DPFT: Dual Perspective Fusion Transformer for Camera-Radar-Based Object Detection

Fent, F.; Palffy, A.; Caesar, H.

DOI
10.1109/TIV.2024.3507538

Publication date
2025

Document Version
Final published version

Published in
IEEE Transactions on Intelligent Vehicles

Citation (APA)
Fent, F., Palffy, A., & Caesar, H. (2025). DPFT: Dual Perspective Fusion Transformer for Camera-Radar-Based Object Detection. *IEEE Transactions on Intelligent Vehicles*, *10*(11), 4929-4941. https://doi.org/10.1109/TIV.2024.3507538

DPFT: Dual Perspective Fusion Transformer for Camera-Radar-Based Object Detection

Felix Fent, Andras Palffy, and Holger Caesar

Abstract—**The perception of autonomous vehicles has to be efficient, robust, and cost-effective. However, cameras are not robust against severe weather conditions, lidar sensors are expensive, and the performance of radar-based perception is still inferior to the others. Camera-radar fusion methods have been proposed to address this issue, but these are constrained by the typical sparsity of radar point clouds and often designed for radars without elevation information. We propose a novel camera-radar fusion approach called Dual Perspective Fusion Transformer (DPFT), designed to overcome these limitations. Our method leverages lower-level radar data (the radar cube) instead of the processed point clouds to preserve as much information as possible and employs projections in both the camera and ground planes to effectively use radars with elevation information and simplify the fusion with camera data. As a result, DPFT has demonstrated state-of-the-art performance on the K-Radar dataset while showing remarkable robustness against adverse weather conditions and maintaining a low inference time.**

Index Terms—**Perception, object detection, sensor fusion, radar, camera, autonomous driving.**

I. Introduction

Autonomous driving is a promising technology that has the potential to increase safety on public roads and provide mobility to people for whom it was previously not accessible. However, leveraging this technology requires autonomous vehicles to operate safely within a multitude of different environmental conditions. These conditions include everyday driving situations such as nighttime driving or driving under severe weather conditions, but also critical situations where the autonomous vehicle (AV) has to react quickly or maintain general functionality after a sensor failure.

The perception of most autonomous driving systems is based on either camera or light detection and ranging (lidar) sensors. While camera sensors are cost-effective, they depend on ambient light and do not provide depth information [1]. In contrast, lidar sensors provide accurate measurements of the surroundings but come at a high cost. More importantly, neither camera nor lidar sensors are robust against severe weather conditions like rain, fog, or snow [2]. On the other hand, radio detection and ranging (radar) sensors are cost-effective and robust against challenging environmental conditions but do not yet provide comparable object detection qualities as lidar or camera-based perception methods due to their low spatial resolution and high noise level [3].

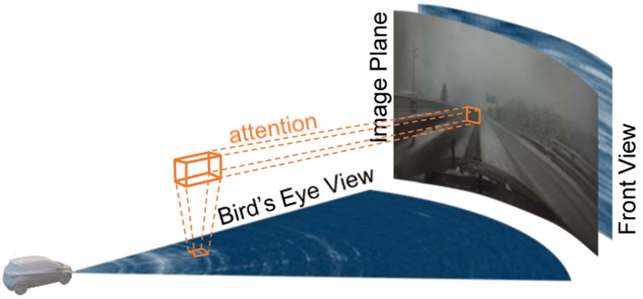


Fig. 1. Illustration of the dual perspective fusion procedure. The 4D radar cube is projected onto a front and bird's eye view to create a parallel and perpendicular perspective to the camera image. This simplifies the camera-radar fusion and maintains the complementary sensor features. Object features are queried from these perspectives via an attention mechanism and used to regress 3D detections.

A potential solution to overcome the limitations of individual sensor technologies is the combination of multiple sensor modalities, also referred to as sensor fusion. Nevertheless, sensor fusion remains challenging due to inherent differences between the camera and radar sensors, such as the perceived dimensionality (2D vs. 3D), data representation (point cloud vs. grid), and sensor resolution [4].

In this paper, we propose a novel camera and radar sensors fusion method to provide a robust, performant, yet cost-effective method for 3D object detection. While camera-radar fusion has been done before [4], previous methods mostly rely on radar point cloud data, thus suffering from a sparse data representation and facing the challenge of combining images with point clouds. On the other hand, fusion approaches that utilize raw radar data solely rely on radar data in a bird's eye view (BEV) representation. Therefore, they are fusing data from the image plane with data from a perpendicular BEV plane on one side and discarding the advantages of modern 4D radar sensors on the other.

Received 16 October 2024; revised 18 November 2024; accepted 25 November 2024. Date of publication 27 November 2024; date of current version 5 December 2025. *(Corresponding author: Felix Fent.)*

Felix Fent is with the Technical University of Munich, School of Engineering and Design, Institute of Automotive Technology, Munich Institute of Robotics and Machine Intelligence (MIRMI), 80992 München, Germany (e-mail: felix.fent@tum.de).

Andras Palffy is with the Microwave Sensing, Signals and Systems (MS3) Group, Department of Microelectronics, Delft University of Technology, and Perciv AI, 2628, BC Delft, The Netherlands.

Holger Caesar is with the Intelligent Vehicles (IV) Section, Department of Cognitive Robotics, Delft University of Technology, 2628, BC Delft, The Netherlands.

The code is made available as open-source software under https://github.com/TUMFTM/DPFT.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2024.3507538.

Digital Object Identifier 10.1109/TIV.2024.3507538

Our proposed method overcomes these limitations by fusing camera data with raw radar cube data to mitigate the differences in sensor resolution and benefit from a structured grid representation for both sensor modalities. However, directly consuming the raw radar cube would be unfeasible due to its high demand for computational resources. Therefore, we developed a projection method that reduces the 4D radar cube to two 2D grids while maintaining important features and providing a low sensitivity to input noise. As a result, the proposed fusion architecture utilizes radar data from both a BEV and a front-view perspective as shown in Fig. 1. With this dual perspective approach, we create a corresponding data source to the image plane to support camera-radar fusion and incorporate data from the BEV plane to exploit all radar dimensions. All three data inputs are then fed through a ResNet feature extractor and subsequent Feature Pyramid Network (FPN) neck before they are combined in the fusion module. However, our method does not require a combined feature space but queries 3D objects directly from these individual perspectives, thus preventing the loss of information caused by a uniform feature space or raw data fusion [5]. To enable this, we introduce a modified deformable attention [6] mechanism that allows both cartesian and spherical reference point projection to realize a modality-agnostic sensor fusion.

In summary, our main contributions are three-fold:

- We propose an efficient sensor fusion approach that projects the radar cube onto two perspectives, thus simplifying the camera-radar fusion, avoiding the limitations of sparse radar point clouds, and leveraging the advantages of 4D radar sensors.
- We are the first to fuse 4D radar cube data with image data by proposing a novel fusion method that does not rely on a common BEV representation to fuse camera and radar data.
- Experiments show that our method achieves state-of-the-art results in severe weather conditions on the challenging K-Radar dataset thus offering greater robustness and lower inference times than previous methods.

II. Related Work

The proposed method combines the complementary features of camera and radar sensors to create a robust, performant, and cost-effective method for 3D object detection. However, to understand the motivation behind the proposed Dual Perspective Fusion Transformer (DPFT), it is important to understand the concepts and limitations of unimodal object detection methods and available datasets.

A. Camera-Radar Datasets

While there are many datasets within the autonomous driving domain, most of them do not include radar sensor data [4]. The nuScenes [7] dataset only provides 3D radar point clouds and has been criticized for its limited radar data quality [8], [9]. The RadarScenes [9] dataset provides higher-quality radar data but only on a point cloud level and does not provide object annotations. Both the View-of-Delft [10] as well as the TJ4DRadSet [11] datasets provide 4D radar data and corresponding bounding boxes but do not include raw radar data. The CARRADA [12], RADIATE [13], and CRUW [14] datasets are one of the few proving cube-level radar data but are limited to 3D radar data and do not provide 3D object annotations. The RADIal [15] dataset provides raw 4D radar data but originally only included 2D bounding box annotations. Even if 3D annotations were recently added by Liu et al. [3], the RADIal dataset does not support the retrieval of 4D radar cube data, has a limited extent, and does not include data within severe weather conditions, which is one of the main motivations for radar applications. For these reasons, the K-Radar [16] dataset is the only suitable dataset for our experiments. The dataset itself includes raw (cube-level) radar data from a 4D radar sensor as well as the data from two lidar sensors, 4 stereo cameras, one GNSS, and two IMU units. In addition, it provides 3D annotated bounding boxes for 34994 frames sampled from 58 different driving scenes and is split into 49.9% train and 50.1% test data.

