Cross-Modal Supervision-Based Multitask Learning With Automotive Radar Raw Data

Yi Jin, *Graduate Student Member, IEEE*, Anastasios Deligiannis, *Member, IEEE*, Juan-Carlos Fuentes-Michel, and Martin Vossiek, *Fellow, IEEE*

***Abstract*—With the rapid development of autonomous driving technology, radar sensors play a vital role in the perception system due to their robustness under harsh environmental conditions, exact range and velocity perception capability. However, the state-of-the-art performance of algorithms solely based on radar to achieve various perception tasks, such as classifying road users and infrastructures, still lags far behind expectation. Their failure can mainly be accounted for the extreme sparseness of radar point cloud for objects, low angular resolution, and the issue of ghost targets. In this work, we propose a novel network that employs the complex range-Doppler matrix as input to achieve radar-tailored panoptic segmentation (i.e., *free-space segmentation* and *object detection*). Our network surpasses previous works in *free-space segmentation* and *object detection* tasks, and the improvement in the former task is especially notable. During training, the segmented camera image with radar customized adaption is utilized as the ground truth. Through such a cross-modal supervision method, the labeling expense is alleviated considerably. Based on it, we further design an innovative camera-radar system concept that is able to automatically train deep neural networks with radar measurement.**

***Index Terms*—Machine learning, cross-modal supervision, multitask learning, radar, raw data, free-space segmentation, object detection.**

I. Introduction

NOWADAYS, autonomous driving (AD) technology has attracted enormous attention from academia and industry due to its disruptive potential to the economy and society [1]. To achieve AD, a well-accepted concept regarding environment perception is using multiple sensor modalities, such as camera, radar, and lidar, to improve system redundancy [2]. Based on this consensus, different modality sensors must be able to independently complete various perception tasks, such as detecting different kinds of road users and understanding driving-related, critical infrastructure information.

With the boom in machine learning, more specifically deep learning (DL), camera-based or lidar-based driving systems have achieved huge development, but they are sensitive to the weather and illumination conditions. In addition, it is extremely challenging for these systems to distinguish moving and static objects with a single measurement. In contrast, their radar-based counterpart is advantageous because it is capable to work robustly under harsh environmental conditions and simultaneously measure the precise relative radial velocity of objects. However, the common radar sensor is not on par with other two modalities sensors, considering its angular resolution. Concretely, current automotive-grade radar can usually achieve the best angular resolution at about 1° [3], which is much worse than lidar or camera. Also, radar data are much noisier and have many ghost targets [4]. As a result, radar usually has a low confidence level in the entire perception system to avoid the frequent false alarm issue. Furthermore, the radar sensor does not have a uniform data format. It can be analog-to-digital converter (ADC) signal, range-azimuth-Doppler tensor (or called radar cube), or point cloud data. In addition, compared with the camera or lidar, the release of the public dataset containing radar measurement is still limited. And the availability of radar cube data is even harder, which, owing to its abundant information, is believed to have tremendous potential for perception tasks. All these reasons hinder the development of radar-based perception systems.

In this work, we utilize the camera image and radar range-Doppler matrix from the freshly released RADIal dataset [5] to address two key perception tasks, namely *free-space segmentation* and *object detection*. Since most road users in this dataset are various moving vehicles, the scope of *object detection* is limited to vehicle class. However, given an appropriate dataset, the extension of our work to other road users is also feasible.

