy = \frac{m_{\mathrm{T}}}{M_{\mathrm{T}}}$$
$$recall = \frac{TP}{TP + FN} \tag{28}$$

TABLE II
THE EFFECT OF GAZE PREDICTING MODULE

| | KLD | AUC-J |
|---|---|---|
| Dr(eye)ve | 1.63 | 0.84 |
| Our dataset | 1.58 | 0.88 |

where $m_{\mathrm{T}}$ is the number of samples that are correctly identified during the test, $M_{\mathrm{T}}$ is the total number of test set samples, $TP$ is the number of positive samples that are correctly identified, $FN$ is the number of positive samples that are incorrectly identified as negative. In a multi-class class problem, when calculating the recall rate of one category, that category will be regarded as positive, and other categories will be regarded as negative.

*2) Baseline Methods:* During the test, $M_{\mathrm{T}}$ is the total number of test set samples. The proposed method is compared with the following methods:

1. *Scene understanding (SU)*: It only contains the part of trajectory information, without considering the gaze information of driver.
2. *Gaze analysing (GA)*: It only contains the gaze information of driver, without considering the trajectory of the surrounding traffic participants.
3. *Spatial-Temporal fusion convolutional neural network (STFCNN)*: It is a framework that has two streams called spatial stream and temporal stream. The spatial stream is fed with a single original image, and the temporal stream is fed with optical flow images in series. The outputs of the two streams are fused by concatenating [36].
4. *Social LSTM(SL)*: It is a network composed of LSTM, using "Social" pooling layers to handle the information of neighbouring traffic participants [29].

### D. Training and Validation

*1) Effect of Gaze Predicting Method:* The BLVD does not contain gaze images of the driver. Therefore, the gaze predicting method was used to generate the gaze images. The selection of parameters referred to [16], and the gaze predicting method was trained on Dr(eye)ve dataset [15]. The Dr(eye)ve dataset contains 74 segments of videos, each of which contains 7500 camera images and corresponding gaze images (probability map). During the training, the input was a sequence of 16 consecutive frames, and the prediction result was the gaze image of the final frame. To prove that the gaze predicting module was robust enough to be applied to generate gaze images for BLVD, the module was verified on the dataset collected using the eye tracking glasses. The eye tracking dataset contained 165 segments of videos, each of which exceeded 300 frames.

Two metrics were adopted to measure the prediction effect: the Kullback-Leibler divergence (KLD) between the predicted images and the gaze images collected by eye tracking glasses, and the AUC-J [37] between the predicted images and ground truth. Table II compares the performance of the gaze predicting module on two datasets: Dr(eye)ve and the dataset collected by us. It can be seen from Table II that the gaze predicting

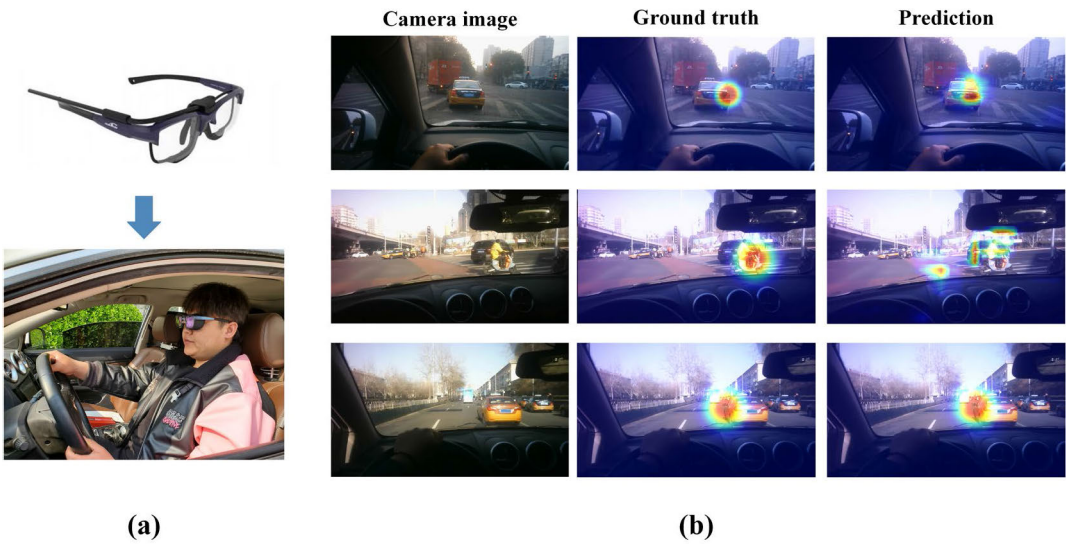


Fig. 6. Gaze predicting results on the images collected by eye tracking glasses.