B. Camera-Based 3D Object Detection

Camera-based monocular 3D object detection methods can be divided into three major categories: data lifting, feature lifting, and result lifting methods [17], [18].

Data lifting methods directly lift 2D camera data into 3D space to detect objects within it [17]. Out of those, pseudo-lidar methods [19] are most commonly used to transform camera images into 3D point clouds. Besides that, learning-based approaches [20] can be used for data lifting, and even most feature lifting methods [21], [22] can directly be applied to image data.

Feature lifting methods first extract 2D image features, which are then lifted into 3D space to serve as the basis for the prediction of 3D objects [17]. Within this category, there are two dominant lifting strategies: one “pushes” (splatting) the features from 2D into 3D space [21] and the other “pulls” (sampling) the 3D features from the 2D space [22].

Result lifting methods are characterized by the fact that they first estimate the properties of the objects in the 2D image plane and then lift the 2D detections into 3D space [17]. Inspired by the taxonomy used within the field of 2D object detection, these methods can be further divided into one-stage and two-stage detectors. One-stage detectors regress 3D objects directly from 2D image features and are typically characterized by fast inference speeds. Representative methods of this category are anchor-based detectors [23] or anchor-free models like [24]. Two-stage detectors first generate region proposals before they refine those proposals to predict 3D objects [17]. Methods from this category can use either geometric priors [25], [26] or model-based priors [27].

Even if different strategies have been developed over the years, the biggest challenge for camera-based 3D object detection remains the lifting from 2D to 3D space due to the inability of camera sensors to directly measure depth information [1]. Furthermore, camera sensors are susceptible to illumination changes and severe weather conditions [1], limiting their robustness in field applications.

C. Radar-Based 3D Object Detection

Radar sensors, in contrast to cameras, are robust against severe weather conditions [2] and are able to measure not only depth information but also intensities and relative velocities via the Doppler effect. This is due to the fact that radar sensors perceive their environment by actively emitting radio wave signals and analyzing their responses [4]. However, this analysis requires multiple processing steps, which is why radar-based 3D object detection methods are categorized by the data level they are operating on [3]. The first category of methods operates directly on the raw analog-to-digital converted (ADC) radio wave signals. These ADC signals are then converted from the temporal to the spatial domain using a Discrete Fast Fourier Transformation (DFFT). The resulting data representation is a discrete but dense radar cube and the basis for the second type of detection methods. Finally, this data can be further reduced by only considering data points with high response values, leading to a spare point cloud representation and the input to the third (and most common) type of methods [4].

Methods operating on the raw ADC signals are rare due to limited data availability, high memory requirements, and the abstract data format. Even if Yang et al. [28] achieved promising results on the RADIal [15] dataset, Liu et al. [3] showed that ADC data has no advantages over radar cube data. Thus, the benefits of replacing the DFFT with a neural network remain questionable.

Detection methods utilizing cube-level radar data can be subdivided into those using 2D, 3D, or 4D radar data. Methods utilizing 2D radar data use either range-azimuth (RA) [29], [30], [31] or range-doppler (RD) [32], [33] measurements, while 3D methods either use multiple 2D projections [34], [35] or the whole range-azimuth-doppler (RAD) cube [36], [37]. However, none of the above mentioned methods are used for 3D, but only 2D object detection, and neither of those utilizes the elevation information of modern 4D (3+1D) radar sensors.

Methods relying on radar point clouds are the most common type of detectors and can be further divided into grid, graph, and point-based methods. Grid-based methods [38], [39], [40], [41] discretize the point cloud space to derive a regular grid from the sparse point cloud. Graph-based methods [42], [43], [44] create connections (edges) between the points (vertices) to utilize graph neural networks (GNNs) for object detection tasks. Lastly, point-based methods [45], [46], [47], [48], [49] use specialized network architectures to directly detect objects within the sparse irregular radar point clouds.

Generally, radar-based object detection methods are robust against severe weather conditions but do not yet achieve competitive performance values. This is mainly due to the radar's lower spatial resolution, higher noise level, and limited capability to capture semantic information.

D. Camera-Radar Fusion for 3D Object Detection

The complementary sensor characteristics of camera and radar sensors make them promising candidates for sensor fusion applications. These fusion methods can be divided into data-level, object-level, and feature-level fusion methods.

Data-level fusion aims to directly combine the raw data from both sensor modalities. Following this approach, Nobis et al. [50] was the first to propose a camera-radar fusion model that projected the radar points into the camera image and used a hierarchical fusion strategy to regress objects from it. On the other side, Bansal et al. [51] projected the semantic information of the camera image onto the radar point cloud (similar to PointPainting [52]) and detected objects within the enriched radar data. Nevertheless, data-level fusion is associated with a high loss of information due to the differences in sensor resolution [5] and challenging due to different data representations and dimensionalities.

Object-level fusion addresses these challenges by using two separate networks for both modalities independently and only combining their detection outputs. Using this technique, Jha et al. [53] fused 2D objects from a camera and radar branch, while Dong et al. [54] combined an object-level with a data-level fusion approach to detect 3D objects on a proprietary dataset. Most recently, Zhang et al. [55] fused the outputs of a radar-based method [37] with the detections of a camera-lidar fusion method and achieved state-of-the-art results on the K-Radar dataset [16]. However, their method, solely relying on camera and lidar data, outperformed the radar fusion method only slightly, thus showing that the capabilities of object-level fusion are limited. This is because object-level fusion exclusively depends on the final detection outputs, neglecting any intermediate features [4]. As a result, the final detection quality relies heavily on the performance of the individual modules and does not fully utilize complementary sensor features [4].

Feature-level fusion aims to combine the advantages of both methods by first extracting features from each modality separately, fusing them at an intermediate level, and finally predicting objects based on their combined feature space. Therefore, it allows to address individual sensor aspects and benefits from a combination of their unique properties. However, finding a suitable feature space to combine both modalities remains challenging. Besides early attempts to combine region proposals from camera and radar branches [56], [57], [58] or feature-level fusion on the image plane [59], [60], most recent methods focus on a bird's eye view (BEV) feature representation.

Using a BEV feature representation, Harley et al. [22] proposed a method to combine rasterized ("voxelized") radar point cloud data with camera data and outperformed their camera baseline on the nuScenes [7] dataset. Similarly, Zhou et al. [61] fused rasterized and temporally encoded radar point cloud data with image data in the BEV space and reported an increased detection quality. However, both methods utilize only 3D radar data, not considering modern 4D radar sensors. Addressing this issue, both Xiong et al. [62] as well as Zheng et al. [63] proposed a method to fuse camera and 4D radar point cloud data in a BEV space. While achieving good results on the TJ4DRadSet [11] and View-of-Delft [10] dataset, these methods solely rely on radar point cloud data. However, radar point cloud data is not only difficult to fuse due to its irregular, sparse data structure but also contains significantly less information, which is lost during signal processing and adverse to accurate environment perception [64].

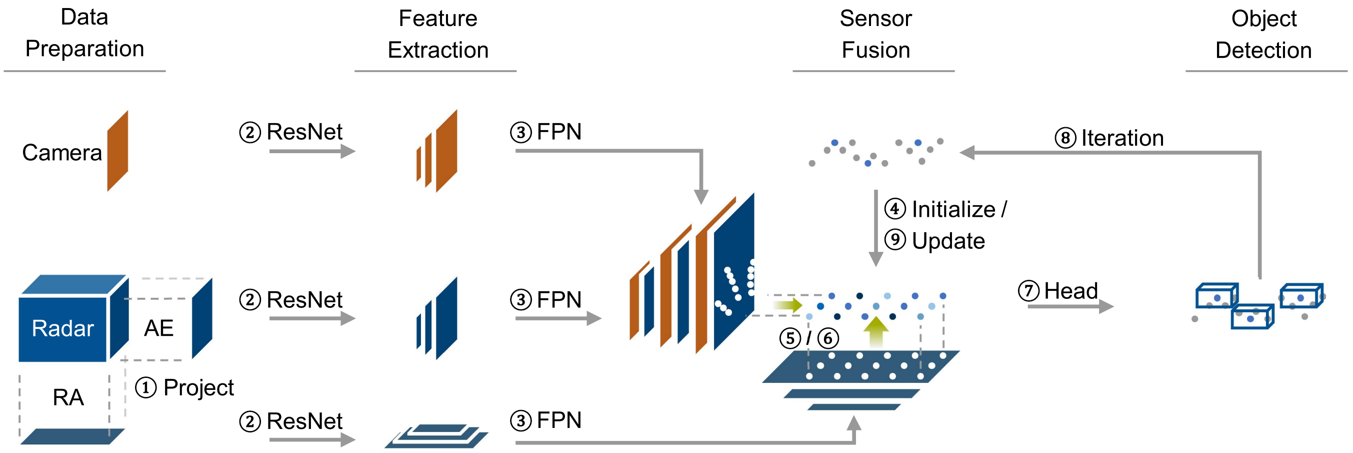


Fig. 2. The DPFT model overview shows the essential steps to fuse camera data with raw 4D radar data and retrieve objects from it. First ①, the data of the 4D radar cube is projected onto the range-azimuth (RA) and azimuth-elevation (AE) plane. Second ②, the two radar perspectives and the camera data are fed through individual ResNet backbones to extract essential features from them. In the ③ step, Feature Pyramid Networks (FPN) are used to align the dimensions of the multi-level feature maps. To fuse the features of the different perspectives, a set of query points is initialized in 3D space in the ④ step and projected onto the different perspectives in the ⑤ step. After that, the features hit by the projection points are fused in the associated query points, using deformable attention ⑥. A classification and regression head is used in ⑦ to retrieve bounding boxes from the queried features. Finally, the regressed bounding box positions are used as new query points in step ⑧ and their features are updated ⑨ in an iterative process to refine the bounding box proposals.