For supervised learning, access to labeled data is the prerequisite. Usually, manual labeling, which is time consuming and labor intensive, is adopted to the radar point cloud like in [6]. However, when it comes to radar cube data, manual labeling is rather challenging or even infeasible since this data is not well understandable for humans. To solve the labeling issue for radar data, we pay attention to the camera image, which is time synchronized and spatially overlapped with radar measurement. We adopt the innovative, cross-modal training methodology from [7] and extend its labeling range from sole free space (semantic level) to free space and road users (panoptic level). More concretely, panoptic segmentation masks are generated from the camera image, and it means the ground truth and

Manuscript received 28 September 2022; revised 8 November 2022 and 16 December 2022; accepted 27 December 2022. Date of publication 6 January 2023; date of current version 19 May 2023. This article was produced by the IEEE Publication Technology Group, Piscataway, NJ, USA. *(Corresponding author: Yi Jin.)*

Yi Jin and Martin Vossiek are with the Institute of Microwaves and Photonics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Bavaria, Germany (e-mail: yi.y.jin@fau.de; martin.vossiek@fau.de).

Anastasios Deligiannis and Juan-Carlos Fuentes-Michel are with the BMW Group, 80788 Munich, Germany (e-mail: anastasios.deligiannis@bmw.de; juan-carlos.fuentes-michel@bmw.de).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2023.3234583.

Digital Object Identifier 10.1109/TIV.2023.3234583

final output from our model are in the camera perspective instead of bird's eyes view (BEV) of the radar. By this training methodology, we can employ radar range-Doppler matrix to address the infrastructure classification task like [7] and the more challenging moving road user detection. The segmented camera images, which can be automatically generated by an off-the-shelf panoptic segmentation network like [8] provide supervision signals for the training of radar data. Since all the steps from ground truth preparation to training the deep neural network (DNN) can be achieved automatically, we proposed a camera-radar system concept that can automatically utilize radar data to train DNN with camera's supervision signal.

Based on our endeavor, we find that handling road users as free space is not a wise option. Therefore, we adjust the segmented image to realize the radar-tailored panoptic segmentation, which can be decomposed into *free-space segmentation* and *object detection* task. The performance of our network overpasses previous works in terms of *free-space segmentation* and *object detection* tasks, and the improvement in the former task is significantly notable.

Our contribution in this work can be summarized in the following three aspects:

1) We extend and adjust the camera image-based training methodology to road users. Thus, a radar-tailored panoptic segmentation in the camera image is achieved.
2) Based on this training methodology, we propose a camera-radar system concept that can automatically utilize segmented image data as supervision signal to train DNN with radar data.
3) We design and analyze our novel radar-only model, which addresses *free-space segmentation* and *object detection* problems, and it surpasses the performances of previous works.

This work is organized as follows: Chapter II introduces some background information related to our work; Chapter III talks about how we prepare the ground truth data; from Chapters IV to VII, we elaborate the concrete work like the cross-modal training methodology, camera-radar system concept, and our network as well as results and analysis; in Chapter VIII, we discuss the issue and future work; Finally, we conclude the entire work in Chapter IX.

II. Related Work

A. Radar Signal-Processing Chain

The signal-processing chain for the state-of-the-art frequency-modulated continuous wave (FMCW) automotive radar can generally be divided into two steps: preprocessing and detection list generation. In the preprocessing step, the baseband samples (ADC) as input for the entire signal-processing chain first undergo twice fast Fourier transform (FFT) along the fast-time and slow-time axes, respectively, and the result is called range-Doppler matrix. Then, beamforming algorithm is applied to the range-Doppler matrix to transform it into the radar cube. For a usual 2D imaging radar, its radar cube has three dimensions, namely range, Doppler, and azimuth, while its counterpart, 3D radar, further has elevation dimension. In the detection generation step, the constant false alarm rate (CFAR) [9] is employed to detect peaks in the radar cube, and the output is basically the detection list (point cloud data), including radar cross section (RCS) channel.

B. Radar Grid Map Generation

After signal processing, the detection list is available. Using the Doppler velocity, ego-motion and radar mounting position, we can distinguish static and dynamic detection. The dynamic detection is usually filtered out while the static counterpart is preserved for the grid map. To generate the grid map, static detections from different measurement moments are accumulated. This way, the shape and energy information of the infrastructure on the map become clear. Generally, there are several types of radar grid maps, for example, occupancy grid map and RCS grid map [10].

