CARB-Net: Camera-Assisted Radar-Based Network for Vulnerable Road User Detection

Wei-Yu Lee, Martin Dimitrievski, David Van Hamme, Jan Aelterman, Ljubomir Jovanov, and Wilfried Philips

TELIN-IPI, Ghent University-IMEC, Gent, Belgium

Abstract. Ensuring a reliable perception of vulnerable road users is crucial for safe autonomous driving. Radar stands out as an appealing sensor choice due to its resilience in adverse weather, cost-effectiveness, depth sensing capabilities, and established role in adaptive cruise control. Nevertheless, radar's limited angular resolution poses challenges in object recognition, especially in distinguishing targets in close proximity. To tackle this limitation, we present the Camera-Assisted Radar-Based Network (CARB-Net), a novel and efficient framework that merges the angular accuracy of a camera with the robustness and depth sensing capabilities of radar. We integrate camera detection information through a ground plane feed-forward array, entangling it with the early stages of a radar-based detection network. Furthermore, we introduce a unique context learning approach to ensure graceful degradation in situations of poor radar Doppler information or unfavorable camera viewing conditions. Experimental validations on public and our proposed datasets, along with benchmark comparisons, showcase CARB-Net's superiority, boasting up to a 12% improvement in mAP performance. A series of ablation studies further emphasize the efficacy of the CARB-Net architecture. Our proposed dataset is released at https://github.com/weiyulee/RadVRU.

Keywords: Micro-Doppler Signature · Vulnerable Road User Detection

1 Introduction

Robust Vulnerable Road User (VRU) perception is vital for driving safety. Among the myriad of sensors used, millimeter-wave radar stands out due to its resilience in challenging light and weather conditions, such as rain, fog, and snow, and its depth and Doppler velocity capability. Its cost-effectiveness has also contributed to its widespread adoption, making it a fundamental component in many Advanced Driving Assistance Systems (ADAS) applications.

However, radar sensors have inherent limitations: Their angular resolution is limited by the size of the antenna array and operating wavelength. Compared to HD cameras, typical radars have two orders of magnitude lower angular resolution [30, 36, 40], making it challenging to distinguish between closely spaced objects. Beyond detection, this low signal resolution makes classifying between object categories even more difficult. In cluttered environments, multipath interference noise further complicates the interpretation of radar data.

Fig. 1: An example of micro-Doppler patterns. *left*: The radar signal is projected and overlaid with the synchronized camera image. *right*: The radar Doppler signal is plotted over time. The micro-Doppler pattern of the cyclist is highlighted by a dotted line.

In the literature, radar is often used as a means to improve 3D object detection in monocular systems. By fusing imaging information from cameras with radar, researchers have proposed to exploit the strengths of both technologies [13, 14, 42]. In these prior studies, cameras play a key role in providing rich visual cues that can be used to train effective object detection and classification models. However, in low-light environments, cameras cannot provide the recognition of road users, and detection of targets by only using *radar point clouds* becomes difficult due to their sparsity. One of the possible alternative cues for VRU detection is using Doppler information from dense radar tensors, alternatively referred to as Radar Frequency (RF) images. This radar data representation provides the unique motion patterns in the time-frequency domain [5,16,17]. For example, in Figure 1, a cyclist exhibits an oscillatory motion of the limbs, which results in a RF signal with sufficient features to classify these road users. These periodically changing motion patterns are commonly referred to as *micro-Doppler patterns*, providing fundamental features for identifying road users. However, the low azimuth resolution of radar sensors still limits the accuracy of the predicted object positions in these studies.

Motivated by these challenges, we propose a RF-based, camera-assisted fusion network with two significant novelties that enhance detection accuracy in nominal conditions (*i.e.*, sensors are working as expected), but more importantly, under conditions of sensor failure. A new cooperative fusion approach introduces summarized camera observations into the early stages of a radar-based network, helping in better feature extraction over the Doppler signal in the RF images. Furthermore, we propose a robust way of entangling camera and radar data, helping the network enhance robustness against scenarios where a sensor fails to detect the targets. By applying these two improvements, we achieve detector robustness without the need for excessive training in unforeseen situations.

We focus on the task of Bird's Eye View (BEV) vulnerable road user perception, where we train a radar-based network on radar frequency images, or more specifically, Range-Azimuth-Doppler (RAD) tensors. The RAD encodes the micro-Doppler pattern of targets at individual range-azimuth spatial cells and is used as the core feature for both detection and classification. Camera information is appended to the RAD tensor in the form of a summarized detection array in the BEV plane, which assists the radar-based network when it is available. We use an off-the-shelf camera-based object detection model to produce detec-

tions of road users and populate the feed-forward array using back-projection. The input layer of the network then applies pixel-shuffle [37] to reorganize the high-resolution camera observations into a common representation, matching the low-resolution radar data. This allows a seamless cooperative fusion with radar tensors without loss of information.