TABLE III
RECALL RATE OF FOUR CLASSIFIERS IN 8 KINDS OF BEHAVIOURS RECOGNITION

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Average |
|---|---|---|---|---|---|---|---|---|---|
| SVM | 0.915 | **0.945** | 0.940 | 0.923 | 0.861 | 0.901 | 0.951 | 0.931 | 0.915 |
| MLP | **0.974** | 0.900 | **0.945** | **0.930** | **0.925** | **0.923** | **0.965** | **0.934** | **0.938** |
| Decision Tree | 0.846 | 0.825 | 0.834 | 0.846 | 0.803 | 0.825 | 0.858 | 0.868 | 0.842 |
| Random Forest | 0.906 | 0.882 | 0.880 | 0.896 | 0.885 | 0.878 | 0.925 | 0.905 | 0.892 |

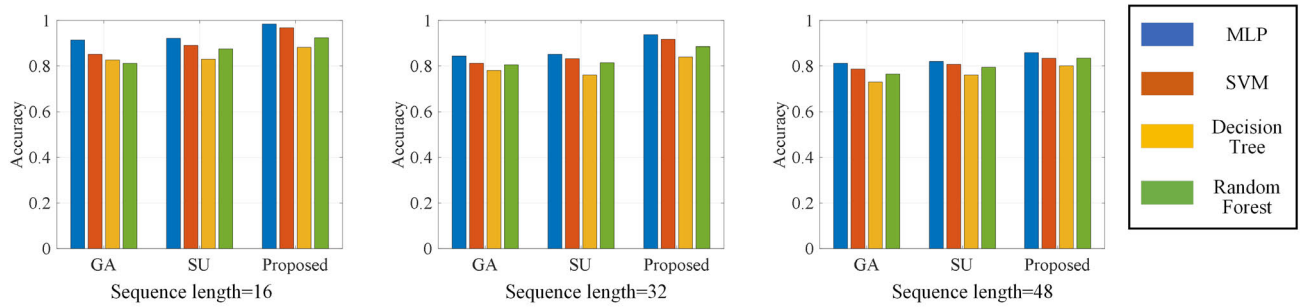


Fig. 7. The verification results on GA, SU and the proposed framework.

module maintained a small error for the eye tracking dataset, showing good robustness. Fig. 6 shows the results of gaze predicting modules for some sample images from the eye tracking dataset. Fig. 6 (a) shows the data acquisition equipment, and (b) overlaps the camera images with corresponding gaze images.

*2) Parameters Adjustment and Classifier Selection:* There are some adjustable parameters in the network that affect the performance of recognition. The parameters of VGG and GNN were set according to [25] and [28]. In the LSTM layer of the GA, the dropout value was set to 0.5 [38].

Experiments were conducted to compare four widely used classifiers: MLP, SVM, Decision Tree and Random Forest [39], [40], [41], [42], [43], [44], [45], [46]. The recognition effects of the four classifiers are shown in Table III. It can be seen from the table that the MLP achieves the highest recall rate in recognition of 7 kinds of driving behaviours, while the recall rate of the Decision Tree is the lowest. The Random Forest overcomes the overfitting problem, which makes it perform better than the Decision Tree. To observe the performance of the framework under different observation sequence lengths with the four classifiers, the length of the input sequences was set as 16, 32 and 48 frames. The results are illustrated in Fig. 7. The Figure graphically compares the recognition accuracy of the different classifiers. The recognition accuracy of the MLP is always higher than that of the SVM, Decision Tree and Radom Forest. Therefore, the MLP is selected as the classifier of the proposed framework.