To prevent this loss of information, Liu et al. [3] proposed a method to fuse raw radar data with camera image data, similar to our approach. However, their method relies on an intermediate BEV representation, which increases the demand on computational resources and limits their ability to encode various 3D structures [65]. Moreover, their method does not utilize the elevation information of modern 4D radar sensors, but solely relies on radar data in the range-azimuth (BEV) plane. To overcome these limitations, we propose a novel method that does not require a uniform feature representation and exploits all radar dimensions.

III. Methodology

The Dual Perspective Fusion Transformer (DPFT) is designed to address the main challenges of multimodal sensor fusion, which are caused by the differences in the perceived dimensionality, data representations, and sensor resolutions. First, it utilizes raw cube-level radar data to preserve as much information as possible and lower the resolution differences between camera and radar data. Second, cube-level radar data is given in a structured grid representation, thus avoiding the fusion of point cloud and image data. Third, two projections are created from the 4D radar cube. One parallel to the image plane to support the fusion between camera and radar and another perpendicular to it to preserve the complementary radar information. Besides that, the model design aims to achieve a low inference time and is designed with no interdependencies between the two modalities such that the overall model remains operational even if one sensor modality fails. However, to achieve that, multiple steps are required, which are shown in Fig. 2 and explained in the following.

A. Data Preparation

The input data itself poses the greatest challenge for multimodal sensor fusion due to the differences in data resolution and dimensionality. Camera sensors capture the environment as a projection onto the 2D image plane, while radar sensors typically capture measurements in the range-azimuth (BEV) plane. Broadly speaking, these two perception planes are perpendicular to one another, which makes them difficult to fuse due to their small intersection. To counteract this, our method is built on 4D radar data with three spatial dimensions and one Doppler dimension. This allows us to create a physical relationship between the two data sources. However, working with 4D data is not ideal for two reasons. First, lifting camera data into 3D space is challenging due to the missing depth information, and second, processing high dimensional data has a high demand on computational resources. Resolving this dilemma, the radar data is projected onto the range-azimuth plane as well as the azimuth-elevation plane. This way, we can create a complementary data source to the camera data while reducing the data size and creating a physical relationship between the image and the BEV plane to regress 3D objects.

To address the challenges associated with diverging data formats and sensor resolutions, our method is based on raw (cube-level) radar data. Usually, radar data is given as an irregular, sparse point cloud with a few hundred points per sample, while camera data is represented in a structured grid format with millions of pixels. Not only is it difficult to fuse these two data formats, but a fusion is also associated with a high loss of information or computational overhead [5]. Furthermore, radar point clouds are the results of a multistage signal processing chain (explained in Section II) during which a lot of information is lost and which deteriorates perception performance [3]. Therefore, our method utilizes raw (cube-level) radar data, avoiding the loss of information, creating a uniform data representation, and lowering the differences in data resolution.

Following this idea, the 4D radar cube is projected onto the range-azimuth (RA) and azimuth-elevation (AE) plane. However, to avoid the loss of important information and minimize the sensitivity to input noise, the design of the projection (dimensional reduction) follows a three-step process. First, a set of 30 initial radar features was defined that were proven to be significant to radar-based perception by previous studies [66], [67]. Secondly, a model was trained on all 30 radar features before the weights of the first model layer were analyzed to determine the importance of individual features to the converged model. Lastly, a sensitivity analysis was conducted where noise was added to individual input features and the changes in the output were monitored to determine the sensitivity of the model to input noise. As a result, the maximum, median, and variance of the amplitude and Doppler values were chosen to be extracted during the radar data projection. In addition, the first and last three cells of the radar cube are cut off to avoid DFFT artifacts in the AE projection. Besides that, the image data is rescaled to an input height of 512pixels using bilinear interpolation to lower the demand on computational resources.

B. Feature Extraction

The multimodal input data is fed to consecutive backbone and neck models to deduce expressive features for the desired detection task. Every input is fed to an individual backbone model resulting in three parallel backbones. The purpose of the backbone networks is the extraction of expressive, higher-dimensional features for the subsequent sensor fusion and is chosen to be a ResNet [68] architecture. Since the standard ResNet implementation resizes the inputs to a height of 256, the resulting feature maps of all inputs have similar spatial dimensions. In addition, multi-scale feature maps are extracted from intermediate backbone layers (to detect objects at different scales) and skip connections are used to directly pass the input data to the neck models [69]. More specifically, a ResNet-101 is used for the camera data and a ResNet-50 for both radar data inputs. The larger image backbone is chosen because of the higher image data resolution compared to the radar data. All backbones have been pre-trained on the ImageNet database [70] and a single 1x1 convolution layer is added in front of the radar backbones to make them compatible with the six feature dimensions of the radar data.

The neck models are responsible for feature alignment and ensuring homogeneous feature dimensions. They align the feature dimensions of the multi-scale feature maps and the sensor raw data, which is required for the subsequent sensor fusion. In addition, it also exchanges information between the four feature maps (from three backbone models and the raw input data). For this purpose, a Feature Pyramid Network [71] with an output feature dimension of 16 is used.

C. Sensor Fusion

Our sensor fusion model allows the direct querying of fused features from the individual inputs and the retrieval of objects from them. Therefore, a combined intermediate feature space is not required. To achieve this, multi-head deformable attention [6] is used, which was originally developed for camera-based object detection. This method projects reference points onto camera images to query features from the surrounding pixels. Therefore, this method allows to attend to a fixed number of keys in an image (or feature map), regardless of their spatial size. While this projection was originally designed for a pinhole camera model, we introduce a spherical reference point projection to utilize it for low-level radar data.

Our resulting fusion module consists of five distinct steps. First, the reference points are initialized as a set of 400 evenly distributed 3D points in a polar space with feature values sampled from a uniform distribution and cover the entire field of view (FoV) of the sensor. Next, the reference points are fed to a self-attention layer to allow the exchange of information between queries, which becomes important during the iterative refinement. After that, the reference points are projected onto the camera and the dual radar perspectives in the third step. Based on these projections, deformable cross-attention is used to query features from the (positional encoded) multi-level feature maps. In the last step, the queried features are passed through a feed-forward network (FFN) before they are combined in a max pooling layer. Besides that, each of the attention and FFN layers includes dropout, addition, and normalization layers. With this approach, multiple sensors from different modalities can be fused as long as a projection of the query points onto the sensor feature maps exists.

D. Object Detection

The detection head predicts object bounding boxes based on the fused query features and is separated from the fusion module to allow for multi-task applications. Following [6], [72], [73], we use an interactive output refinement process where the predicted bounding box centers and the previous query features are used for another three attention cycles. As a result, we get object bounding boxes represented by their 3D center point (x, y, z), size (l, w, h), heading angle θ, and class label. The detection head design follows the example of other sparse detectors [72], [73] and consists of three consecutive linear layers. However, DPFT uses a specific activation function for each bounding box component. The center point prediction utilizes an identity function due to its unrestricted value range, the bounding box size uses a ReLu [74] activation function, and the heading angle is predicted by a hyperbolic tangent function. This is due to the fact that the heading angle is not predicted directly but rather split into its $\sin\theta$ and $\cos\theta$ components, since it is shown that the model training benefits from a continuous output space [75]. The class label is predicted by a sigmoid activation function and chosen to be the maximum across all classes.

