DinoRADE: Full Spectral Radar-Camera Fusion with Vision Foundation Model Features for Multi-class Object Detection in Adverse Weather

Christof Leitgeb[1,2] Thomas Puchleitner[1] Max Peter Ronecker[2,3] Daniel Watzenig[2,3]
[1]Infineon Technologies AG, Austria [2]Graz University of Technology, Austria
[3]Virtual Vehicle Research GmbH, Austria
{christof.leitgeb, thomas.puchleitner}@infineon.com
max.ronecker@v2c2.at daniel.watzenig@tugraz.at

Abstract

Reliable and weather-robust perception systems are essential for safe autonomous driving and typically employ multi-modal sensor configurations to achieve comprehensive environmental awareness. While recent automotive FMCW Radar-based approaches achieved remarkable performance on detection tasks in adverse weather conditions, they exhibited limitations in resolving fine-grained spatial details particularly critical for detecting smaller and vulnerable road users (VRUs). Furthermore, existing research has not adequately addressed VRU detection in adverse weather datasets such as K-Radar. We present DinoRADE, a Radar-centered detection pipeline that processes dense Radar tensors and aggregates vision features around transformed reference points in the camera perspective via deformable cross-attention. Vision features are provided by a DINOv3 Vision Foundation Model. We present a comprehensive performance evaluation on the K-Radar dataset in all weather conditions and are among the first to report detection performance individually for five object classes. Additionally, we compare our method with existing single-class detection approaches and outperform recent Radar-camera approaches by 12.1%. The code is available under https://github.com/chr-is-tof/RADE-Net.

1. Introduction

Safe autonomous driving requires robust and reliable perception systems to interact with other road users and navigate complex scenarios, often in adverse weather conditions like fog, rain, and snow, which additionally disturb the view of optical sensors [4]. To tackle these challenges, automotive systems must rely on a variety of sensor technologies to guarantee reliable perception. Cameras provide high structural details which contribute to reliable object classification, but they lack inherent depth information which can lead to depth estimation errors, especially for textureless regions, reflective surfaces, varying illumination conditions or perspective illusions [35]. LiDAR sensors combine high structural information with accurate depth information but rely on expensive hardware and are limited in adverse weather conditions, where atmospheric particles like smoke, fog, rain, and snow block their relatively short wavelength signal [4, 16]. In contrast, Radar sensors emit a signal with significantly longer wavelength, which results in a constant perception performance across various adverse weather conditions [5, 17]. However, while providing relatively accurate distance information, the large antenna array structures result in low angular resolution, which makes precise object separation challenging, especially for small objects like pedestrians and cyclists [47]. Another significant benefit of Radar sensors is the direct measurement of the relative radial velocity via the Doppler effect. This information can be employed for tasks like object separation and motion classification [47]. Additionally, the full Doppler spectrogram contains characteristic micro-Doppler signatures caused by micro-motions, such as rotation and vibration from object parts, which are unique for different types of road users and motions [47]. These signatures contain rich information for human and vehicle motion classification, which subsequently benefits close object separation [34, 47].

Given the complementary strengths and weaknesses of different automotive sensors, it becomes evident that a combination of modalities will benefit robust perception. The fusion of Radar and camera combines fine structural details with reliable performance in adverse weather conditions, direct position, and velocity measurements as well as low costs [4]. Consequently, Radar-camera approaches have been widely explored for object detection and segmentation tasks and reported results demonstrate superior performance [5, 18, 20, 33, 43]. However, while visual images from a camera are fairly compact, raw Radar data con-

sists of sampled time-series data which contains frequency and phase information about objects in the scene and comes with large data sizes for each single frame [47]. To obtain a portable format, they are usually Fast-Fourier transformed (FFT) and reduced by adaptive thresholding techniques, which results in a 3-dimensional Radar point cloud with additional Doppler values. However, this process results in a significant loss of structural detail, yielding only few data points per object, especially for smaller classes like pedestrians and cyclists. To counteract data sparsity, recent approaches incorporate denser Radar information extracted from lower-level processing stages. Examples include raw time-series analog-to-digital (ADC) converted data [8, 44], sparse FFT-processed range-azimuth-Doppler-elevation (RADE) tensors [15, 28] as well as 2D and 3D projections from RADE tensors [5, 10, 17, 30].

Conventional vision-based perception pipelines utilize task-specific architectures, such as ResNet [11] and YOLO [32], which are often initialized with pre-trained weights or trained from scratch to accommodate domain-specific requirements. However, the advancement of Vision Foundation Models (VFMs) [26, 36] has introduced a paradigm shift by providing general-purpose pre-trained feature extractors that demonstrate superior generalization capabilities. These models can extract semantically meaningful representations from visual scenes, including rare scenarios potentially absent from task-specific training datasets. Consequently, VFMs have demonstrated significant efficacy in automotive perception systems, offering enhanced robustness and adaptability compared to conventional approaches.

Many state-of-the-art Radar-camera or Radar-only approaches are evaluated on datasets without severe weather conditions like fog or snow [1, 46], thereby limiting the full potential of Radar sensing. Conversely, approaches evaluated on adverse weather datasets like K-Radar [28] predominantly focus on vehicle detection [5, 15, 21, 28], thus missing the effect of high resolution visual details that might be crucial for VRU detection and classification. To the best of our knowledge, we are the first to combine dense Radar representations derived from 3D FFT tensor projections with high-resolution semantic features extracted from Vision Foundation Models, while providing comprehensive evaluations across diverse weather conditions and multiple object classes including VRUs. This enables full exploitation of the complementary strengths of both sensing modalities. Specifically, our contributions are as follows:

- A weighted query lifting method to refine Radar queries based on their true distribution in the range-azimuth-elevation projection and aggregate inter-perspective features from synchronized DINOv3 [36] VFM representations.
- Adaptive fusion strategy to selectively refine features from RADE-Net [17] Radar backbone to increase detection performance with special focus on smaller object classes.
- Comprehensive performance evaluation in adverse weather conditions and individually for 5 object classes of the K-Radar [28] dataset.