*3) Driving Behaviour Recognition Effect of Framework:* Fig. 8 visually shows the recognition effects of the proposed framework under four scenes. The orange dots and lines in the figure represent the historical trajectories of participants. The recognition result of Fig. 8 (a) is deceleration, where the vehicles in the view slow down under the command


Fig. 8. The recognition effect of the proposed framework of four scenarios.

of traffic lights and stop in front of the zebra crossing. The driver stares at the vehicle in front of him or her, slows down and keeps a certain distance from the vehicle in front. Fig. 8 (b) shows the acceleration behaviour when the traffic lights change from red to green. The driver still gazes at the vehicle ahead. It is difficult to distinguish deceleration and acceleration behaviours by just analysing the gaze information. The difference between Figs. 8 (b) and (a) is the length of historical trajectories of the vehicles ahead. Since the vehicles in Fig. 8 (b) are in the acceleration phase and the vehicles in Fig. 8 (a) tend to stop, the historical trajectories of the vehicles in Fig. 8 (b) are longer than those in Fig. 8 (a), which becomes the key for the proposed framework to distinguish the two scenarios. In Fig. 8 (c), the driver is turning left. The gaze map shows that the gaze of the driver falls on the van turning at the same time as him or her on the left. This characteristic is used by the proposed framework to successfully recognise the turning behaviour of the driver. Fig. 8 (d) shows an overtaking scenario, where the driver is changing the lane to the right. The black vehicle in front is stopping on the road and flashing taillights, indicating that it has failed. Therefore, the driver of the ego vehicle ignores the static vehicle and focuses on the motorcycle running on the right side in the process of overtaking.

*4) Comparative Experimental Results and Discussion:* Fig. 9 compares the increasing recognition accuracy of the five methods in the training process. It can be seen from the figure that the training effect of the proposed method is the best, the curves of SU, GA and SL see a similar trend beneath the proposed method, and the STFCNN is the worst, staying at the level of 0.8 after 13 epochs. In the comparative experiment, the sequence lengths were set as 4, 8, 16, 32 and 48. The results are illustrated in Fig. 10. The recognition accuracy of

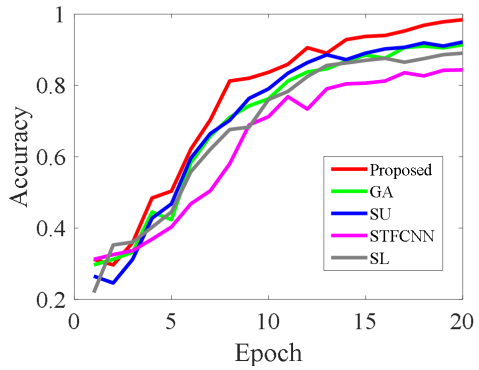


Fig. 9. The variation of recognition accuracy of five methods with epoch in the training process.

driver behaviour in SU is lower than the proposed method because the SU only considers the historical trajectories of traffic participants around the ego vehicle and ignores the gaze information of drivers. Although the GA considers the gaze of the driver, it lacks the accurate location information of the surrounding traffic participants, meaning an insufficient understanding of the traffic scene. Therefore, the accuracy of GA is also lower than the proposed method. When the information of two branches is combined, a higher accuracy is achieved than a single module.

Although compared to the method with scene information only, with the help of gaze prediction, the proposed method can achieve a superior performance, it does not mean that the model will become more accurate as long as the driver's gaze information is included. In our experiments, parameters of gaze predicting module have been adjusted to guarantee a high performance with the KLD around 1.6 (as mentioned in Section IV-D. 2)). With the increase of gaze prediction

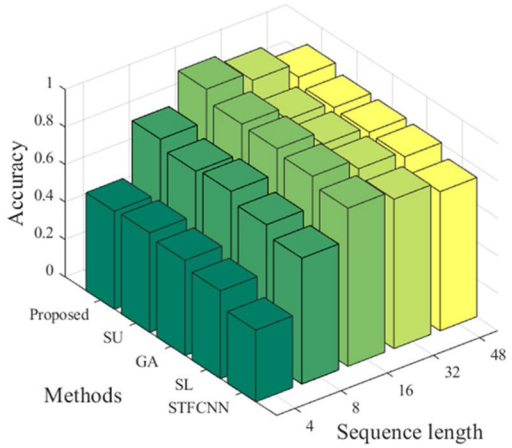


Fig. 10. Comparison of the proposed framework with and four other methods.