E. Model Training

The model training uses a set-to-set loss with a one-to-one matching as introduced by DETR [76]. The loss function itself is composed of a focal loss [77] for classification and an L1 regression loss for all bounding box components [72]. The loss weights for these two terms are set to one such that the final loss function can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{box}}. \tag{1}$$

TABLE I
3D Object Detection Results for the Test Data of the K-Radar Dataset Revision v1.0

Method	Modality	Norm.	Overcast	Fog	Rain	Sleet	Light Snow	Heavy Snow	Total mAP
RTNH [16]	R	49.9	56.7	52.8	42.0	41.5	50.6	44.5	47.4
Voxel-RCNN [80]	L	81.8	69.6	48.8	47.1	46.9	54.8	37.2	46.4
CasA [81]	L	82.2	65.6	44.4	53.7	44.9	62.7	36.9	50.9
TED-S [82]	L	74.3	68.8	45.7	53.6	44.8	63.4	36.7	51.0
VPFNet [83]	C + L	81.2	76.3	46.3	53.7	44.9	63.1	36.9	52.2
TED-M [82]	C + L	77.2	69.7	47.4	54.3	45.2	64.3	36.8	52.3
MixedFusion [55]	C + L	**84.5**	**76.6**	53.3	**55.3**	49.6	**68.7**	44.9	55.1
EchoFusion [3]	C + R	51.5	65.4	55.0	43.2	14.2	53.4	40.2	47.4
DPFT (ours)	C + R	55.7	59.4	**63.1**	49.0	**51.6**	50.5	**50.5**	**56.1**

The optimization scheme uses an AdamW [78] optimizer with a learning rate of 1×10^{-4} and a constant learning rate throughout the training. All models are trained with a batch size of 4 and a maximum of 200 epochs (∼72 h).

IV. Results

All reported results are achieved on the K-Radar [16] test set and are in line with the official evaluation scheme, which is based on the KITTI [79] protocol. For comparability, the benchmark results of Table I were obtained on the original version (revision v1.0) of the dataset, while all other results were achieved on the revised version (revision v2.0) of the dataset, which is preferred since it includes corrected object heights and previously missing object labels. For development purposes, we split the train data into 80% train and 20% validation data, the test set remains unmodified.

Since the published version of EchoFusion [3] was only evaluated on the first 20 scenes of the K-Radar dataset, limited to a field of view (FoV) of ±20° (instead of ±50°) and did not use the official evaluation script, we retrained the EchoFusion [3] model on the full dataset and evaluated it in accordance with the official evaluation scheme. All other results are in line with the literature.

The results of Table I show that our Dual Perspective Fusion Transformer achieves state-of-the-art performance on the challenging K-Radar dataset. The DPFT model achieves a mean average precision (mAP) value of 56.1% at an intersection over union (IoU) threshold of 0.3 for 3D bounding box detection across all scene types. To account for any non-deterministic training behavior, the model is trained multiple times with different random seeds, such that 56.1% represents the mean across three runs with a standard deviation of 1.1%. Our proposed camera-radar fusion model outperforms both the radar-only RTNH [16] baseline model as well as the recently proposed EchoFusion [3] camera-radar fusion. In comparison to state-of-the-art lidar or camera-lidar fusion models, it shows a significantly lower performance in normal conditions but outperforms them in particularly difficult weather conditions like fog, sleet, or heavy snow. This is most likely due to the radar's lower spatial resolution but higher robustness against environmental influences.

TABLE II
3D Object Detection Results for Different Input Modalities on the Test Data of the K-Radar Dataset Revision v2.0

	C	R_{AE}	R_{RA}	R	C + R_{AE}	C + R_{RA}	C + R
mAP	8.9	4.4	35.0	36.2	11.1	48.5	50.5

The subscripts AE and RA describe the usage of just a single input perspective, namely azimuth-elevation or range-azimuth.

The comparison of different sensor modalities, as shown in Table II, provides evidence for the effectiveness of our sensor fusion approach. It can be shown that the detection quality of the sensor fusion method exceeds even the combined performance of the individual sensor modalities, thus highlighting the effective use of the complementary sensor features. While the camera-only (C) configuration is similar to DETR3D [72] it struggles with the multitude of severe weather scenarios, the small backbone size, and the inability to utilize multi-view camera images. The results for the fusion of camera data with radar data from the range-azimuth (RA) plane in comparison to the fusion with data from the azimuth-elevation (AE) plane demonstrate the importance of the different perception planes for 3D object detection. However, the results with both radar perspectives, in comparison to only one perspective, suggest that the correspondence of the radar data and the camera data in the image plane, in combination with the physical relationship between the two radar perspectives, supports the fusion of the two sensor modalities. This shows the importance of the complementary information from the RA perception plane on one side and the benefits of the additional AE plane for the association between camera and radar on the other.

A. Robustness

The experimental results show the robustness of the DPFT model in two aspects: robustness against severe weather conditions and robustness against sensor failure. The robustness against severe weather conditions can be seen in Fig. 3 and shown by comparing the model performance under normal (norm.) conditions with the performance under different weather conditions of the K-Radar [16] dataset. As shown in Table I, the highest performance decrease for the DPFT model can be

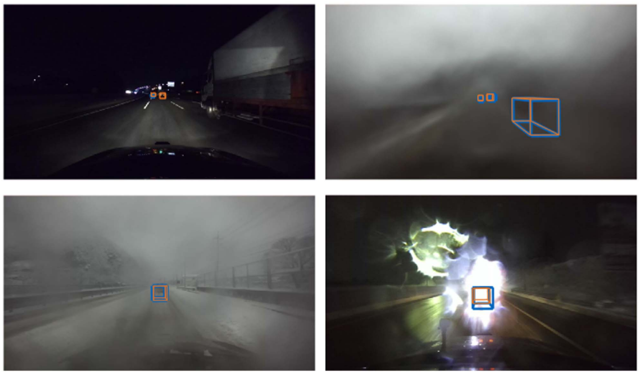


Fig. 3. Exemplary results of the model performance under night, rain, snow, and backlight conditions. The ground truth is shown in blue and the model prediction in orange.

TABLE III
RESULTS WITH SIMULATED SENSOR FAILURE ON THE TEST SET OF THE K-RADAR DATASET REVISION V2.0

Trained	Tested	mAP	$\text{mAP}_{\text{pre-trained}}$	$\text{mAP}_{\text{dropout}}$
C	C	8.9	-	-
R	R	36.2	-	-
C + R	C	1.1	0.0	9.2
C + R	R	11.1	12.8	37.5
C + R	C + R	50.5	51.4	38.3

observed for the sleet condition, where a decrease of 6.8% can be measured in comparison to the normal condition. In comparison to that, the performance of the MixedFusion [55] model decreased by 41.3% and the performance of EchoFusion [3] decreased by 76.3%. In general, our proposed DPFT method shows an average performance difference of −2.5% between the normal and all other conditions. In comparison, the RTNH [16], MixedFusion [55], and EchoFusion [3] models show a decrease of −3.8%, −31.3%, and −12.8%, respectively. The analysis of the average and maximum decrease suggests that models that are considering radar data are less affected by varying weather conditions than those that are not considering radar data. Ultimately, it can be shown that our proposed DPFT model shows high robustness against server weather conditions and is equally robust as the radar-only RTNH [16] method. However, the unimodal RTNH [16] model performance is significantly lower and it cannot deal with a sensor modality failure.

The robustness of our method against sensor failure is achieved by a model design without interdependencies between the different modalities. While this prevents a complete failure of the model if a single sensor modality fails during runtime, the model performance still drops significantly, as shown in Table III. To counteract this, we used the pre-trained weights of the camera and radar-only models as initialization [84], but could not observe any significant changes. Besides that, we trained the model with modality dropout [85] and were able to improve the performance for the sensor failure cases, but observed a significant decrease under nominal conditions, which is in contrast to [85], [86].

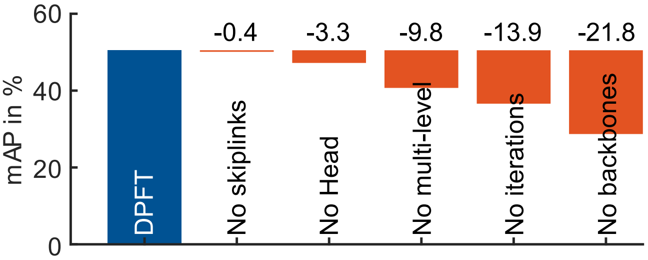


Fig. 4. Performance loss due to the ablation of individual model components on the test data of the K-Radar dataset revision v2.0.

B. Complexity

Our model is designed for real-world applications, which is why inference time and memory consumption measurements are conducted. All tests are executed on a dedicated benchmark sever equipped with an NVIDIA V100 GPU and isolated in a containerized environment. The DeepSpeed [87] framework is used for reliable and accurate measurements.