Robustness against scenarios where a sensor fails to detect the targets is achieved by augmenting the data and expanding the ground truth labels into sub-categories of possible fusion contexts of nominal and deteriorated sensor performance for each object. The novel aspect of training is to generate and encode these conditions as control variables, that augment the ground truths during training. This strategy encourages the model to learn more sophisticated and finer features for each modality without additional human annotation, enhancing its adaptability to diverse fusion contexts. Our experiments, conducted on both public and our proposed RadVRU datasets, demonstrate the effectiveness of our approach. Our method not only achieves state-of-the-art detection performance but also exhibits a remarkable up to 12% improvement in challenging sensor failure cases. Our contributions are outlined as follows:

1. Novel camera-assisted radar-based network based on pixel-shuffling of high-resolution vision feed-forward and low-resolution micro-Doppler signatures.
2. Novel context learning strategy, utilizing context-specific control variables to train a model that is robust to conditions where a sensor is unable to detect the targets.
3. An extensive experimental evaluation on two datasets for BEV detection where we objectively compare to existing state-of-the-art methods.

2 Related Work

Radar-Based Object Detection Prior studies utilize radar as the primary sensor for object detection, employing various data representations: point clouds and dense radar tensors. For radar point clouds, the sparsity and lack of context pose detection challenges [33, 35, 41] for neural networks to learn features effectively. On the other hand, image-like radar tensors provide rich but noisy contextual data [8, 15]. Many prior studies utilize deep learning to recognize objects by their motion characteristics [3, 9]. For instance, radar's unique micro-Doppler signature, a time-varying Doppler frequency shift captured in RAD tensors, enables accurate recognition of road users like pedestrians and cyclists [5, 6, 16, 17]. This distinctive motion pattern advantage motivates our use of RAD tensors as inputs, harnessing the micro-Doppler signature for robust object detection.

On the other hand, radar sensors, while effective at detecting moving road users, struggle to separate closely packed objects, such as groups of pedestrians, due to their low angular resolution [41]. Additionally, radar sensors fail to detect stationary objects due to interference from other reflections and a lack of motion information. As a result, the existing radar-only methods cannot accurately localize and recognize all objects. In this study, we utilize observations from the camera as assistance to improve the precision of azimuth localization.

Radar-Camera-Based Object Detection Recently, using BEV representation to merge images from multiple camera and radar sensors has sparked a rising wave of research in autonomous driving. Prior studies adopt large pretrained image networks, such as LSS [32], BEVFormer [20], or BEVDepth [19], to project 2D image features onto the BEV for further fusion with radar data. For instance, in [13], the authors use proposals from a camera-based network to associate and merge with radar point clouds. In [14], a radar-assisted transformation method is proposed to project 2D images onto BEV and aggregate image and radar feature maps for detection. However, the perspective transformation of camera views highly relies on accurate calibration parameters and good light conditions. Numerous objects visible to cameras may yield only a few radar points, posing significant challenges when cameras are unreliable and sparse point clouds cannot provide sufficient evidence for recognition.

In contrast, our method interfaces with the radar at a much earlier stage, utilizing 3D dense radar tensors as input to learn the micro-Doppler patterns for recognition instead of relying on camera features. Therefore, detection can be performed only on radar data in low-light conditions without cameras. Moreover, existing fusion methods often necessitate training a huge network on a large amount of both camera and radar data, which poses challenges due to the lack of publicly available datasets containing both modalities and the high computing effort. To address this, we utilize detections from an efficient off-the-shelf camera-based detection model, reducing the training and inference burden. Since our method's core is a radar-based detector using micro-Doppler patterns to recognize VRUs, only aided by the camera when it is available, we mainly consider the dataset providing 3D radar tensors for experiments, and the existing BEV radar-camera fusion methods cannot be directly applied to this task.

Robustness of Multimodal Fusion In practice, it is quite common for sensors to become unreliable in specific contexts. This negatively affects the robustness of the perception system, such as occlusions or low-light environments for cameras. Prior studies introduce sensor dropout [4, 11, 38, 39] during training to enhance the network's ability to handle these conditions with single modality features. However, simply removing the sensor data from the input constrains the fusion potential and usually degrades the overall detection performance. In this work, we introduce a novel context learning strategy that augments the data by creating challenging scenarios that render sensor observations unreliable, and we instruct the network to recognize these situations through auxiliary queries, enhancing model robustness in detection tasks.

3 Methodology

In this paper, we introduce Camera-Assisted Radar-Based Network (CARB-Net), a novel neural network model for integrating low spatial resolution radar tensors and high-resolution camera images. As depicted in Fig. 2, the proposed framework operates in the spatio-temporal domain, taking sequential camera

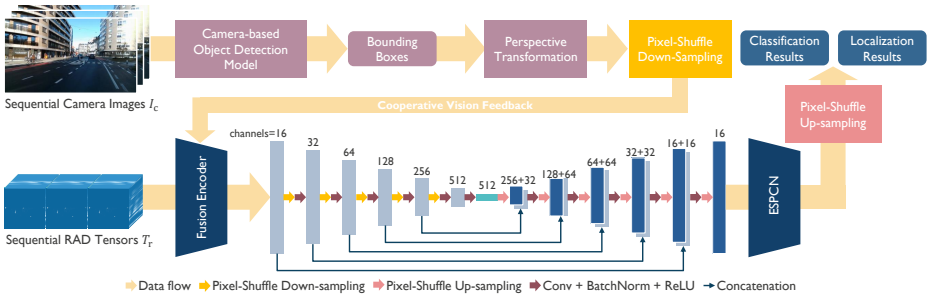

Fig. 2: Overview of the proposed Camera-Assisted Radar-Based Network (CARB-Net).

images $I_c \in \mathbb{R}^{NC_c \times H_c \times W_c}$ and sequential RAD radar tensors $T_r \in \mathbb{R}^{NC_r \times H_r \times W_r}$, where N denotes the number of tensors and images in the sequence, C_c, C_r represent channel sizes, and (H_c, W_c) and (H_r, W_r) are the spatial dimensions of the RAD tensors and camera images.