2. Background and Related Works

2.1. Automotive Camera-Radar Datasets

Automotive datasets, which provide camera images and Radar data, have strong variations in size, class categories, use of sensor modalities, data representation, environment scenarios, and annotation quality [46]. While camera data is almost exclusively represented in RGB format, with the exception of RaDICal [19] providing RGB-depth data, Radar experiences strong variations in data representation. RaDICal and RADIal [31] provide raw time-series data. CARRADA [27], UWCR [6], and RADDet [48] offer processed range-azimuth-Doppler (RAD) tensors. K-Radar [28] provides higher resolution range-azimuth-Doppler-elevation (RADE) tensors. Furthermore, very prominent automotive datasets like nuScenes [1], View-of-Delft [29], and Astyx [24] only provide sparse Radar point clouds, but instead include up to 23 different object classes along with high quality 3D bounding box annotations. Most time-series and FFT tensor-based datasets lack annotation quality with the exception of K-Radar, which provides 3D rotated bounding boxes for 7 object classes. To our knowledge, K-Radar [28] is the only large-scale Radar-camera dataset capturing 7 weather conditions, 8 road types, and day/night variations, making it suitable for exploring the complementary advantages of Radar and camera across diverse scenarios.

2.2. Radar-only Detection

Radar-only object detection and segmentation approaches can be grouped in three primary categories based on their underlying data representation [47]: Raw time-series ADC data [8, 44], FFT-processed spectral data [15, 17, 28, 30], and sparse point cloud data [25, 29]. ADCNet [44] uses a distillation method to learn RAD tensors from raw ADC data and optimize the network to predict the range-azimuth coordinate of objects in the scene. Similarly, T-FFTRadNet [8] learns the transformation from ADC to FFT representation and utilizes hierarchical Swin Vision transformers on the patched FFT data. FFT-RadNet [30] reduces memory requirements by recovering angular information from the range-Doppler (RD) spectrum. RADE-Net [17] creates 3D projections from RADE tensors which preserve rich Doppler and elevation features, thus achieving superior performance. RTNH [28] and RTNH+ [15] employ a sparse RADE tensor by reducing the full tensor to the top 10% of

power measurements. RadarPillars [25] uses a pillar-based approach similar to LiDAR processing and introduces self-attention as well as radial velocity encoding to handle sparsity of Radar data, thus surpassing the Radar baseline on the VoD dataset [29].

2.3. Camera-only Detection

3D object detection from monocular RGB-cameras presents a challenge due to the lack of inherent depth information, which requires network architectures to recover depth information from perspective features in the image [14]. Image-based 3D detectors can be categorized based on 2D or 3D features and further grouped into result lifting, feature lifting, or data lifting methods [23]. First, result lifting methods use 2D features to estimate the location of objects in the image plane and then lift them into 3D. Therefore, they are similar to classical 2D detectors, which include region-based methods like R-CNN [9] and single-shot methods like CenterNet [50]. Second, feature lifting methods generate 3D features from lifted 2D features and predict the result in 3D. Examples like BEVDet [12] collapse the 3D features to BEV before predicting the final result. Last, data lifting methods lift the input data from 2D to 3D and generate the results directly. This concept is widely associated with pseudo-LiDAR approaches [40] where image-based depth maps are converted to pseudo-LiDAR representations, which allows to apply existing LiDAR-based detection algorithms [23, 40].

2.4. Vision Foundation Models

Recent advancements in foundation models have demonstrated remarkable performance in a variety of computer vision and autonomous driving applications, including object detection, classification, and segmentation [2, 22, 26, 36]. VFMs are large scale models pre-trained on massive, diverse datasets which enables them to capture complex patterns and features from unseen images and provide a reusable representation across various tasks and domains [22].

2.5. Radar-Camera Fusion

The fusion of Radar and camera combines the most complementary features of both modalities which makes it a popular research direction for perception in adverse weather [4]. DPFT [5] creates multi-view projections from Radar FFT data in range-azimuth and azimuth-elevation view. Both Radar and camera are processed by ResNet [11] backbones, view-transformed and fused using deformable attention [52]. BEVCar [33] encodes surround-view images using a DINOv2 [26] pretrained VFM in combination with a vision transformer adapter [2]. Radar is encoded from point-cloud format followed by deformable self- and cross-attention [52] where the cross-attention takes lifted 3D queries from the VFM output. WRCFormer [10] utilizes multi-view Radar encoders in range-azimuth and elevation-azimuth perspectives which are subsequently decomposed by a Wavelet-transform based attention module and adaptively fused.

3. Methodology

Our architecture in Fig. 1 is designed to extract 3D positional and spectral Doppler features from Radar 3D projections and aggregate structural details from inter-perspective VFM features. The RAD and RAE projections are processed utilizing a modified RADE-Net backbone from [17], yielding a BEV feature map that preserves the inherent range-azimuth coordinates of Radar data while Doppler and elevation information is encoded into the 128 feature channels. Subsequently, the BEV features are lifted to receive a 3D query representation, which is further refined through elevation weights, learned from the spectral distribution of the RAE projection. The query positions provide 3D information while removing the need for computationally expensive 3D CNNs during feature extraction [38].