The proposed DPFT model achieves an inference time of 87 ±1 ms, which is lower than the 100 ms cycle time of the radar sensor. This is important to be able to process every sensor image and not have to drop any. In comparison, MixedFusion [55], and EchoFusion [3] have an inference time of 143 ms, and 348 ms, respectively. Therefore, our DPFT model achieves the lowest inference time among all tested methods.

The overall model complexity is mainly driven by the backbone selection, while the memory consumption is mainly caused by the input image size, as shown in Table IV. The baseline implementation of or DPFT model requires 4.0 GiB of GPU memory during inference and has a measured computational complexity of 0.16 TFLOPs. In comparison, EchoFusion [3] requires 3.5 GiB of GPU memory, but has a computational complexity of 1.52 TFLOPs, explaining its higher inference time. This comparison shows the computational efficiency of our proposed method even without any runtime optimization like TensorRT. Moreover, the modular design of our implementation allows the usage of different backbones and input image sizes. As shown in Table IV, altering these parameters can significantly decrease the computational complexity but influence the model performance. As a consequence, these parameters have to be chosen in accordance with the desired application.

C. Ablation Study

The results of the ablation study show the contribution of the individual model components on the overall detection performance and are shown in Fig. 4. It can be seen that the ablation of the backbones causes the greatest performance decreases, whereas the contribution of the skiplinks is not significant. Besides that, the iterative refinement process and the usage of multi-level feature maps have a significant effect on the detection performance. Moreover, in consideration of the conducted experiments on the input data modalities (Table II) and the analysis of different backbones (Table IV), the results suggest that the sensor fusion is the most important factor for the model performance.

TABLE IV
PERFORMANCE AND COMPLEXITY FOR DIFFERENT BACKBONES AND INPUT IMAGE RESOLUTIONS

	ResNet101	ResNet50	ResNet34	720px	512px	256px
mAP in %	50.5	49.8	47.2	50.5	50.5	45.4
Inference time in ms	87	69	64	94	87	81
Complexity in TFLOPs	0.16	0.09	0.08	0.30	0.16	0.04

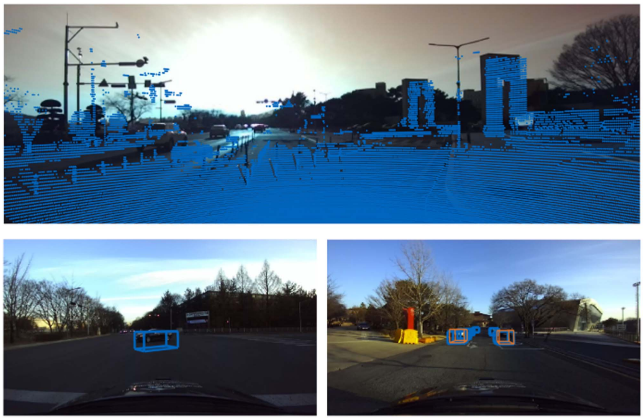


Fig. 5. Visualization of the dataset's sensor miscalibration (left) and two failure cases of the model. One shows a missing detection of a crossing object (center) and the other shows false negatives for partially occluded objects (right). The ground truth is shown in blue and the model prediction in orange.

D. Discussion

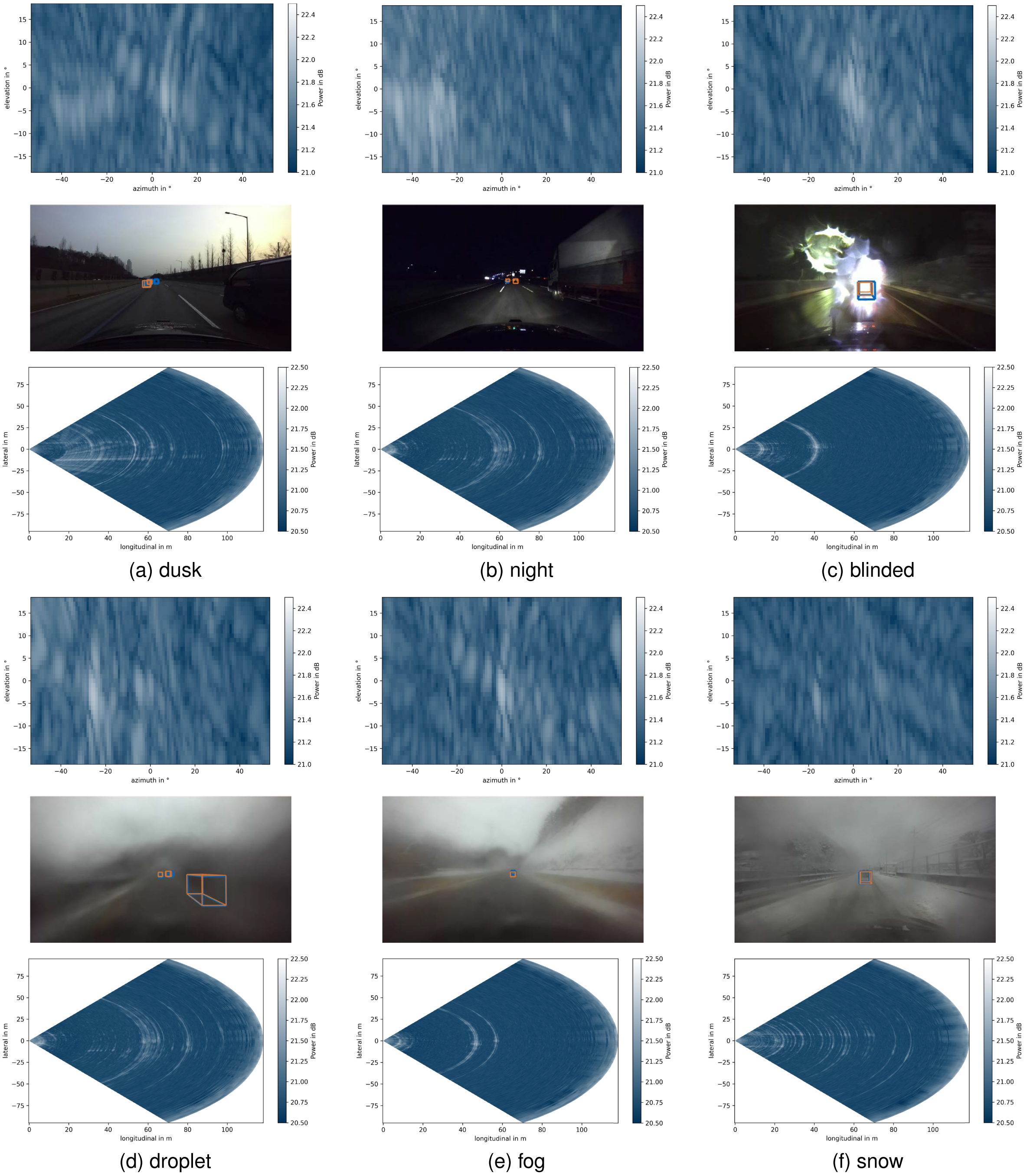
While our model achieved state-of-the-art results in the conducted experiments, there are certain limitations to it. Firstly, it outperforms lidar and camera-lidar fusion methods only in severe weather conditions while showing a significantly lower performance in normal conditions. Secondly, the model has difficulties detecting objects that are moving tangential to the ego vehicle's direction of travel and correctly predicting their heading angle, as shown in Fig. 5. This is probably caused by the fact that crossing objects are heavily underrepresented in the dataset on the one side and the inability of the radar sensor to measure tangential velocities on the other. Furthermore, it struggles to detect or differentiate between multiple objects that are behind each other or close to each other, which can be seen in Fig. 5. We believe that this is due to the partial occlusion and the limited resolution of the radar sensor in the azimuth-elevation plane. Last but not least, the generalization capability of the model could only be tested within the scope of the K-Radar [16] dataset. Since the K-Radar [16] dataset is the only dataset that provides raw 4D radar data for different weather conditions and the only large-scale dataset with radar cube data in general, the transferability of the model to different datasets is yet to be shown. Nevertheless, comparable model architectures [3] that only rely on 3D radar data show promising generalization results, which is a first indicator for the transferability of these model types.