Unified Data Representation The data representations of radars and cameras are different, which poses challenges in sensor fusion as association requires relating objects in different coordinate systems. Therefore, it is essential to establish a common representation suitable for both fusion and detection. Front View and Bird's Eye View are the two mainstream representations in prior studies. For front view, due to the low resolution in the azimuth angle provided by the radar as well as camera calibration errors, projected radar signals may deviate from the object in the image, which makes the observations of the same object contradictory and hard to be fused by networks. On the other hand, BEV provides a view that accurately portrays the scale and positions of objects, making it a suitable representation for autonomous driving tasks. In this study, we use BEV as the unified data representation.

Given the image coordinates (u, v), we rely on the intrinsic camera parameters and extrinsic parameters to perform transformation from camera reference frame into BEV coordinates. We assume the point where the target touches the ground plane (foot or wheel) is with zero height, then the image coordinates (u, v) can be projected to BEV [10, 27]. However, this assumption can be inaccurate in cases when the camera cannot clearly capture the bottom-most point of an object, the target is jumping in the air, or the ground is not flat. This inaccuracy severely degrades the detection performance when using perspective transformation to project the RGB image to BEV for aligning radar and camera signals. Furthermore, perspective transformation also introduces image distortion because of the zero height assumption. The distortion can severely corrupt objects' visual appearance, such as changing the shape and structure, which causes difficulties for the network to learn from the camera images [18].

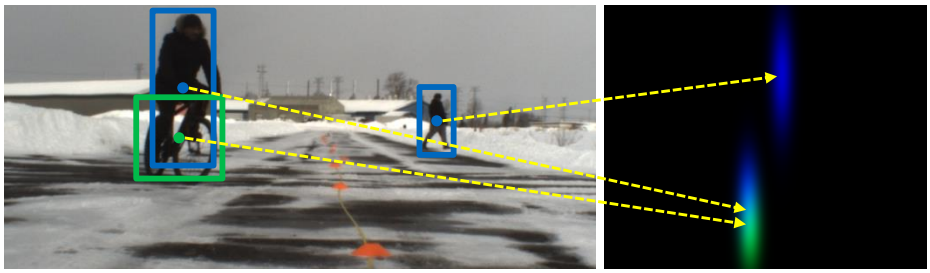

Fig. 3: A camera image with predicted bounding boxes and the cooperative vision feed-forward image in BEV. Blue and green colors represent the pedestrian and bicycle classes. Yellow arrows indicate the mapping after perspective transformation.

Cooperative Vision Feed-Forward In order to address the aforementioned problems, we propose Cooperative Vision Feed-Forward to merge radar data uncertain in azimuth but certain in range with RGB data uncertain in range but certain in azimuth. Despite the uncertainties of radar and camera observations, there is usually a small region on the BEV plane where the radar and camera observations exhibit a strong response, which we use for road user detection.

As depicted in Figure 3, we introduce the detections made by using camera data in the form of a vision feed-forward array on the BEV plane. Specifically, we utilize the off-the-shelf camera-based detection network to produce 2D detections of road users based on the RGB images I_{c}, then we project the center of the predicted bounding boxes of the road users to the BEV to make vision feed-forward images I_{f}. The projected values on the BEV are the confidence scores, and these projected centers can be seen as summarized information from the pre-trained network using RGB images. By doing this, little to no spatial distortion will be caused by the back-projection.

In addition to the projection based on the zero height assumption, we estimate the depth of a target from its width and height in the image frame and its average size (*i.e.*, averaged human heights and shoulder widths) in the real-world to estimate robust depth values. As a result, we use the median value of these estimations as the depth value. To compensate for the projection errors caused by inaccurate perspective transformation, we enlarge the projected positions from points to regions by applying a Gaussian kernel to cover the uncertainty of the depth estimation. Specifically, we never know the actual heights and widths of pedestrians in practice, but instead we use a statistical model of the mean height and width of humans, 1.2m to 2.0m height and 0.36m to 0.44m width, respectively. That means the projection ends up on a segment of a perspective line of x meters, which becomes the sigma of our Gaussian kernel. As long as the sigma is not underestimated, the result will not critically depend on its value. Therefore, the inaccuracy problem caused by perspective transformation can be alleviated.

Spatial Resolution Alignment Efficiently aligning low resolution radar data and high resolution camera signals on the ground plane is a challenge. A straight-

forward solution for aligning is to use a BEV space with the same low spatial resolution as the radar sensor [21]. However, the low spatial resolution of the radar causes this approach to discard high-resolution camera information, negatively affecting the fusion outcomes. On the other hand, upscaling low-resolution radar to align with high-resolution cameras is also not practical because of the high computing effort. Therefore, we introduce Pixel-shuffle [37] Downsampling (PD) to tackle this resolution misalignment problem. PD is to reorganize the high-resolution image data into several low resolution channels. This operation only rearranges the order of the elements to down-scale the image instead of encoding or pooling. Hence, all the information of the image can be preserved, and the network's time and space complexity will not be increased. Specifically, we make vision feed-forward images $I_{\mathrm{f}} \in \mathbb{R}^{N \times H_f \times W_f}$, where $(H_f \times W_f)$ is an intermediate spatial resolution between RGB images and radar cubes. In this instance, we apply PD on the vision feed-forward images I_{f} to obtain $I'_{\mathrm{f}} \in \mathbb{R}^{ND^2 \times H_r \times W_r}$, where D is the down-scaling factor, such that the subsequent network may jointly extract features from the radar tensors and vision feed-forward images with varying resolutions using convolution kernels. The solution has two main advantages: (1) no additional network training effort on camera images is needed, and (2) resolution alignment without information loss.