The camera images are processed utilizing a pretrained DINOv3 ViT-S/16 [36], which extracts generalized image features that are subsequently up-sampled, thus resulting in a perspective-view (PV) feature map. To aggregate features in the PV map with query information from the BEV map, we utilize deformable attention [52]. With known transformation matrices, each 3D query position can be mapped to a 2D location in the PV feature map. To exploit this relationship, we transform a set of reference points, visualized in Fig. 2, for all queries from Radar to camera PV coordinates. Furthermore, an offset prediction network predicts offsets for each 3D query to learn where to aggregate features in the PV feature map as described in [52]. The aggregated features are utilized to update the BEV Feature map with refined visual information. Camera images are severely affected by adverse weather conditions, resulting in occlusions, noise and obstructed views. This leads to the PV feature map lacking informative content for the BEV queries to aggregate from. To address this limitation, we introduce a gated residual fusion method, which enables the model to selectively refine the BEV feature map with visual information or rely exclusively on Radar data, depending on the quality, robustness and relevance of the visual input.

After adaptive fusion, the 3D features are again mapped onto a BEV map, which is then forwarded to a detection head adapted from [17]. Specifically, we modified the focal loss to better match the true shape and size of different object bounding boxes in the range-azimuth format of the feature maps. This results in better detection performance especially for smaller road users like pedestrians and cyclists.

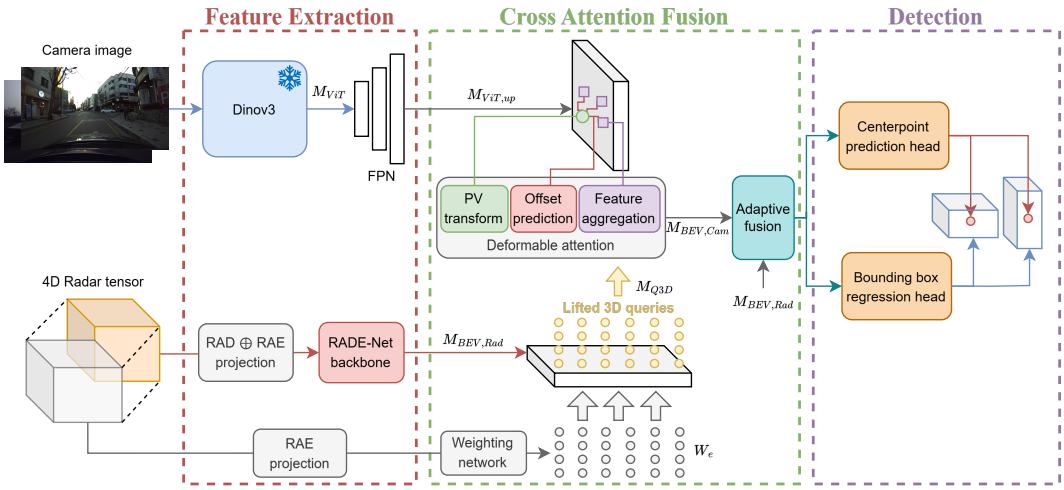

Figure 1. Overview of the DinoRADE architecture.

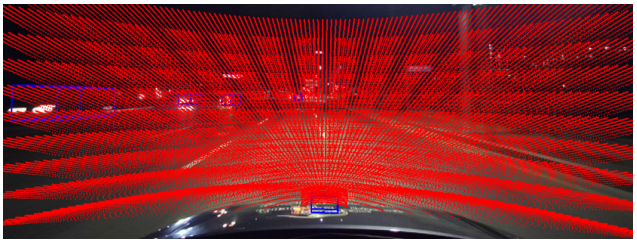

Figure 2. Reference points projected from 3D Radar queries to camera image (red) and Radar sensor location (blue).

3.1. Radar Feature Extraction

We adapt the RADE-Net backbone [17] to employ a dual encoding architecture in which both 3D projections undergo independent processing before concatenation along the feature dimension. This approach allows the model to selectively integrate Doppler and elevation information prior to feature extraction, thereby enhancing flexibility when supplementary elevation data is introduced during the subsequent lifting stage. Upon completion of processing, the backbone generates a BEV feature map $M_{BEV,Rad} \in \mathbb{R}^{256\times112\times128}$, which encodes each bin in the 256×112 Radar range-azimuth domain with a 128-dimensional feature representation.

3.2. Camera Feature Extraction

We employ a DINOv3 ViT-S/16 [36] pretrained on 1.689 billion images to extract general-purpose visual feature representations. The input image, with a resolution of 720x1280 pixels, is divided into non-overlapping 16x16 pixel patches, yielding 3600 patches in total. These patches are processed by the frozen vision transformer, which produces an output feature map $M_{ViT} \in \mathbb{R}^{45\times80\times384}$. To obtain features at a higher spatial resolution, the output feature map undergoes spatial upsampling using a two-layer Feature Pyramid Network (FPN) which reduces the feature dimension, resulting in an upsampled feature map $M_{ViT,up} \in \mathbb{R}^{180\times320\times128}$.

3.3. Weighted Feature Lifting

To account for the distribution of Radar features along the elevation dimension, we expand the BEV feature map into E=10 elevation segments and apply appropriate weighting to the bins within each segment. This yields the lifted 3D query map $M_{Q3D} \in \mathbb{R}^{256\times112\times128\times10}$ according to:

$$M_{Q3D} = [M_{BEV,Rad}]_{i=1}^{E} \times W_e^T \tag{1}$$

where the weights W_e are learned from the RAE projection.

$$W_e = [\text{Softmax}(\text{MLP}(\mathcal{P}^{RAE}))]_{i=1}^{128} \tag{2}$$

This concept exploits the property that the RAE projection $\mathcal{P}^{RAE}$ provides a direct representation of the spectral power distribution along the elevation dimension for each range-azimuth bin within the feature map.