C. Image Segmentation and Object Detection

Image segmentation has been a crucial computer vision task for a long time. Along with the boom in DL, image segmentation has also developed from semantic to instance segmentation and was finally unified by panoptic segmentation. Semantic segmentation assigns a class information to each pixel in the image and usually utilizes a fully convolutional network like [11] with a dilation structure [12]. However, a fatal weakness of semantic segmentation is that it treats everything as stuff. In other words, object instances within the same class are not differentiable, which is not enough for AD scenarios.

In contrast, instance segmentation or object detection can be adopted to handle objects. Mask R-CNN [13] is a classic network of instance segmentation that utilizes a bounding box or segmentation mask to denote an object. A bounding box is used for object detection to represent the object's size and position. Generally, object detection networks can be classified into the following two categories: region based [14] and anchor based [15], [16]. However, neither instance segmentation nor object detection processes the stuff information like free space. Finally, panoptic segmentation [8] realizes the reconciliation between object and stuff. Concretely, each pixel in the image is assigned with a semantic label and an instance ID. Pixel with the same label and ID is classified as one object, and the instance ID is neglected for stuff. Naturally, panoptic segmentation offers the opportunity to utilize one DNN to handle not only the infrastructure related but also road user related information in AD scenarios.

D. Cross-Modal Supervision

To apply supervised learning-based DL algorithms to radar data, the ground truth must be provided. The authors from [17] leveraged manual labeling to generate labels for the radar point cloud. However, serious flaws for this method are time consuming and labor intensive. Furthermore, when dealing with the radar range-Doppler-azimuth tensor, manual labeling fails because the presentation of the range-Doppler-azimuth tensor is not well understood by annotators.

In light of these issues, the cross-modal supervision approach becomes a good solution. For cross-modal supervision, apart from the goal modality data, data from one or multiple other modalities are involved to generate labels for the goal modality. For instance, [18] employed radar detection as supervision signal for image-based object detection. Authors from [19], [20], [21] utilized camera and lidar measurements to form the ground truth for radar signals. In contrast, apart from the depth estimation, *free-space segmentation* and *object detection* as the main part of our work require only supervision signal from a camera, which is easy and cheap to achieve. Furthermore, a joint extrinsic calibration between the camera and radar is no longer necessary. Considering these, our method is fairly advantageous.

E. RADIal Dataset

The RADIal dataset [5] is a freshly released dataset that contains synchronized camera, lidar, and radar measurements and includes three scenarios: urban roads, countryside, and highways. The entire measurements of the RADIal last 2 hours and are divided into 91 sequences of 1 to 4 min duration. Among a total of 25000 synchronized frames, 8252 frames are labeled containing all together 9550 vehicles, which are the most road users in the dataset.

Its radar is a prototype high-definition (HD) front radar with 12 transmitter antennas and 16 receiver antennas. Different from the frequently adopted time-division multiplexing method, this HD radar leverages Doppler-division multiplexing technology. To better understand this technology, assuming that the radar detects a vehicle, the signal of this vehicle appears N_{tx} times on the range-Doppler matrix of a single receiver channel, where N_{tx} means the number of transmitter antennas. Mathematically, these signals appear at $(R, (D + n \times \delta) \bmod D_{\max})_{n=1,\ldots,N_{\text{tx}}}$ positions on the range-Doppler matrix, where R and D are the range and Doppler velocity of the vehicle, respectively, δ is 16, and $D_{\max}$ means the largest measured Doppler value. Furthermore, this HD radar is alleged to have $0.2\,\text{m}$ range resolution, 0.1° and 1° angle resolution in azimuth and elevation directions, respectively. In addition to the radar point cloud data, this dataset provides range-Doppler-azimuth tensors, which can significantly boost the radar community to research and develop new algorithms based on raw data. The camera is placed below the interior mirror behind the windshield and has 5 Mpix resolution. For more details, readers can refer to [5].