Camera-Assisted Radar-Based Network The aim of the proposed network is to identify road users by using sequential RAD tensors T_{r} and vision feed-forward images I_{f}. The RAD tensors capture *micro-Doppler patterns* and precise range measurements but have limited azimuth resolution. Our network interprets these tensors to pinpoint and classify various target categories. Conversely, the vision feed-forward images offer detailed azimuth localization of VRUs from a pre-trained camera-based network but lack accuracy in range estimation. Despite the inherent uncertainties, there is a tiny region on the BEV plane where all sensors exhibit a strong response, which we use for road user detection.

Our CARB-Net is developed based on the principles of multi-resolution feature decomposition in U-Net [34] with several improvements. First, we use a small convolution network, which we call the Fusion Encoder, to encode the RF information from the radar and vision feed-forward images into a compact channel-representation as the input for U-Net. In addition, we apply the concept of pixel-shuffle [37] again to replace the max-pooling and up-scaling modules inside U-Net. The feature maps are easily up- or down-scaled by reorganizing the image channels without adding additional model parameters. Compared to deconvolution, max-pooling, or bilinear interpolation, pixel-shuffle is used as an effective way to prevent resolution loss during the encoding and decoding steps.

To obtain high-resolution output, we apply ESPCN [37] and pixel-shuffle up-sampling by factor D to reorganize the channels of the network output. Ultimately, the spatial resolution of the output tensor becomes $(H_f \times W_f)$. With this architecture, we allow the networks to process input with different resolutions.

Context Learning Radars or cameras might not be able to provide recognition of road users in specific contexts, which is unacceptable for real-world deployment. For instance, a stationary object is poorly visible to radars, and adverse weather conditions may make objects hard-to-spot by cameras. However, most existing methods do not focus on such cases during training but only emphasize average detection performance, causing significant performance drops under such contexts compared to favorable circumstances.

In order to improve robustness in these specific contexts, we make the network answer 2 additional auxiliary questions to distinguish the sensing capability of each sensor. For each position in the output confidence map, the network has to determine: (1) is there a VRU or not; (2) **can** a VRU be detected by the radar; and (3) **can** a VRU be detected by the camera. During training, we apply data augmentation and increase the variability of available labeled data in order to cover more combinations of the possible answers to these questions.

We define $y \in \{y_0, y_1, y_2\}$ to represent the 3 correct answers to the 3 questions. Specifically, y_0 indicates whether this position is occupied by a road user. The network should use all the sensor data as input to detect VRUs. y_1 describes if the road user's micro-Doppler pattern is detectable by the radar sensor. In this case, we focus on this specific context that makes the targets poorly visible to radar. Since the targets are stationary or moving without micro-Doppler patterns, the network cannot distinguish their radar signals from other obstacles' reflections. Finally, y_2 indicates whether the pre-trained camera-based network can detect the target or not. In this case, we focus on the conditions that make the camera sensor unable to see the target. By training the network to answer these 3 questions, we use the network's first answer as the final fusion decision to distinguish if the road user is present, and we use the others to describe the specific fusion contexts for each target.

For instance, $[y_0, y_1, y_2] = [1, 1, 1]$ is the common case that both radar and camera can sense the targets, and $[y_0, y_1, y_2] = [1, 0, 1]$ means the road user is only detectable by the camera. This way, our network learns from the correct answers to distinguish these circumstances and understand the scene in more detail. During training, this approach additionally allows us to consider and penalize each class (*e.g.*, pedestrian, cyclist, etc.) independently. This is achieved by tasking the network to perform this classification as well, thus penalizing solutions that fail to interpret their input data or context. That is, for each class, the network is required to output the 3 answers to the 3 questions, which are independent of the other classes' answers.

In order to teach the network to distinguish these specific contexts, we control the input data to generate the corresponding contexts for training. For radar, the ideal way to simulate targets moving without micro-Doppler patterns or stationary targets is to remove their micro-Doppler patterns and retain only the mean velocities. However, this solution is not practical because the radar signals usually contain reflections from other objects and clutter. Moreover, micro-Doppler information from one object can spill over into nearby cells, depending on scene geometry. It is challenging to only remove a certain target's motion information

while not affecting the others. Instead, we propose an alternative way to simulate such cases: the latest RAD tensor is repeated inside the sequential RAD tensors T_{r} to replace the other $N-1$ radar tensors. That is, all the objects in the scene are preserved at their mean velocities, and most of their micro-Doppler patterns are eliminated. So, the objects' motion inside the repeated tensors is seen as an extension of the current movement without introducing any periodic motion. However, there are several downsides to this solution. First, if the target is moving very fast, the radar can still capture its micro-Doppler patterns in a single radar cube. Secondly, repeating the latest radar cube results in radar data where objects are static in RA plane but have non-zero velocity, a situation that is impossible in reality. However, this data augmentation is a close enough simulation of the most important radar limitations: its weakness in detecting slow motion and people with low micro-Doppler variation amplitude. Furthermore, it is easy to implement, and we do not need any additional computing effort to carefully remove the micro-Doppler patterns. For RGB cameras, we switch on and off the cooperative vision feed-forward to simulate the contexts in which the camera sensor cannot detect the target. In this case, the network should only use radar signals to recognize road users.