3.4. Deformable Cross Attention

The cross attention module employs deformable attention [52] to aggregate camera features for a set of 3D queries derived from Radar input. To guide this aggregation process, we generate reference points for each 3D query and project them onto $M_{ViT,up}$. This projection leverages Radar sampling parameters, sensor positions, and camera intrinsics/extrinsics from [28], providing spatial priors that indicate relevant regions in the perspective view. Figure 2 illustrates the resulting reference point distribution on the camera image. Following [52], we predict four offsets

around each reference point to sample features, which are then aggregated to update M_{Q3D}. Finally, we average the updated 3D query map along the height dimension to obtain the BEV refined map $M_{BEV,Cam}$.

3.5. Adaptive Fusion

The adaptive fusion model is designed to learn, for each range-azimuth bin, whether features should be extracted from the Radar-only feature map $M_{BEV,Rad}$ or from the VFM-refined feature map $M_{BEV,Cam}$. Specifically, we employ a gated fusion approach in which a learned gate $\Gamma \in \mathbb{R}^{256\times112\times128}$ dynamically regulates the composition of the fused Radar-camera feature map M_f as follows:

$$M_f = \Gamma M_{BEV,Rad} + (1-\Gamma) M_{BEV,Cam} \tag{3}$$

with

$$\Gamma = \text{Sigmoid}(\,\text{MLP}(M_{BEV,Rad} \oplus M_{BEV,Cam})) \tag{4}$$

where $\oplus$ represents the concatenation operation and MLP a multilayer perceptron with two layers.

3.6. Detection Head

The detection head performs center-point detection, object classification, and bounding box regression based on the fused BEV feature map M_f, thereby exploiting the Radar's native range-azimuth coordinate representation [17]. Multi-class detection is achieved through class-specific heatmaps, with each object class represented in a separate channel.

3.7. Loss

We adapt the loss strategy from [17] by employing a combination of loss components: a focal loss for center-point convergence, and a composite regression loss that integrates the Gaussian-Wasserstein Distance (GWD) [45] with a smooth L1 term. While the focal loss performs adequately for larger objects such as cars, busses and trucks, it presents challenges for smaller objects like pedestrians and cyclists. The isotropic Gaussian distribution generated at the ground truth center point in range-azimuth coordinates with a fixed $\sigma = 3$ covers a disproportionately large physical area at greater range values due to the polar coordinates of the Radar data. This issue is particularly problematic for small objects, as the model struggles to converge to the true object center. We conducted an analysis of various σ values and determined that $\sigma = 0.75$ yields optimal center-point convergence across all object classes while simultaneously enhancing the detection performance for smaller road users.

4. Experiments

We conducted a comprehensive experimental evaluation of our approach, with particular emphasis on the detection of vulnerable road users and demonstrate the robustness of our method by evaluating Average Precision (AP) and mean Average Precision (mAP) metrics across various weather conditions. Following the K-Radar benchmark [15], we evaluate our trained model on single-class detection for the 'Sedan' class, which exhibits strong representation across all weather conditions in the dataset and thus serves as a reliable performance indicator. We employ this single-class evaluation to compare our approach against other methods utilizing different sensor modality configurations, as presented in Table 1.

To provide more comprehensive insights, we present the AP and mAP detection performance for different road user classes across all weather conditions in Table 2. Following the K-Radar benchmark [28], existing works primarily report AP performance for the 'Sedan' and 'Bus or Truck' classes which are represented strongly in the dataset [28, 37]. To our knowledge, only [17] additionally evaluates detection performance on 'Pedestrian' and 'Bicycle' classes, limiting direct comparison for multiclass detection. To facilitate a more comprehensive evaluation, we establish an additional baseline by retraining DPFT [5] and extending the single-class detection approach to all five object classes, with results presented in Table 3. Following K-Radar [28], we consider detections within a Region of Interest (ROI) of $x \in [0, 72m]$, $y \in [-6.4m, 6.4m]$ and $z \in [-2m, 6m]$.

4.1. Dataset

Our evaluation is conducted on the large-scale K-Radar dataset [28] and adheres to the official KITTI evaluation protocol [7]. Results reported in Table 1 are derived from the earlier dataset label version 1.1 [28] to ensure comparability with existing approaches. Table 2– 4 employ the updated labels from v2.1, which include additional annotations missing in v1.1 [39]. The dataset provides 35k Radar frames along with synchronized camera images. Annotations consist of rotated 3D bounding boxes with seven class labels, of which we utilize five: 'Sedan', 'Bus or Truck', 'Pedestrian', 'Bicycle', and 'Motorcycle'. The additional classes 'Pedestrian Group' and 'Bicycle Group' are severely underrepresented in the dataset and introduce classification ambiguity between the relatively similar 'Pedestrian' and 'Bicycle' classes.

4.1.1. Training

We train our model on a NVIDIA A30 GPU for 11 epochs using a batch size of 6. For optimization, we employ the AdamW optimizer with an initial learning rate of 0.001 and a weight decay of 0.01. We apply a cosine annealing learning rate scheduler with a minimum learning rate of 10^{-4}. Our model has a total number of 31M learnable parameters and 21M frozen weights of the DINOv3 ViT-S/16 [36].