error, the behaviour prediction accuracy of the framework will decrease. When the gaze prediction error increases to a certain extent, for example, when the driver is looking to the left, but the gaze prediction shows that the driver is looking to the right, the effect of the fusion framework will be even worse than that of using only the scene information. Thus, for using the proposed framework, the accuracy of gaze predicting module should be guaranteed first.

It can be seen from Fig. 10 that when the number of observation frames is fixed, the accuracy of the proposed method is improved by 6.51%∼11.11% and 3.78%∼9.10% relative to the SU and GA, respectively. When the length of the observation sequence increases, the identification accuracy of all models decreases.

The structure of the STFCNN is similar to that of the GA. However, the STFCNN has just two streams, without the stream to process the gaze images. A lack of gaze information makes it difficult to capture the intention of the driver. The recognition accuracy of GA is 6.86%∼10.52% higher than that of the STFCNN, which shows that the gaze image of the driver can improve the recognition accuracy when only using the image information for the recognition of behaviour. The Social LSTM can just manage the information of trajectories. Compared with the SU, the SL has the social pooling layer to analyse spatial influence between the traffic participants, which is analogous to the spatial edge of the SU. However, the SL does not have a category layer. Hence, it is difficult for the SL to learn the movement patterns of different types of traffic participants. Therefore, the performance of SU is better than SL.

It should be noted that the recognition accuracy does not always increase with the increase in the length of the sequence. When the observation length is 4, behaviour recognition becomes extremely difficult because the input only contains information of 0.4 seconds, resulting in recognition accuracy of less than 0.5 for each model. As the observation length increases from 4 to 16, the input sequence becomes longer, and the information fed to the model becomes ampler. Therefore, the recognition accuracy is improved accordingly. However, as the length of observed sequence continues to increase, the performance of models declines. The reason is that when the observation length is too long, the behaviour of the driver in the sequence may change, such as changing back and forth between going straight and lane changing, which brings disturbance to behaviour recognition. However, the performance of the proposed method is still better than others.

## V. Conclusion and Future Work

In this paper, a framework for fusing the gaze and scene information and recognizing the behaviour of drivers is proposed. The proposed framework is composed of three main modules, namely gaze analysing (GA), scene understanding (SU) and information fusion (IF). The GA module extracts the characteristics of the gaze images with VGG and LSTM. The SU module processes the trajectories of traffic participants with different types at the same time using a hierarchical graph neural network. The IF module uses the attention mechanism to assign different weights to the trajectories of surrounding traffic participants according to the areas where the drivers pay attention and realises the fusion of two types of information. The proposed framework is evaluated using the BLVD dataset. The performance of the proposed framework is compared with the state-of-the-art approaches. The comparative analysis demonstrates and validates the superior performance of the proposed framework driving behaviour recognition.

At the current stage, the traffic scenes with only a small amount of interactive traffic participants are considered. More complex interactive behaviour between different traffic participants will be involved in future work.

## References

[1] X. Ma and I. Andreasson, "Behavior measurement, analysis, and regime classification in car following," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 1, pp. 144–156, Mar. 2007.

[2] B. Higgs and M. Abbas, "Segmentation and clustering of car-following behavior: Recognition of driving patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 81–90, Feb. 2015.

[3] W. Wang, J. Xi, and D. Zhao, "Learning and inferring a driver's braking action in car-following scenarios," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 3887–3899, May 2018.

[4] Z. Ouyang, J. Niu, Y. Liu, and X. Liu, "An ensemble learning-based vehicle steering detector using smartphones," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1964–1975, May 2019.

[5] B. Zhu, J. Zhao, S. Yan, and W. Den, "Personalized lane-change assistance system with driver behavior identification," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10293–10306, Nov. 2018.