Anchor Points and Offsets In the output layer, we incorporate anchor points and offsets to precisely refine VRU positions at any resolution. This technique increases the achievable resolution without increasing the data or grid dimensions by using an additional channel of output data to encode highly precise spatial offsets to a target from a low-dimensional grid, low-precision position. We predefine our anchor points over a uniformly spaced range and azimuth positions to allow for an equal capacity for detecting objects at all distances. Since there is a low variation in the size of the VRUs in BEV, we refrain from predicting object width and height. Specifically, we utilize range and azimuth bins from the radar configuration to pre-define the anchor points, representing certain positions on the ground plane. The network predicts 2 offsets for two directions relative to the default anchor point. To ensure that each VRU only be matched by a subset of anchor points, we restrict the offsets to being only inside the 4 surrounding anchor points. Finally, we perform non-maximum suppression to merge nearby anchor points using a greedy approach that favors the highest confidence scores within a pre-defined radius. This is a standard methodology employed in image object detection, which we found also works well in BEV detection using radar.

Loss Functions For a certain position $\mathbf{r}$ in the output tensor, it contains $(3k+2)$ different output values: 2 offsets and $3k$ classes (original k different categories from ground truths and 3 answers for context learning), which results in a total of $(3k+2)H_fW_f$ outputs. We define $O_{c,y}(\mathbf{r})$ that refers to the probability of category c with answer y, where $c \in \{c_0, c_1, ..., c_k\}$ and $y \in \{y_0, y_1, y_2\}$. In order to reduce the inherent class imbalance between the road users and background in the sparsely populated training BEV maps, we apply Focal loss [22] to compute the classification loss $\mathcal{L}_{\mathrm{cls}}$:

$$\mathcal{L}_{\text{cls}}(\mathbf{r}) = -\sum_{c}\sum_{y} \alpha_c[(1-O_{c,y}(\mathbf{r}))^{\gamma} \log\left(O_{c,y}(\mathbf{r})\right) + O_{c,y}(\mathbf{r})^{\gamma}(1-y)\log\left(1-O_{c,y}(\mathbf{r})\right)], \tag{1}$$

where α_c represents the weighting factor for class c, and $\gamma \geq 0$ is the hyper-parameter to down-weigh easy samples. We use smooth L1-norm as our localization loss $\mathcal{L}_{\text{loc}}$. This is similar to the bounding box regression in SSD [23] but we remove the multiple boxes and only predict the offsets for every anchor point.

4 Experiment

4.1 Datasets and Implementation Details

In most public datasets, data representations of radar signals are in the form of a processed point cloud containing targets [2] or projected RD or RA maps [1, 28, 31, 36, 40]. Since our method is based on the micro-Doppler signature to recognize road users, RAD tensors with synchronized camera images are necessary. Hence, we identified the publicly available CARRADA dataset [30] to be best suited for conducting experimental evaluation. In addition, we present results from experimentation using our proposed RadVRU dataset.

CARRADA Dataset The CARRADA dataset [30] includes RAD tensors, RD, RA, and AD maps, all synchronized with camera images. Annotations comprise sparse points, dense masks, and bounding boxes for object detection, tracking, and radar semantic segmentation. Mounted on a stationary vehicle, the hardware captures data in a controlled environment. This dataset encompasses 12k annotated frames across three object categories: pedestrians, cyclists, and vehicles. Each RAD tensor has dimensions of $256 \times 256 \times 64$, and the accompanying camera images boast a resolution of 1232×1028.

Proposed RadVRU Dataset Our RadVRU dataset is captured in several cities in Western Europe. The acquisition hardware consists of a 77 GHz FMCW radar (TI AWR1243), a full HD RGB camera (Realsense D435), and a 3D LiDAR (Ouster OS1-128) mounted on the top of a vehicle. This dataset contains total 78k frames, including 317 sequences of duration from 10 s to 20 s. These frames are representative of the challenging driving scenarios, including different lighting conditions, complex traffic with cars, pedestrians, and cyclists. The annotations are performed by using a camera- and a LiDAR-based object detection model to pre-label the instances, and then these labels are manually corrected by human annotators. The pedestrians are annotated as single points on the ground plane. The dimensions of range, azimuth, and Doppler of every RAD tensor are $128 \times 16 \times 128$, and the resolution of the camera images is 1920×1080.

Table 1: Comparisons with other methods on CARRADA dataset.