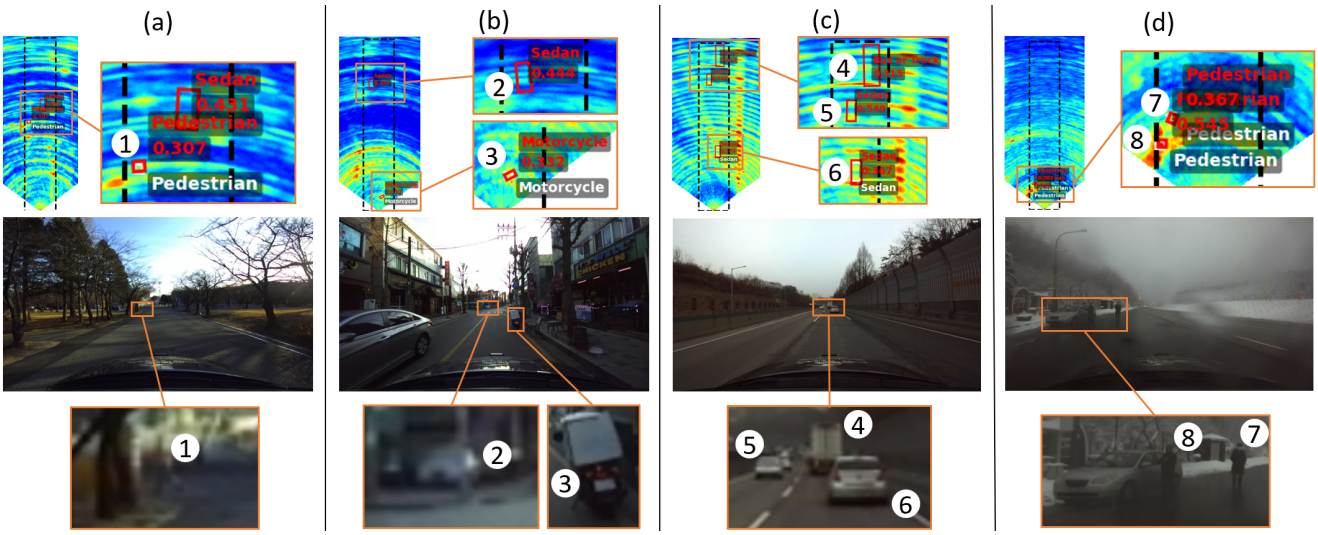

Figure 3. DinoRADE performance visualization in four different scenarios: (a) university campus, (b) alleyway, (c) highway, and (d) road shoulder. Ground truth bounding boxes and class labels are shown in white. Predicted bounding boxes with predicted class and confidence score are shown in red. The ROI where detections are considered is marked with a dashed black rectangle. Magnified views are indicated in orange.

Table 1. 3D Average Precision (AP) comparison on 'Sedan' class in all weather conditions using K-Radar v1.1.

Methods	Modality	Total	Normal	Overcast	Fog	Rain	Sleet	Light Snow	Heavy Snow
Voxel-RCNN [3]	L	46.4	81.8	69.6	48.8	47.1	46.9	54.8	37.2
CasA [41]		50.9	82.2	65.6	44.4	53.7	49.9	62.7	36.9
TED-S [42]		51.0	74.3	68.8	45.7	53.6	44.8	63.4	36.7
RTNH [28]	R	47.4	49.9	56.7	52.8	42.0	41.5	50.6	44.5
RTNH+ [15]		57.6	-	-	-	-	-	-	-
RADE-Net [17]		66.7	66.3	74.5	**82.1**	57.2	**69.4**	63.9	**68.9**
VPFNet [51]	LC	52.2	81.2	76.3	46.3	53.7	44.9	63.1	36.9
TED-M [42]		52.3	77.2	69.7	47.4	54.3	45.2	64.3	36.3
MixedFusion [49]		55.1	**84.5**	76.6	53.3	55.3	49.6	68.7	44.9
EchoFusion [21]	RC	47.4	51.5	65.4	55.0	43.2	14.2	53.4	40.2
DPFT [5]		56.1	55.7	59.4	63.1	49.0	51.6	50.5	50.5
WRCFormer [10]		58.7	55.9	59.7	70.1	52.2	54.7	59.2	54.0
DinoRADE (ours)		**70.8**	72.8	**84.9**	69.7	**68.4**	58.4	**70.4**	65.0

4.1.2. Inference

Inference speed was evaluated on an NVIDIA A30 GPU, achieving an average processing time of 190.34 ms per frame. It is worth noting that the framework has not been optimized for inference speed, indicating substantial potential for improvement and further acceleration in real-time deployment scenarios.

4.2. Results

We present qualitative results in Fig. 3 across four scenarios for different road users. Scenario (a) demonstrates the correct detection of a pedestrian (1) at 37m from the ego vehicle, right next to a false positive (FP) prediction. In scenario (b), while a parked motorcycle (3) is accurately detected, an unlabeled parked vehicle (2) causes a FP prediction. Scenario (c) exhibits similar behavior, where a vehicle (6) traveling ahead is correctly classified, whereas the unlabeled car (5) and truck (4) preceding it result in FP predictions due to missing annotations in the dataset. Finally, scenario (d) demonstrates the successful detection of two pedestrians (7,8), with the adjacent vehicle positioned beyond the ROI.

Examining quantitative results in Table 1, we observe

Table 2. AP and mAP (Total) evaluation of all road users in all weather conditions using K-Radar v2.1. (-) indicates no representation in the test set and (*) indicates a severe under-representation of under 2% of total objects in the train set of the respective weather condition.