Despite being the only dataset with 4D radar cube data, the K-Radar [16] dataset shows some labeling inconsistencies (especially between the sedan and bus or truck classes) even within the revision v2.0. In addition, the test set is sampled from the same driving sequences and contains similar scenarios to the train set, which limits the ability to test the generalizability of models, even if the test split is formally independent. Furthermore, we observed a misalignment between the camera and lidar frame, as shown in Fig. 5, which is important because the labels are created on the lidar data, and which is why EchoFusion [3] used their own calibration. However, a recalibration of the sensors is difficult and would limit the comparability to previous methods, which is why we used the official calibration. Nevertheless, further investigations would be needed to quantify the model's sensitivity to miscalibrations. Last but not least, the calculation of the total mAP metric in the official evaluation scheme could be misleading since it is calculated as the weighted average of the individual categories weighted by the number of ground truth objects. In general, the usage of the KITTI [79] evaluation protocol could be questioned due to the problem of average precision distortion [88] and since recent studies show that other metrics, like the nuScenes detection score (NDS), correlate better with the fulfillment of the autonomous driving task [89].

V. CONCLUSION

We proposed a novel method to fuse camera and cube-level radar data to achieve a performant, robust, yet cost-effective method for 3D object detection. We are the first to fuse raw 4D radar data with camera data and demonstrate the importance of the different input perspectives. Our proposed DPFT method achieves state-of-the-art results in the challenging environmental conditions of the K-Radar dataset. Experimental results show that our proposed method is robust against severe weather conditions and is able to maintain general functionality even after a sensor failure. Finally, we provided a comprehensive analysis of the computational complexity of our method and were able to show that our method has the fastest inference time among all tested fusion methods.

Despite the great potential of camera and radar fusion for 3D object detection, new research questions emerge from this work. While we proposed a novel dual perspective fusion approach, the general question of how to utilize the high dimensional radar data most efficiently remains open for research. Moreover, balancing the performance of different sensor modalities within a fusion method to exploit the input data most effectively and avoid significant performance losses during the event of a sensor failure remains challenging. Even if we used different methods to counteract the performance degradation after a sensor failure, further research is needed to mitigate this effect. Moreover, sensor-specific challenges like target separation in the radar domain or depth estimation in the camera domain remain open for



(a) dusk
(b) night
(c) blinded
(d) droplet
(e) fog
(f) snow

Fig. 6. Exemplary results of the model performance under night, rain, and snow conditions. The camera data is shown in the center, the radar range-azimuth (RA) data at the bottom, and the radar azimuth-elevation (AE) data at the top. The ground truth is shown in blue and the model prediction in orange.

research. Beyond that, temporal information could be considered to increase the performance and a different detection head could be used to realize an instance-free detection method in future work.

Appendix A
Additional Result Details

The appendix presents additional details on the results on the K-Radar [16] dataset. Following Section IV, the results of Table V were obtained on the original version (revision v1.0) of the dataset, while all other additional results are based on the revised version (revision v2.0) of the dataset.

The results of Table V show decreasing mAP values with increasing IoU thresholds for all tested methods and both the 3D and the BEV object detection tasks. It is worth mentioning that the results of the DPFT method, listed in Table V, are the mean values of three independent model trainings to mitigate the effects of any non-deterministic training behavior. It can be seen that our proposed method outperforms all previous radar- and camera-radar-based methods for 3D object detection at all IoU thresholds and performs on par with the RTNH [16] method for

TABLE V
OBJECT DETECTION RESULTS FOR THE K-RADAR TEST SET REVISION V1.0

Method	Modality	3D mAP AP@0.3	3D mAP AP@0.5	3D mAP AP@0.7	BEV mAP AP@0.3	BEV mAP AP@0.5	BEV mAP AP@0.7
RTNH [16]	R	47.4	15.6	0.5	**58.4**	43.2	11.5
EchoFusion [3]	C + R	47.4	28.1	6.4	48.9	39.7	25.7
DPFT (ours)	C + R	**56.1**	**37.0**	**8.0**	57.5	**48.5**	**26.3**

TABLE VI
3D OBJECT DETECTION RESULTS FOR DIFFERENT DETECTION RANGES

Modality	Total	0 - 10 m	10 - 30 m	30 - 50 m	50 - 72 m
C	8.9	27.3	15.5	4.7	3.4
R	36.2	35.5	42.7	37.1	25.2
C + R	50.5	44.8	54.6	53.4	35.3

TABLE VII
3D OBJECT DETECTION RESULTS FOR DIFFERENT DAYTIMES

Modality	Day	Night	Total
C	9.8	3.0	8.9
R	36.9	29.1	36.2
C + R	52.7	39.8	50.5

BEV detections at a low IoU threshold. However, our proposed method archives higher mAP values for BEV detection at higher IoU thresholds.

The experimental results of Table VI show the performance of our DPFT model for different sensor modalities and detection range bins. It can be seen that the general performance of the model decreases with increasing range. This is especially true for the camera-only model, which shows a significant performance decrease with increasing detection range. The observed behavior is probably caused by the inability of the camera sensor to measure depth information and its decreasing spatial resolution with increasing distance. In contrast, the radar-only model shows a lower performance for the range between 0–10 m and achieves the highest performance in a range between 10–30 m, with a decreasing performance over increasing distance. This phenomenon is probably caused by the higher noise level of the radar in close range and the decreasing spatial resolution with increasing distance. The performance of the camera-radar fusion model shows a similar behavior to the radar-based model, but a higher performance overall and seems to be less affected by increasing distance. We believe that this is a result of the already discussed sensor properties and the distribution of objects in the dataset that contains the most objects in a range of 20–40 m and the least for distances greater than 60 m [16].

APPENDIX B
ADDITIONAL DETAILS ON ROBUSTNESS

In addition to the differentiation into different weather conditions, the K-Radar dataset allows the separate determination of the performance values for day and night conditions. The results of Table VII show that all configurations of the DPFT model perform better under daytime conditions than nighttime conditions. Nevertheless, the performance of the camera-only model is affected the most, while the radar-only model shows the smallest decrease of all tested configurations. This is probably because camera sensors are dependent on ambient light, while radar sensors are active sensors and, therefore, independent from external sources. However, the general tendency could also be explained by the data distribution of the K-Radar dataset, which consists of 63% daytime scenes, which results in an imbalanced training and test set [16].

The analysis of individual models shows that the camera-based model fails if the camera lens is covered by raindrops or sleet (as shown in 6), which only gets worse in night-time conditions. However, these problems cloud be avoided by a different camera positioning or cleaning mechanism. The radar-based performance seems to be less affected by environmental conditions but more dependent on the number of available training samples. Nevertheless, target separation remains challenging in dense traffic or city scenarios.

APPENDIX C
ADDITIONAL DETAILS ON COMPLEXITIY

In this section, we provide more detailed results on the model complexity analysis discussed in Section IV. The appended Table VIII is an extension of Table IV and includes additional metrics on the computational complexity of the different model configurations as well as the memory requirements based on the model parameters. In general, it provides evidence for the claim that a larger backbone size and higher input resolution lead to a higher model performance but an increased computational complexity.

In addition, the model has been tested with different numbers of query points to analyze the effects of different query point resolutions on the model performance and computational complexity. The results for 100, 400, and 900 query points show that an increased query point resolution leads only to a marginal increase in computational complexity of 156, 156, and 157 GFLOPs, but a larger impact on the memory consumption. In contrast, the best model performance seems to be achieved with 400 query points, whereas a query point resolution of 100 and 900 leads to a result of 47.8% and 44.5% mAP, respectively. During model development, quadratic and exponentially distributed query point initializations in both cartesian and polar coordinates as well as a learnable query point initialization were also tested with no significant performance increases. Besides that, Table IX provides inference time measurements on different hardware accelerators, using the same method as described in Section IV and demonstrates that significantly lower inference times can be achieved on more modern GPUs.

TABLE VIII
PERFORMANCE AND COMPLEXITY FOR DIFFERENT BACKBONES AND INPUT IMAGE RESOLUTIONS

	ResNet101	ResNet50	ResNet34	720px	512px	256px
mAP in %	50.5	49.8	47.2	50.5	50.5	45.4
Time in ms	87	69	64	94	87	81
FLOPs in 10^9	156	86	75	302	156	44
MACs in 10^9	78	43	37	150	78	22
Parameters in 10^6	90	66	44	90	90	90

TABLE IX
INFERENCE TIME ON DIFFERENT NVIDIA GPU UNITS

	3090	4090	V100	A40	A100
Time in ms	74±1.2	32±0.4	87±1.2	52±1.0	41±0.1

APPENDIX D
EXAMPLES

Fig. 6 shows the model predictions and ground truth data plotted onto the camera images and the associated radar data in the range-azimuth (RA) and azimuth-elevation (AE) planes under different environmental conditions.

REFERENCES

[1] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, “A survey on 3D object detection methods for autonomous driving applications,” *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3782–3795, Oct. 2019.