Method	Training Data	AP			mAP	mean Dist. Error
		Ped.	Cycl.	Car		
TMVA-Net [29]	Radar	0.44	0.43	0.37	0.37	0.67
PeakConv [25]	Radar	0.45	0.43	0.38	0.38	0.67
ARC [17]	Radar	0.62	0.57	0.50	0.54	0.67
RAMP-CNN [7]	Radar	0.75	0.64	0.50	0.61	0.64
CARB-Net-baseline	Radar	0.77	0.65	0.53	0.63	0.65
U-Net [34]	Radar + Camera	0.57	0.48	0.47	0.51	0.70
FCN [26]	Radar + Camera	0.57	0.43	0.47	0.46	0.69
Lim et al. [21]	Radar + Camera	0.61	0.53	0.51	0.53	0.66
CARB-Net	Radar + Camera	**0.82**	**0.75**	**0.60**	**0.73**	**0.63**

Implementation Details We use the data from CARRADA as an example to explain our implementation details. First, we down-sample the original radar tensor to $128 \times 128 \times 64$ to reduce the high computing effort. Similar to [29], we use 5 sequential radar tensors and camera images as input, so the sizes become $[NC_r, H_r, W_r] = [512, 128, 128]$ and $[NC_c, H_c, W_c] = [15, 1232, 1028]$. The cooperative vision feed-forward is made by a pre-trained "yolov8m" [12] model. We filter the detections from YOLOv8 with a confidence threshold of 0.1, and project the detections to the size of vision feed-forward images $[N, H_f, W_f] = [5, 256, 256]$. After pixel-shuffle down-sampling by factor $D = 2$, it becomes $[ND^2, H_r, W_r] = [20, 128, 128]$. Similarly, after the ESPCN [37], we apply pixel-shuffle up-sampling by factor $D = 2$ to restore the resolution $[256, 256]$. For the proposed anchor offsets and context learning, we have $(3k + 2) = 11$ predictions. Hence, the size of the final output tensor is $[11, 256, 256]$. We use 0.8 meter as the radius of our non-maximum suppression algorithm. Please refer to the Supplementary Materials for more implementation details and network architecture.

4.2 Experiment Results

Evaluation Metric Our experiments employ Average Precision (AP) and mean distance error as primary evaluation metrics. We define true positives as detections falling within a 1.5-meter boundary around the ground truth. We avoid multiple matches within the same boundary. Additionally, we measure localization precision by calculating the distance between true positives and ground truths in meters. Averaging these distance errors yields the mean distance error, providing an evaluation of localization accuracy.

Evaluation on CARRADA Dataset We compare the object detection performance of our proposed method and the other state-of-the-art approaches on the CARRADA dataset. Since some of the previous methods output semantic segmentation masks instead of detections, we carefully find the centers of the predicted masks as their localization results. For FCN [26] and U-Net [34] baselines, we use inverse projection mapping to project the RGB images to BEV

Table 2: Comparisons with other methods on our proposed RadVRU dataset.

Method	Training Data	mAP	mean Dist. Error
ARC [17]	Radar	0.51	0.64
CARB-Net-baseline	Radar	0.55	0.62
U-Net [34]	Radar + Camera	0.52	0.87
FCN [26]	Radar + Camera	0.35	0.94
Lim et al. [21]	Radar + Camera	0.67	0.63
EchoFusion* [24]	Radar + Camera	0.69	0.54
CARB-Net	Radar + Camera	**0.81**	**0.47**

down to a cell resolution of 128×128. Then, we concatenate the projected images and radar tensors along the channel and output the BEV map. For fair comparisons, we ensure the resolutions of the radar data and projected image are the same for all methods. Furthermore, we remove the cooperative vision feed-forward and train our CARB-Net-baseline model only on the radar tensors to compare with the radar-based methods. Although our method is designed for VRUs, we still involve cars for training and testing. We use the cars' center points in BEV space as the ground truth for detection. The vehicles' orientations are not considered during evaluation by all the methods. As observed in Tab. 1, compared to the others, our proposed framework can not only clearly improve the detection performance but also achieve a smaller mean distance error on the RA plane. The results demonstrate that the proposed method successfully integrates the detections from camera images to achieve better accuracy.

Evaluation on Proposed RadVRU Dataset Similarly, we conduct the same experiment on our RadVRU dataset. We evaluate our framework against the other four different models. For EchoFusion* [24], we use RA tensors as their input and re-implement their method with reduced network size due to our limited computing hardware. As illustrated in Tab. 2, our proposed approach also outperforms the other methods by a significant margin.

Visualization In Fig. 4, we present camera images with bounding boxes predicted by YOLOv8 alongside the radar-based model ARC's [17] and our method's detections in BEV. Notably, our detections outperform ARC's, exhibiting fewer false positives and negatives, leading to superior detection performance. Even in low-light conditions, where YOLOv8 fails, our model successfully identifies pedestrians using radar data alone. Additionally, when road users are stationary, ARC fails due to the absence of micro-Doppler patterns, whereas our method utilizes the vision feed-forward to accurately identify pedestrians.

4.3 Ablation Study

Effects of Proposed Context Learning To further investigate the impact of the proposed context learning, we conduct several experiments to show the robustness of our model when one of the sensors fail. In Tab. 3, the first row model