Weather	Total		Sedan		Bus or Truck		Pedestrian		Motorcycle		Bicycle	
	3D	BEV	3D	BEV	3D	BEV	3D	BEV	3D	BEV	3D	BEV
Total	36.99	39.61	71.38	75.32	54.92	58.93	33.01	38.12	3.77	3.77	21.89	21.89
Normal	29.04	30.08	71.52	74.13	46.61	49.17	*0.99	*0.99	*3.77	*3.77	22.32	22.32
Overcast	70.89	74.33	74.20	75.16	67.57	73.51	-	-	-	-	-	-
Fog	67.41	75.68	83.84	89.91	-	-	50.98	61.46	-	-	-	-
Rain	24.10	25.97	71.32	76.92	*0.99	*0.99	*0.00	*0.00	-	-	-	-
Sleet	47.14	49.63	66.89	70.43	44.69	46.41	29.85	32.04	-	-	-	-
Light Snow	70.73	76.63	69.29	76.72	72.17	76.55	-	-	-	-	-	-
Heavy Snow	71.93	76.34	66.43	66.75	77.42	85.94	-	-	-	-	-	-

Table 3. AP and mAP (Total) comparison with Radar-Camera and Radar-only methods for multiple object classes using K-Radar v2.1.

Methods	Mod.	Total		Sedan		Bus or Truck		Pedestrian		Motorcycle		Bicycle	
		3D	BEV	3D	BEV	3D	BEV	3D	BEV	3D	BEV	3D	BEV
RADE-Net [17]	R	20.93	23.96	56.75	63.21	40.99	48.29	5.40	6.79	0.00	0.00	1.49	1.49
DPFT [5]	RC	11.65	13.54	38.99	41.02	17.60	20.55	1.66	6.14	0.00	0.00	0.00	0.00
Ours	RC	36.99	39.61	71.38	75.32	54.92	58.93	33.01	38.12	3.77	3.77	21.89	21.89

that purely optical methods utilizing LiDAR and camera demonstrate strong performance under normal and overcast weather conditions but exhibit substantial degradation in adverse weather. In contrast, Radar-only performance remains relatively consistent across all weather conditions, while Radar-camera fusion enhances performance under normal and overcast conditions while maintaining robustness in adverse weather. Notably, our approach outperforms all prior methods in total AP as well as in overcast, rain, and light snow conditions. Additionally, it can be observed that Radar-based methods perform slightly better in adverse weather compared to normal conditions, which may appear counterintuitive but is attributed to the distribution of labeled bounding boxes across weather conditions and detection distances, as discussed in [13].

Results presented in Table 2 show the detection performance for different road users across weather conditions. Pedestrian detection performs strong in fog and sleet, attributed to simple scenes, well-represented in both train and test set. Road scenarios in normal conditions are more challenging, including a high number of different road users and complex moving paths. This introduces classification mismatches, for instance where the model mistakes a person standing on a scooter for a pedestrian which generates a false positive in the 'Pedestrian' class and a false negative in the 'Motorcycle' class. Additionally, pedestrians are strongly underrepresented in the training set with 1.6% in 'normal' and 0% in 'rain' conditions while representing 28.4% in sleet, 24% in fog and 6.9% in light snow. Similarly, 'Motorcycle' and 'Bicycle' class only represent 1% and 3.1% respectively of training data in normal weather condition.

In Table 3 we compare our model with RADE-Net [17] and DPFT [5]. Our method achieves substantially higher performance across multiple classes. However, it should be noted that DPFT was originally designed and optimized for single-class detection rather than multi-class scenarios. Consequently, the multi-class evaluation may not fully reflect DPFT's capabilities within its intended operational scope.

4.3. Ablation Study

In Table 4 we demonstrate the performance of different module configurations using both 3D and BEV AP on the 'Sedan' class, which exhibits consistent representation throughout the dataset. We begin with the Radar-only (R) configuration, where no camera-based feature refinement is applied. Subsequently, we incorporate the lifted 3D query cross-attention and adaptive fusion module (R+C), yielding an approximately 8% improvement in both metrics. Furthermore, we integrate the azimuth-elevation-based weighting network (R+C+W), which redistributes features along the elevation domain based on their true distribution in the original Radar spectrum. This addition naturally provides greater improvement to AP_{3D} with a 1.77% gain, as it facilitates elevation refinement, while improving AP_{BEV} by

Figure 4. Examples for partially occluded (1), heavily occluded (2), and fully occluded (3).

0.36%. Finally, we replaced the DINOv3 ViT with a pre-trained ResNet50 [11] (R+C*+W), fine-tuning only the final layer. This substitution results in approximately a 3% performance decrease in both metrics, demonstrating the superior feature representation capabilities of the DINOv3 VFM.

Table 4. Ablation Study

Metric	R	R+C	R+C+W	R+C*+W
AP_{3D}	61.65	69.61	**71.38**	68.43
AP_{BEV}	66.68	74.96	**75.32**	72.42

5. Discussion

5.1. Camera Images in Adverse Weather Conditions

Upon examination of the dataset, we discovered a substantial number of frames where the camera image is partially, heavily, or fully occluded by adverse weather effects, including rain, fog, sleet, and snow. We estimate 7.5k partially occluded, 4.8k heavily occluded, and 4.9k fully occluded frames within the whole dataset (35k frames) and present three examples in Fig. 4. These weather-related effects pose a significant challenge to the training pipeline, likely contributing to the suboptimal performance of camera-based approaches in existing methods, where the incorporation of camera data yields limited performance improvements.

5.2. Dataset Annotation

The K-Radar dataset has undergone several iterations to improve annotation completeness and address visibility limitations. Version 2.0 addressed missing labels through an automated labeling pipeline utilizing LiDAR and camera data [39], while versions 1.1 and 2.1 provide additional information regarding the physical visibility for Radar and LiDAR modality due to sensor placement [28]. However, missing annotations in the dataset still pose a challenge for model training and evaluation. Upon examination of false positive predictions, we identified instances where our model correctly detects objects in the scene that lack corresponding ground truth bounding boxes. This phenomenon is particularly prevalent in scenarios involving parked vehicles and oncoming traffic. Consequently, this negatively affects both evaluation metrics and training dynamics, as the model receives incorrect penalty for accurate predictions, thereby degrading the learning process.