[2] K. Yoneda, N. Suganuma, R. Yanase, and M. Aldibaja, “Automated driving recognition technologies for adverse weather conditions,” *IATSS Res.*, vol. 43, no. 4, pp. 253–262, Dec. 2019.

[3] Y. Liu, F. Wang, N. Wang, and Z. Zhang, “Echoes beyond points: Unleashing the power of raw radar data in multi-modality fusion,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 53964–53982.

[4] S. Yao et al., “Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review,” *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 2094–2128, Jan. 2024.

[5] Z. Liu et al., “BEVFusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation,” in *Proc. 2023 IEEE Int. Conf. Robot. Automat.*, 2023, pp. 2774–2781.

[6] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable transformers for end-to-end object detection,” 2020.

[7] H. Caesar et al., “nuScenes: A multimodal dataset for autonomous driving,” in *Proc. 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11618–11628.

[8] F. Engels, P. Heidenreich, M. Wintermantel, L. Stacker, M. Al Kadi, and A. M. Zoubir, “Automotive radar signal processing: Research directions and practical challenges,” *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 4, pp. 865–878, Jun. 2021.

[9] O. Schumann et al., “RadarScenes: A real-world radar point cloud data set for automotive applications,” in *Proc. 2021 IEEE 24th Int. Conf. Inf. Fusion*, 2021, pp. 1–8.

[10] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila, “Multi-class road user detection with 3 1D radar in the view-of-delft dataset,” *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 4961–4968, Apr. 2022.

[11] L. Zheng et al., “TJ4DRadSet: A 4D radar dataset for autonomous driving,” in *Proc. 2022 IEEE 25th Int. Conf. Intell. Transp. Syst.*, 2022, pp. 493–498.

[12] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Perez, “CARRADA dataset: Camera and automotive radar with range- angle- doppler annotations,” in *Proc. IEEE 2020 25th Int. Conf. Pattern Recognit.*, 2021, pp. 5068–5075.

[13] M. Sheeny, E. De Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, “RADIATE: A radar dataset for automotive perception in bad weather,” in *Proc. 2021 IEEE Int. Conf. Robot. Automat.*, 2021, pp. 1–7.

[14] Y. Wang, G. Wang, H.-M. Hsu, H. Liu, and J.-N. Hwang, “Rethinking of radar’s role: A camera-radar dataset and systematic annotator via coordinate alignment,” in *Proc. 2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 2809–2818.

[15] J. Rebut, A. Ouaknine, W. Malik, and P. Perez, “Raw high-definition radar for multi-task learning,” in *Proc. 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17000–17009.

[16] D.-H. Paek, S.-H. KONG, and K. T. Wijaya, “K-Radar: 4D radar object detection for autonomous driving in various weather conditions,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 3819–3829.

[17] X. Ma, W. Ouyang, A. Simonelli, and E. Ricci, “3D object detection from images for autonomous driving: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3537–3556, May 2024.

[18] R. Qian, X. Lai, and X. Li, “3D object detection for autonomous driving: A survey,” *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108796.

[19] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, “Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving,” in *Proc. 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8437–8445.

[20] S. Srivastava, F. Jurie, and G. Sharma, “Learning 2D to 3D lifting for object detection in 3D for autonomous vehicles,” in *Proc. 2019 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 4504–4511.

[21] J. Philion and S. Fidler, “Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D,” in *Computer Vision*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Berlin, Germany: Springer, 2020, pp. 194–210.

[22] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, “Simple-BEV: What really matters for multi-sensor BEV perception?,” in *Proc. 2023 IEEE Int. Conf. Robot. Automat.*, 2023, pp. 2759–2765.

[23] G. Brazil and X. Liu, “M3D-RPN: Monocular 3D region proposal network for object detection,” in *Proc. 2019 IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9287–9296.

[24] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” 2019, *arXiv:1904.07850*.

[25] X. Chen et al., “3D object proposals for accurate object class detection,” in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 424–432.

[26] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, “Monocular 3D object detection for autonomous driving,” in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2147–2156.

[27] A. Naiden, V. Paunescu, G. Kim, B. Jeon, and M. Leordeanu, “Shift R-CNN: Deep monocular 3D object detection with closed-form geometric constraints,” in *Proc. 2019 IEEE Int. Conf. Image Process.*, 2019, pp. 61–65.

[28] B. Yang, I. Khatri, M. Happold, and C. Chen, “ADCNet: Learning from raw radar data via distillation,” 2023, *arXiv:2303.11420*.

[29] Y. Wang, Z. Jiang, Y. Li, J.-N. Hwang, G. Xing, and H. Liu, “RODNet: A real-time radar object detection network cross-supervised by camera-radar fused object 3D localization,” *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 4, pp. 954–967, Jun. 2021.

[30] X. Dong, P. Wang, P. Zhang, and L. Liu, “Probabilistic oriented object detection in automotive radar,” in *Proc. 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 458–467.

[31] P. Kaul, D. de Martini, M. Gadd, and P. Newman, “RSS-Net: Weakly-supervised multi-class semantic segmentation with FMCW radar,” in *Proc. 2020 IEEE Intell. Veh. Symp. (IV)*, 2020, pp. 431–436.

[32] W. Ng, G. Wang, Z. Siddhartha, Lin, and B. J. Dutta, “Range-doppler detection in automotive radar with deep learning,” in *Proc. IEEE 2020 Int. Joint Conf. Neural Netw.*, 2020, pp. 1–8.

[33] C. Decourt, R. VanRullen, D. Salle, and T. Oberlin, “DAROD: A deep automotive radar object detector on range-doppler maps,” in *Proc. 2022 IEEE Intell. Veh. Symp. (IV)*, 2022, pp. 112–118.

[34] B. Major et al., “Vehicle detection with automotive radar using deep learning on range-Azimuth-Doppler tensors,” in *Proc. 2019 IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 924–932.

[35] X. Gao, G. Xing, S. Roy, and H. Liu, “RAMP-CNN: A novel neural network for enhanced automotive radar object recognition,” *IEEE Sensors J.*, vol. 21, no. 4, pp. 5119–5132, Feb. 2021.

[36] A. Palffy, J. Dong, J. F. P. Kooij, and D. M. Gavrila, “CNN based road user detection using the 3D radar cube,” *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1263–1270, Apr. 2020.

[37] A. Zhang, F. E. Nowruzi, and R. Laganiere, “RADDet: Range-azimuth-doppler based radar object detection for dynamic road users,” in *Proc. 2021 18th Conf. Robots Vis.*, 2021, pp. 95–102.

[38] B. Tan et al., “3-D object detection for multiframe 4-D automotive millimeter-wave radar point cloud,” *IEEE Sensors J.*, vol. 23, no. 11, pp. 11125–11138, Jun. 2023.

[39] M. Dreher, E. Ercelik, T. Banziger, and A. Knoll, “Radar-based 2D car detection using deep neural networks,” in *Proc. 2020 IEEE 23 rd Int. Conf. Intell. Transp. Syst.*, 2020, pp. 1–8.

[40] D. Köhler, M. Quach, M. Ulrich, F. Meinl, B. Bischoff, and H. Blume, “Improved multi-scale grid rendering of point clouds for radar object detection networks,” in *Proc. 26th Int. Conf. Inf. Fusion*, 2023, pp. 1–8.

[41] J. Liu, Q. Zhao, W. Xiong, T. Huang, Q.-L. Han, and B. Zhu, “SMURF: Spatial multi-representation fusion for 3D object detection with 4D imaging radar,” *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 799–812, Jan. 2024.

[42] M. Ulrich et al., “Improved orientation estimation and detection with hybrid object detection networks for automotive radar,” in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst.*, 2022, pp. 111–117.

[43] P. Svenningsson, F. Fioranelli, and A. Yarovoy, “Radar-PointGNN: Graph based object recognition for unstructured radar point-cloud data,” in *Proc. 2021 IEEE Radar Conf.*, 2021, pp. 1–6.

[44] F. Fent, P. Bauerschmidt, and M. Lienkamp, “RadarGNN: Transformation invariant graph neural network for radar-based perception,” in *Proc. 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2023, pp. 182–191.

[45] J. F. Tilly et al., “Detection and tracking on automotive radar data with deep learning,” in *Proc. IEEE 23rd Int. Conf. Inf. Fusion*, 2020, pp. 1–7.

[46] N. Scheiner et al., “Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar,” in *Proc. 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2065–2074.

[47] A. Dubey, A. Santra, J. Fuchs, M. Lübke, R. Weigel, and F. Lurz, “HARadNet: Anchor-free target detection for radar point clouds using hierarchical attention and multi-task learning,” *Mach. Learn. Appl.*, vol. 8, Jun. 2022, Art. no. 100275.