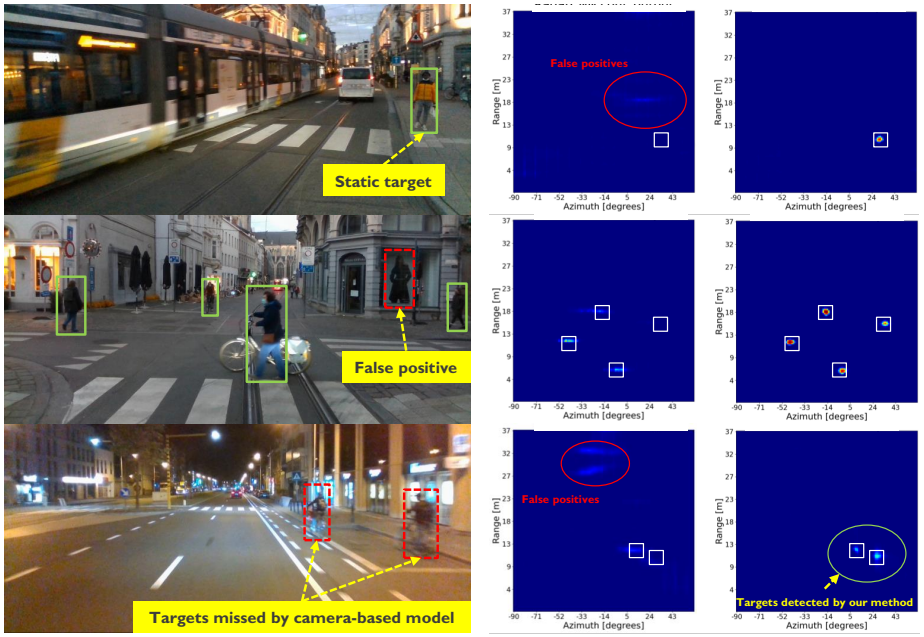

Fig. 4: Visual comparisons of the detection results for three different scenarios in our proposed dataset. Green boxes in (a) represent the true positives detected by camera-based model. Red boxes/circles indicate false positives and negatives, and white boxes denote ground truth locations. Our model (c) exhibits enhanced detection performance compared to ARC [17] (b), integrating radar and camera data effectively.

Table 3: Effects of proposed context learning.

Method	Training Data	mAP w/ Radar Only	mAP w/ Camera Only	mAP w/ Radar & Camera
CARB-Net-baseline	Radar	**0.55**	-	-
CARB-Net wo/ Context learning	Radar + Camera	0.10	0.45	0.75
CARB-Net w/ Mix training	Radar + Camera	0.40	0.52	0.75
CARB-Net w/ Context learning	Radar + Camera	0.54	0.65	**0.81**

is only retrained on radar data without cooperative vision feed-forward signals. The purpose of this model is to show the upper bound of the detection performance when there is only radar data available. The second row model is retrained without context learning. The third row model is retrained by a straightforward solution to consider sensor failure cases during training. We randomly dropout the radar or camera from the input to mix-train the model, which forces the model to learn how to detect in only one modality.

The model without context learning shows a significant performance drop when the camera-based model cannot detect the targets. In other words, this model is sensitive to the input and is not able to perform detection when one

Table 4: Effects of proposed components.

Method	Vision Feed-Forward	Improved U-Net	Anchor Offsets	Context Learning	mAP
U-Net [34]	-	-	-	-	0.53
CARB-Net	✓	-	-	-	0.65
	✓	✓	-	-	0.70
	✓	✓	✓	-	0.75
	✓	✓	✓	✓	**0.81**

sensor is not working as expected. The third model shows marginally improved results when the one of the sensor is unable to detect, but the detection performance in nominal cases are affected. In contrast, our proposed method shows significant improvements, no matter in nominal or sensor failure cases.

Effects of Proposed Components To clarify the effect of every proposed component, we conduct several experiments to show the performance difference with or without these components. Tab. 4 demonstrates that all the proposed components contribute to performance improvements. The most notable improvement is brought about by the cooperative vision feed-forward. When it is considered, the model achieves a greater level of performance. The projection of the predicted bounding boxes provided by the off-the-shelf camera-based model not only reduces the training effort but also prevents the distortion problem. The results show that when all the components are incorporated, the performance of the model is considerably enhanced.

5 Conclusion

In this paper, we introduce Camera-Assisted Radar-Based Network (CARB-Net), a novel radar-based framework aimed at addressing the low angular resolution of radar signals and enhancing robustness to conditions where a sensor is unable to detect the targets. CARB-Net effectively integrates detections from an off-the-shelf camera-based object detection model with radar data, enabling precise localization by learning object micro-Doppler patterns from radar tensors and incorporating vision feed-forward signals. Our context learning method ensures reliable performance even when one sensor fails. The paper contributes significantly by showcasing improved performance for vulnerable road users compared to state-of-the-art methods. Validation on our larger radar-camera dataset confirms these achievements. Additionally, our method demonstrates up to a 12% improvement in mAP performance across diverse datasets, with ablation studies highlighting the efficacy of each framework component. Future research directions could explore optimizing fusion mechanisms and incorporating additional sensor modalities like thermal or acoustic sensors for even more accurate object detection and localization. However, as more sensors are integrated, ensuring robustness in multi-sensor object detection remains critical for real-world applications focused on enhancing road user safety.

Acknowledgements

This work was funded by EU Horizon 2020 ECSEL JU research and innovation programme under grant agreement 876487 (NextPerception), by EU Horizon KDT JU research and innovation programme under grant agreement 101139769 (Distrimuse), and by the Flemish Government (AI Research Program).