6. Conclusion

We present DinoRADE, a Radar-camera fusion framework for 3D object detection that performs consistently across all weather conditions and leverages inter-perspective feature aggregation using the output of a DINOv3 VFM, which improves detection especially in good weather conditions. We demonstrate the performance improvement due to cross-attention fusion with VFM image features compared to ResNet features and Radar-only detection, as well as the improvement attributed to our weighted feature lifting module for AP_{3D}. We present qualitative and quantitative results where we outperform existing methods by 12.1% in AP_{3D} on the 'Sedan' class. To emphasize the relevance of VRUs we report and discuss comprehensive evaluation results of multiple road users in different weather conditions where we achieve 51% AP_{3D} for 'Pedestrian' in foggy scenarios. Additionally, we compare our superior multi-class results with two baselines. With this work, we hope to further drive weather-robust object detection while providing a foundation to include performance metrics of vulnerable road user detection in future research.

Acknowledgment

This work is supported by Infineon Technologies Austria AG and the European Union through the Horizon Europe programme (Grant Agreement project 101092834). Funded by the European Union.

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, Los Alamitos, CA, USA, 2020. IEEE Computer Society. 2

[2] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Y. Qiao. Vision transformer adapter for dense predictions. *ArXiv*, abs/2205.08534, 2022. 3

[3] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35: 1201–1209, 2021. 6

[4] Sheng Feng, Xueying Cai, Limin Li, Weixing Wang, and Senang Ying. A review of research on vehicle detection in adverse weather environments. *Journal of Traffic and Transportation Engineering (English Edition)*, 12(5):1452–1483, 2025. 1, 3

[5] F. Fent, A. Palffy, and H. Caesar. Dpft: Dual perspective fusion transformer for camera-radar-based object detection. *IEEE Transactions on Intelligent Vehicles*, 10(11): 4929–4941, 2025. 1, 2, 3, 5, 6, 7

[6] Xiangyu Gao, Youchen Luo, Guanbin Xing, Sumit Roy, and Hui Liu. Raw ADC data of 77GHz MMWave radar for automotive object detection, 2022. Distributed by IEEE Dataport. 2

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 5

[8] James Giroux, Martin Bouchard, and Robert Laganière. T-fftradnet: Object detection with swin vision transformers from raw adc radar signals. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4032–4041, 2023. 2

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 3

[10] Runwei Guan, Jianan Liu, Shaofeng Liang, Fangqiang Ding, Shanliang Yao, Xiaokai Bai, Daizong Liu, Tao Huang, Guoqiang Mao, and Hui Xiong. Wavelet-based multi-view fusion of 4d radar tensor and camera for robust 3d object detection, 2026. 2, 3, 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 8

[12] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View, 2022. arXiv:2112.11790 [cs]. 3

[13] Xun Huang, Ziyu Xu, Hai Wu, Jinlong Wang, Qiming Xia, Yan Xia, Jonathan Li, Kyle Gao, Chenglu Wen, and Cheng Wang. L4dr: Lidar-4dradar fusion for weather-robust 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3806–3814, 2025. 7

[14] JunXin Jin, Wei Liu, Zuotao Ning, Qixi Zhao, Shuai Cheng, and Jun Hu. 3d object detection for autonomous driving: A survey. In *2024 36th Chinese Control and Decision Conference (CCDC)*, pages 3825–3832, 2024. 3

[15] Seung-Hyun Kong, Dong-Hee Paek, and Sangyeong Lee. RTNH+: Enhanced 4D Radar Object Detection Network Using Two-Level Preprocessing and Vertical Encoding. *IEEE Transactions on Intelligent Vehicles*, 10(2):1427–1440, 2025. 2, 5, 6

[16] Akhil M Kurup and Jeremy P. Bos. Dsor: A scalable statistical filter for removing falling snow from lidar point clouds in severe winter weather. *ArXiv*, abs/2109.07078, 2021. 1

[17] Christof Leitgeb, Thomas Puchleitner, Max Peter Ronecker, and Daniel Watzenig. RADE-Net: Robust Attention Network for Radar-only Object Detection in Adverse Weather, 2026. 1, 2, 3, 4, 5, 6, 7

[18] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2025. 1

[19] Teck Yian Lim, Spencer Markowitz, and Minh Do. Radical: A synchronized fmcw radar, depth, imu and rgb camera data dataset with low-level fmcw radar signals. *IEEE Journal of Selected Topics in Signal Processing*, PP:1–1, 2021. 2

[20] Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. RCBEVDet: Radar-Camera Fusion in Bird's Eye View for 3D Object Detection . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14928–14937, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 1

[21] Yang Liu, Feng Wang, Naiyan Wang, and ZHAO-XIANG ZHANG. Echoes beyond points: Unleashing the power of raw radar data in multi-modality fusion. In *Advances in Neural Information Processing Systems*, pages 53964–53982. Curran Associates, Inc., 2023. 2, 6

[22] Siqi Lu, Junlin Guo, James R. Zimmer-Dauphinee, Jordan M. Nieusma, Xiao Wang, Parker VanValkenburgh, Steven A. Wernke, and Yuankai Huo. Vision foundation models in remote sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 13(3):190–215, 2025. 3

[23] Xinzhu Ma, Wanli Ouyang, Andrea Simonelli, and Elisa Ricci. 3d object detection from images for autonomous driving: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3537–3556, 2024. 3

[24] Michael Meyer and Georg Kuschk. Automotive radar dataset for deep learning based 3d object detection. In *2019 16th European Radar Conference (EuRAD)*, pages 129–132, 2019. 2