[6] Y. Xia, Z. Qu, Z. Sun, and Z. Li, "A human-like model to understand surrounding vehicles' lane changing intentions for autonomous driving," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 4178–4189, May 2021.

[7] H. Zhang and R. Fu, "An ensemble learning–online semi-supervised approach for vehicle behavior recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10610–10626, Aug. 2022.

[8] L. Li, W. Zhao, C. Xu, C. Wang, Q. Chen, and S. Dai, "Lane-change intention inference based on RNN for autonomous driving on highways," *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 5499–5510, Jun. 2021.

[9] C. Lu, F. Hu, D. Cao, J. Gong, Y. Xing, and Z. Li, "Transfer learning for driver model adaptation in lane-changing scenarios using manifold alignment," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3281–3293, Aug. 2020.

[10] C. Lu, F. Hu, D. Cao, J. Gong, Y. Xing, and Z. Li, "Virtual-to-real knowledge transfer for driving behavior recognition: Framework and a case study," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6391–6402, Jul. 2019.

[11] L. Chao, W. Huaji, L. Chen, G. Jianwei, X. Junqiang, and C. Dongpu, "Learning driver-specific behavior for overtaking: A combined learning framework," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 6788–6802, Aug. 2018.

[12] A. Bender, G. Agamennoni, J. R. Ward, S. Worrall, and E. M. Nebot, "An unsupervised approach for inferring driver behavior from naturalistic driving data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3325–3336, Dec. 2015.

[13] Y. Zhang, J. Li, Y. Guo, C. Xu, J. Bao, and Y. Song, "Vehicle driving behavior recognition based on multi-view convolutional neural network with joint data augmentation," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4223–4234, May 2019.

[14] J. Gottlieb, M. Hayhoe, O. Hikosaka, and A. Rangel, "Attention, reward, and information seeking," *J. Neurosci.*, vol. 34, no. 46, pp. 15497–15504, Nov. 2014.

[15] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "DR (eye) VE: A dataset for attention-based tasks with applications to autonomous and assisted driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun./Jul. 2016, pp. 54–60.

[16] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "Predicting the driver's focus of attention: The DR(eye)VE project," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1720–1733, Jul. 2019.

[17] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: Driver attention prediction in driving accident scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4959–4971, Jun. 2022.

[18] M. J. Rahman, S. S. Beauchemin, and M. A. Bauer, "Predicting driver behaviour at intersections based on driver gaze and traffic light recognition," *IET Intell. Transp. Syst.*, vol. 14, no. 14, pp. 2083–2091, Dec. 2020.

[19] S. Martin and M. M. Trivedi, "Gaze fixations and dynamics for behavior modeling and prediction of on-road driving maneuvers," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1541–1545.

[20] S. Martin, S. Vora, K. Yuen, and M. M. Trivedi, "Dynamics of driver's gaze: Explorations in behavior modeling and maneuver prediction," *IEEE Trans. Intell. Veh.*, vol. 3, no. 2, pp. 141–150, Feb. 2018.

[21] N. Pugeault and R. Bowden, "How much of driving is preattentive?" *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5424–5438, Dec. 2015.

[22] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scand. Conf. Image Anal.* Berlin, Germany: Springer, 2003, pp. 363–370.

[23] Z. Wu, L. Su, Q. Huang, B. Wu, J. Li, and G. Li, "Video saliency prediction with optimized optical flow and gravity center bias," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[24] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.

[25] K. Simonyan and A. A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[28] Z. Li, C. Lu, Y. Yi, and J. Gong, "A hierarchical framework for interactive behaviour prediction of heterogeneous traffic participants based on graph neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9102–9114, Jul. 2022.

[29] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.

[30] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4601–4607.

[31] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha, "TrafficPredict: Trajectory prediction for heterogeneous traffic-agents," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 6120–6127.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[33] J. Xue et al., "BLVD: Building a large-scale 5D semantics benchmark for autonomous driving," in *Proc. Int. Conf. Robot. Automat. (ICRA)*, May 2019, pp. 6685–6691.