References

1. Barnes, D., Gadd, M., Murcutt, P., Newman, P., Posner, I.: The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. In: ICRA (2020)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
3. Chen, S., He, W., Ren, J., Jiang, X.: Attention-based dual-stream vision transformer for radar gait recognition. In: ICASSP. IEEE (2022)
4. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: CVPR (2017)
5. Dimitrievski, M., Shopovska, I., Van Hamme, D., Veelaert, P., Philips, W.: Weakly supervised deep learning method for vulnerable road user detection in fmcw radar. In: IEEE International Conference on Intelligent Transportation Systems (ITSC). IEEE (2020)
6. Gao, X., Xing, G., Roy, S., Liu, H.: Experiments with mmwave automotive radar test-bed. In: 2019 53rd Asilomar conference on signals, systems, and computers. pp. 1–6. IEEE (2019)
7. Gao, X., Xing, G., Roy, S., Liu, H.: Ramp-cnn: A novel neural network for enhanced automotive radar object recognition. IEEE Sensors Journal (2020)
8. Griffiths, H., Cohen, L., Watts, S., Mokole, E., Baker, C., Wicks, M., Blunt, S.: Radar spectrum engineering and management: Technical and regulatory issues. Proceedings of the IEEE (2014)
9. Gusland, D., Christiansen, J.M., Torvik, B., Fioranelli, F., Gurbuz, S.Z., Ritchie, M.: Open radar initiative: Large scale dataset for benchmarking of micro-doppler recognition algorithms. In: IEEE Radar Conference (RadarConf). pp. 1–6 (2021)
10. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
11. Hwang, J.J., Kretzschmar, H., Manela, J., Rafferty, S., Armstrong-Crews, N., Chen, T., Anguelov, D.: Cramnet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection. In: ECCV (2022)
12. Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (2023), `https://github.com/ultralytics/ultralytics`
13. Kim, Y., Kim, S., Choi, J.W., Kum, D.: Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer. In: AAAI (2023)
14. Kim, Y., Shin, J., Kim, S., Lee, I.J., Choi, J.W., Kum, D.: Crn: Camera radar net for accurate, robust, efficient 3d perception. In: ICCV (2023)
15. Kopp, J., Kellner, D., Piroli, A., Dietmayer, K.: Tackling clutter in radar data–label generation and detection using pointnet++. arXiv preprint arXiv:2303.09530 (2023)

16. Lee, W.Y., Dimitrievski, M., Jovanov, L., Philips, W.: Spatio-temporal consistency for semi-supervised learning using 3d radar cubes. In: IEEE Intelligent Vehicles Symposium (IV). IEEE (2021)
17. Lee, W.Y., Jovanov, L., Kumcu, A., Philips, W.: Arc: Automotive radar consistency regularization for semi-supervised learning. IEEE Transactions on Intelligent Vehicles (2023)
18. Lee, W.Y., Jovanov, L., Philips, W.: Multi-view target transformation for pedestrian detection. In: WACV Workshops (2023)
19. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: AAAI (2023)
20. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV (2022)
21. Lim, T.Y., Ansari, A., Major, B., Fontijne, D., Hamilton, M., Gowaikar, R., Subramanian, S.: Radar and camera early fusion for vehicle detection in advanced driver assistance systems. In: NIPS Workshops (2019)
22. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
23. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV (2016)
24. Liu, Y., Wang, F., Wang, N., ZHANG, Z.X.: Echoes beyond points: Unleashing the power of raw radar data in multi-modality fusion. NeurIPS (2024)
25. Liwen, Z., Xinyan, Z., Youcheng, Z., Yufei, G., Yuanpei, C., Xuhui, H., Zhe, M.: Peakconv: Learning peak receptive field for radar semantic segmentation. In: CVPR (2023)
26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
27. Mallot, H.A., Bülthoff, H.H., Little, J.J., Bohrer, S.: Inverse perspective mapping simplifies optical flow computation and obstacle detection. Biological cybernetics (1991)
28. Meyer, M., Kuschk, G.: Automotive radar dataset for deep learning based 3d object detection. In: European Radar Conference (EuRAD) (2019)
29. Ouaknine, A., Newson, A., Pérez, P., Tupin, F., Rebut, J.: Multi-view radar semantic segmentation. In: ICCV (2021)
30. Ouaknine, A., Newson, A., Rebut, J., Tupin, F., Pérez, P.: Carrada dataset: Camera and automotive radar with range- angle- doppler annotations. In: ICPR (2021)
31. Paek, D.H., Kong, S.H., Wijaya, K.T.: K-radar: 4d radar object detection for autonomous driving in various weather conditions. NeurIPS (2022)
32. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. pp. 194–210. Springer (2020)
33. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. NIPS (2017)
34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer (2015)
35. Schumann, O., Hahn, M., Dickmann, J., Wöhler, C.: Semantic segmentation on radar point clouds. In: International Conference on Information Fusion (FUSION) (2018)

36. Sheeny, M., De Pellegrin, E., Mukherjee, S., Ahrabian, A., Wang, S., Wallace, A.: Radiate: A radar dataset for automotive perception in bad weather. In: ICRA (2021)
37. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR (2016)
38. Vaizman, Y., Weibel, N., Lanckriet, G.: Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (2018)
39. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: CVPR (2020)
40. Wang, Y., Wang, G., Hsu, H.M., Liu, H., Hwang, J.N.: Rethinking of radar's role: A camera-radar dataset and systematic annotator via coordinate alignment. In: CVPR (2021)
41. Yao, S., Guan, R., Huang, X., Li, Z., Sha, X., Yue, Y., Lim, E.G., Seo, H., Man, K.L., Zhu, X., Yue, Y.: Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. IEEE Transactions on Intelligent Vehicles (2023)
42. Zhou, T., Chen, J., Shi, Y., Jiang, K., Yang, M., Yang, D.: Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection. IEEE Transactions on Intelligent Vehicles (2023)