[25] Alexander Musiat, Laurenz Reichardt, Michael Schulze, and Oliver Wasenmüller. Radarpillars: Efficient object detection from 4d radar point clouds. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1656–1663, 2024. 2, 3

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 2, 3

[27] Arthur Ouaknine, Alasdair Newson, Julien Rebut, Florence Tupin, and Patrick Pérez. Carrada dataset: Camera and automotive radar with range- angle- doppler annotations. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5068–5075, 2021. 2

[28] Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. K-radar: 4d radar object detection for autonomous driving in various weather conditions. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 4, 5, 6, 8

[29] Andras Palffy, Ewoud Pool, Srimannarayana Baratam, Julian F. P. Kooij, and Dariu M. Gavrila. Multi-class road user detection with 3+1d radar in the view-of-delft dataset. *IEEE Robotics and Automation Letters*, 7(2):4961–4968, 2022. 2, 3

[30] Julien Rebut, Arthur Ouaknine, Waqas Malik, and Patrick Perez. Raw High-Definition Radar for Multi-Task Learning . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17000–17009, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 2

[31] Julien Rebut, Arthur Ouaknine, Waqas Malik, and Patrick Pérez. Raw high-definition radar for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17021–17030, 2022. 2

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 2

[33] Jonas Schramm, Niclas Vödisch, Kürsat Petek, B Ravi Kiran, Senthil Yogamani, Wolfram Burgard, and Abhinav Valada. Bevcar: Camera-radar fusion for bev map and object segmentation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1435–1442, 2024. 1, 3

[34] Linda Senigagliesi, Gianluca Ciattaglia, Deivis Disha, and Ennio Gambi. Classification of human activities based on automotive radar spectral images using machine learning techniques: A case study. In *2022 IEEE Radar Conference (RadarConf22)*, pages 1–6, 2022. 1

[35] Marcel Simeonov, Andrei Kurdiumov, and Milan Dado. Real-time 3d scene understanding for road safety: Depth estimation and object detection for autonomous vehicle awareness. *Vehicles*, 8(2), 2026. 1

[36] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. 2, 3, 4, 5

[37] Seung-Hyun Song, Dong-Hee Paek, Minh-Quan Dao, Ezio Malis, and Seung-Hyun Kong. Enhanced 3d object detection via diverse feature representations of 4d radar tensor. *IEEE Sensors Journal*, pages 1–1, 2026. 5

[38] Youyi Song, Zhen Yu, Teng Zhou, Jeremy Yuen-Chun Teoh, Baiying Lei, Kup-Sze Choi, and Jing Qin. Learning 3d features with 2d cnns via surface projection for ct volume segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 176–186, Cham, 2020. Springer International Publishing. 3

[39] Min-Hyeok Sun, Dong-Hee Paek, Seung-Hyun Song, and Seung-Hyun Kong. Efficient 4d radar data auto-labeling method using lidar-based object detection network. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 2616–2621, 2024. 5, 8

[40] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8437–8445, 2019. 3

[41] Hai Wu, Jinhao Deng, Chenglu Wen, Xin Li, Cheng Wang, and Jonathan Li. Casa: A cascade attention network for 3-d object detection from lidar point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 6

[42] Hai Wu, Chenglu Wen, Wei Li, Xin Li, Ruigang Yang, and Cheng Wang. Transformation-equivariant 3d object detection for autonomous driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:2795–2802, 2023. 6

[43] Yuhao Xiao, Xiaoqing Chen, Yingkai Wang, and Zhongliang Fu. Radar–camera fusion in perspective view and bird's eye view for 3d object detection. *Sensors*, 25(19), 2025. 1

[44] Bo Yang, Ishan Khatri, Michael Happold, and Chulong Chen. ADCNet: Learning from Raw Radar Data via Distillation, 2023. arXiv:2303.11420 [eess]. 2

[45] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. *CoRR*, abs/2101.11952, 2021. 5

[46] Shanliang Yao, Runwei Guan, Xiaoyu Huang, Zhuoxiao Li, Xiangyu Sha, Yong Yue, Eng Gee Lim, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu, and Yutao Yue. Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *IEEE Transactions on Intelligent Vehicles*, 9(1):2094–2128, 2024. 2

[47] Shanliang Yao, Runwei Guan, Zitian Peng, Chenhang Xu, Yilu Shi, Weiping Ding, Eng Gee Lim, Yong Yue, Hyungjoon Seo, Ka Lok Man, Jieming Ma, Xiaohui Zhu, and Yutao Yue. Exploring radar data representations in autonomous driving: A comprehensive review. *IEEE Transactions on Intelligent Transportation Systems*, 26(6):7401–7425, 2025. 1, 2

[48] Ao Zhang, Farzan Erlik Nowruzi, and Robert Laganiere. Raddet: Range-azimuth-doppler based radar object detection for dynamic road users. In *2021 18th Conference on Robots and Vision (CRV)*, pages 95–102, 2021. 2

[49] Cheng Zhang, Hai Wang, Long Chen, Yicheng Li, and Yingfeng Cai. Mixedfusion: An efficient multimodal data fusion framework for 3-d object detection and tracking. *IEEE Transactions on Neural Networks and Learning Systems*, 36 (1):1842–1856, 2025. 6

[50] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 3

[51] Hanqi Zhu, Jiajun Deng, Yu Zhang, Jianmin Ji, Qiuyu Mao, Houqiang Li, and Yanyong Zhang. Vpfnet: Improving 3d object detection with virtual point based lidar and stereo data fusion. *IEEE Transactions on Multimedia*, 25:5291–5304, 2023. 6

[52] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ArXiv*, abs/2010.04159, 2020. 3, 4