[34] G. E. Batista, R. C. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, 2004.

[35] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[36] Y. Hu, M. Lu, and X. Lu, "Spatial-temporal fusion convolutional neural network for simulated driving behavior recognition," in *Proc. 15th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2018, pp. 1271–1277.

[37] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[39] D. W. Ruck, S. K. Rogers, and M. Kabrisky, "Feature selection using a multilayer perceptron," *J. Neural Netw. Comput.*, vol. 2, no. 2, pp. 40–48, 1990.

[40] G. S. Aoude, V. R. Desaraju, L. H. Stephens, and J. P. How, "Driver behavior classification at intersections and validation on large naturalistic data set," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 724–736, Jun. 2012.

[41] A. D. Gordon, L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," *Biometrics*, vol. 40, no. 3, p. 358, 1984.

[42] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

[43] L. Dong et al., "Very high resolution remote sensing imagery classification using a fusion of random forest and deep learning technique—Subtropical area for example," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 113–128, 2020.

[44] S. Luan, Y. Gao, W. Chen, N. Yu, and Z. Zhang, "Automatic modulation classification: Decision tree based on error entropy and global-local feature-coupling network under mixed noise and fading channels," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1703–1707, Aug. 2022.

[45] M. P. Neto and F. V. Paulovich, "Explainable matrix–visualization for global and local interpretability of random forest classification ensembles," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1427–1437, Feb. 2021.

[46] D. Streeb et al., "Task-based visual interactive modeling: Decision trees and rule-based classifiers," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 9, pp. 3307–3323, Sep. 2022.

**Yangtian Yi** received the B.E. degree from the Beijing Institute of Technology (BIT), Beijing, China, in 2020, where he is currently pursuing the M.A.Sc. degree in mechanical engineering. His research interests include intelligent vehicles, driver behavior modeling, and gaze of drivers.

**Chao Lu** received the B.S. degree in transport engineering from the Beijing Institute of Technology, Beijing, China, in 2009, and the Ph.D. degree in transport studies from the University of Leeds, Leeds, U.K., in 2015. In 2017, he was a Visiting Researcher with the Advanced Vehicle Engineering Centre, Cranfield University, Cranfield, U.K. He is currently an Associate Professor with the School of Mechanical Engineering, Beijing Institute of Technology. His research interests include intelligent transportation and vehicular systems, driver behavior modeling, reinforcement learning, and transfer learning and its applications.

**Boyang Wang** received the B.S. and Ph.D. degrees in vehicle engineering from the Beijing Institute of Technology, Beijing, China, in 2013 and 2020, respectively. He was a Visiting Researcher with the Interaction Digital Human Group, CNRS-UM LIRMM, from 2017 to 2019. He was a Post-Doctoral Researcher with the Key Laboratory of Machine Perception (MOE), Peking University, from 2020 to 2022. He is currently an Assistant Professor with the School of Mechanical Engineering, Beijing Institute of Technology. His research interests include intelligent vehicles, driver behavior, motion planning and control, and vehicle dynamics modeling.

**Long Cheng** received the B.S. degree in transport and traffic and the Ph.D. degree in transport engineering from Southeast University, Nanjing, China, in 2011 and 2016, respectively. He is currently an Associate Professor with the School of Transportation, Southeast University. His research interests include multimodal transport, shared mobility, travel behavior analysis, and transport and land use integration.

**Zirui Li** received the B.S. degree from the Beijing Institute of Technology (BIT), Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree in mechanical engineering. He is a Visiting Researcher with the Delft University of Technology (TU Delft). His research interests include intelligent vehicles, driver behavior modeling, and transfer learning.

**Jianwei Gong** (Member, IEEE) received the B.S. degree in mechanical engineering from the National University of Defense Technology, Changsha, China, in 1992, and the Ph.D. degree in mechanical engineering from the Beijing Institute of Technology (BIT), Beijing, China, in 2002. From 2011 to 2012, he was a Visiting Professor with the Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently a Professor and the Director of the Research Centre of Intelligent Vehicles, BIT. His research interests include intelligent vehicles, environmental perception, motion planning, and control for autonomous driving.