[48] A. Danzer, T. Griebel, M. Bach, and K. Dietmayer, “2D car detection in radar data with PointNets,” in *Proc. 2019 IEEE Intell. Transp. Syst. Conf.*, 2019, pp. 61–66.

[49] O. Schumann, J. Lombacher, M. Hahn, C. Wohler, and J. Dickmann, “Scene understanding with automotive radar,” *IEEE Trans. Intell. Veh.*, vol. 5, no. 2, pp. 188–203, Jun. 2020.

[50] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, “A deep learning-based radar and camera sensor fusion architecture for object detection,” in *Proc. 2019 Sensor Data Fusion: Trends, Solutions, Appl.*, 2019, pp. 1–7.

[51] K. Bansal, K. Rungta, and D. Bharadia, “RadSegNet: A reliable approach to radar camera fusion,” 2022, *arXiv:2208.03849*.

[52] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, “PointPainting: Sequential fusion for 3D object detection,” in *Proc. 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4603–4611.

[53] H. Jha, V. Lodhi, and D. Chakravarty, “Object detection and identification using vision and radar data fusion system for ground-based navigation,” in *Proc. IEEE 6th Int. Conf. Signal Process. Integr. Netw.*, 2019, pp. 590–593.

[54] X. Dong, B. Zhuang, Y. Mao, and L. Liu, “Radar camera fusion via representation learning in autonomous driving,” in *Proc. 2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 1672–1681.

[55] C. Zhang, H. Wang, L. Chen, Y. Li, and Y. Cai, “MixedFusion: An efficient multimodal data fusion framework for 3-D object detection and tracking,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2024, doi: 10.1109/TNNLS.2023.3325527.

[56] R. Nabati and H. Qi, “CenterFusion: Center-based radar and camera fusion for 3D object detection,” in *Proc. 2021 IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1526–1535.

[57] H. Cui, J. Wu, J. Zhang, G. Chowdhary, and W. R. Norris, “3D detection and tracking for on-road vehicles with a monovision camera and dual low-cost 4D mmWave radars,” in *Proc. 2021 IEEE Int. Intell. Transp. Syst. Conf.*, 2021, pp. 2931–2937.

[58] J. Kim, Y. Kim, and D. Kum, “Low-level sensor fusion network for 3D vehicle detection using radar range-azimuth heatmap and monocular image,” in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 388–402.

[59] S. Chang et al., “Spatial attention fusion for obstacle detection using MmWave radar and vision sensor,” *Sensors*, vol. 20, no. 4, Feb. 2020, Art. no. 956.

[60] R. Yadav, A. Vierling, and K. Berns, “Radar RGB fusion for robust object detection in autonomous vehicle,” in *Proc. 2020 IEEE Int. Conf. Image Process.*, 2020, pp. 1986–1990.

[61] T. Zhou, J. Chen, Y. Shi, K. Jiang, M. Yang, and D. Yang, “Bridging the view disparity between radar and camera features for multi-modal fusion 3D object detection,” *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1523–1535, Feb. 2023.

[62] W. Xiong, J. Liu, T. Huang, Q.-L. Han, Y. Xia, and B. Zhu, “LXL: LiDAR excluded lean 3D object detection with 4D imaging radar and camera fusion,” *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 79–92, Jan. 2024.

[63] L. Zheng et al., “RCFusion: Fusing 4-D radar and camera with bird's-eye view features for 3-D object detection,” *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 8503814.

[64] T.-Y. Lim et al., “Radar and camera early fusion for vehicle detection in advanced driver assistance systems,” in *Proc. Mach. Learn. Auton. Driving Workshop, 33rd Conf. Neural Inf. Process. Syst.*, 2019.

[65] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, “Tri-perspective view for vision-based 3D semantic occupancy prediction,” in *Proc. 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9223–9232.

[66] O. Schumann, C. Wöhler, M. Hahn, and J. Dickmann, “Comparison of random forest and long short-term memory network performances in classification tasks using radar,” in *Proc. Sensor Data Fusion: Trends, Sol., Appl.*, 2017, pp. 1–6.

[67] N. Scheiner, N. Appenrodt, J. Dickmann, and B. Sick, “Radar-based feature design and multiclass classification for road user recognition,” in *Proc. 2018 IEEE Intell. Veh. Symp. (IV)*, 2018, pp. 779–786.

[68] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–786.

[69] L. Liu et al., “Deep learning for generic object detection: A survey,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Oct. 2019.

[70] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[71] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[72] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, “DETR3D: 3D object detection from multi-view images via 3D-to-2D queries,” in *Proc. 5th Conf. Robot Learn.*, 2022, pp. 180–191.

[73] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, “FUTR3D: A Unified Sensor Fusion Framework for 3D Detection,” in *Proc. 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2023, pp. 172–181.

[74] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, “Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit,” *Nature*, vol. 405, no. 6789, pp. 947–951, Jun. 2000.

[75] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, “On the continuity of rotation representations in neural networks,” in *Proc. 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5738–5746.

[76] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, *End-to-End Object Detection With Transformers*. Berlin, Germany: Springer, 2020, pp. 213–229.

[77] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[78] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2017, *arXiv:1711.05101*.

[79] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in *Proc. 2012 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[80] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, “Voxel R-CNN: Towards high performance voxel-based 3D object detection,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1201–1209.

[81] H. Wu, J. Deng, C. Wen, X. Li, C. Wang, and J. Li, “CasA: A cascade attention network for 3-D object detection from LiDAR point clouds,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5704511.

[82] H. Wu, C. Wen, W. Li, X. Li, R. Yang, and C. Wang, “Transformation-equivariant 3D object detection for autonomous driving,” in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 2795–2802.

[83] H. Zhu et al., “VPFNet: Improving 3D object detection with virtual point based LiDAR and stereo data fusion,” *IEEE Trans. Multimedia*, vol. 25, pp. 5291–5304, 2023.

[84] T. Liang et al., “BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 10421–10434.

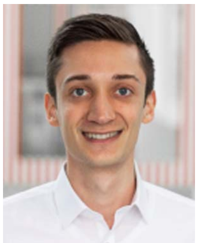
[85] C. Ge et al., “MetaBEV: Solving sensor failures for 3D detection and map segmentation,” in *Proc. 2023 IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8721–8731.

[86] S. Wang, H. Caesar, L. Nan, and J. F. P. Kooij, “UniBEV: Multi-modal 3D object detection with uniform BEV encoders for robustness against missing sensor modalities,” in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2024, pp. 2776–2783.

[87] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, “DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters,” in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 3505–3506.

[88] H. Zhang, A. Rogozan, and A. Bensrhair, “An enhanced n-point interpolation method to eliminate average precision distortion,” *Pattern Recognit. Lett.*, vol. 158, pp. 111–116, Jun. 2022.

[89] T. Schreier, K. Renz, A. Geiger, and K. Chitta, “On offline evaluation of 3D object detection for autonomous driving,” in *Proc. 2023 IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2023, pp. 4086–4091.



Felix Fent received the B.Sc. and M.Sc. degrees in 2018 and 2020, respectively, from the Technical University of Munich (TUM), Munich, Germany, where he is currently working toward the Ph.D. degree in mechanical engineering with the Institute of Automotive Technology. His research interests include radar-based perception and sensor fusion and multimodal object detection approaches with a focus on real-world applications.

Andras Palffy (Member, IEEE) received the M.Sc. degree in computer science engineering from Pazmany Peter Catholic University, Budapest, Hungary, in 2016, the M.Sc. degree in digital signal and image processing from Cranfield University, Cranfield, U.K., in 2015, and the Ph.D. degree from the Delft University of Technology, Delft, Netherlands, in 2022, focusing on radar based vulnerable road user detection for automated driving. From 2013 to 2017, he was an algorithm Researcher with Eutecus, a US based startup developing computer vision algorithms for traffic monitoring and driver assistance applications. In 2022, he Co-founded Perciv AI, a machine perception startup developing AI-driven, next generation machine perception for radars.

Holger Caesar received the Ph.D. degree in computer vision from the University of Edinburgh, Edinburgh, U.K., and also studied in Germany and Switzerland (KIT Karlsruhe, Karlsruhe, Germany, EPF Lausanne, Lausanne, Switzerland, and ETH Zurich, Zurich, Switzerland). He was a Principal Research Scientist with Motional, where he started three teams with more than 20 members that focused on data annotation, autolabeling and data mining. He is currently an Assistant Professor with Intelligent Vehicles Group, TU Delft, Netherlands. His research interests include the area of autonomous vehicle perception and prediction, with a particular focus on scalability of learning and annotation approaches. He is best known for developing the influential autonomous driving datasets nuScenes and nuPlan, and his contributions to the real-time 3D object detection method PointPillars.