III. Data Preparation

In this work, we use all the labeled data provided by RADIal [5], which includes 8252 frames synchronized radar, lidar, and camera measurements from urban road, countryside, and highway. In order to generate the radar grid map, we obtain the ego-motion for the labeled data from separate measurements based on the labeling information. The ego-motion is inevitable; otherwise, the differentiation between static and dynamic detection and the accumulation processing for grid map can not be conducted. In addition, we keep the original data division in RADIal for training, validation and test. Therefore, a single measurement trace only appears in the training or testing.

To generate the segmented images, we first take the pre-trained panoptic feature pyramid network (FPN) [8] from Detectron2 [22] and then use the Cityscapes dataset [23] to fine-tune it. Considering the measurement scenarios in the RADIal dataset, all the labels in the Cityscapes are simplified to the vehicle in the *thing* class, road and nonroad in the *stuff* class. Herein we call them *vehicle*, *free space*, and *nonfree space*. For the *vehicle* category, an instance level annotation (i.e., semantic label, bounding box, and instance ID) is achieved, while *free space* and *nonfree space* have only semantic labels.

Apart from segmentation labels, three heatmaps, namely the keypoint prediction, offset, and depth heatmaps, are prepared for object detection. Since most road users in RADIal are various vehicles, the terminology *object detection* targets vehicles, and *keypoint* is defined in this work as the center point of a bounding box for the vehicle. Considering the processing efficiency, the size of heatmaps is four times smaller than the original image (256×512). For each ground truth keypoint $\boldsymbol{p} \in \mathcal{R}^2$, its position on the heatmap is formulated as:

$$\widetilde{\boldsymbol{p}} = \lfloor \boldsymbol{p}/4 \rfloor \tag{1}$$

As this position must be an integer, the tilde in $\widetilde{\boldsymbol{p}}$ represents the ceiling operation. In addition, on the keypoint prediction heatmap $\boldsymbol{M}_{\text{key}} \in [0,1]^{64\times128}$, we utilize a Gaussian kernel v_{rc} to expand keypoints. The formula of Gaussian kernel is the following:

$$v_{rc} = \exp\left(-\frac{(r-\widetilde{p}_r)^2 + (c-\widetilde{p}_c)^2}{2\sigma_p^2}\right) \tag{2}$$

where r and c are row and column indices, respectively, on the map, and σ_p is a size-adaptive standard deviation. Since the size of a vehicle in the image is largely influenced by its distance to the camera, after several tests, we leverage (3) to determine σ_p, and d is the distance of the keypoint.

$$\sigma_p = \begin{cases} 25 \ \text{if} \ d <= 10\,\text{m} \\ 25 - d/2 \ \text{if} \ 10\,\text{m} < d <= 50\,\text{m} \\ 5 \ \text{if} \ d > 50\,\text{m} \end{cases} \tag{3}$$

If two Gaussian kernels overlap, we keep the element-wise larger value in the heatmap. Since the keypoint detection heatmap is four times smaller than the original input, an offset heatmap is employed to estimate the position offset between them. Specifically, the offset heatmap contains the row and column differences $\boldsymbol{\Delta} = \boldsymbol{p} - \widetilde{\boldsymbol{p}}$ for a keypoint. In addition, the RADIal dataset leverages a lidar-based object detection method to automatically obtain the 3D position of vehicles. We calculate the range d from its 3D position and save the value at $\widetilde{\boldsymbol{p}}$ in the depth heatmap for each keypoint.

IV. Proposed Method

A. Cross-Modal Supervision

To apply supervised learning-based DL algorithms, labeled data must be prepared in advance. Manual labeling is a frequently adopted method for data such as camera images; however, it consumes massive time and labor. Regarding radar data, the labeling is much more challenging than that for camera images, because radar data is much noisier. Frequently, ghost targets caused by factors such as multipath reflection appear with the real targets together. Consequently, it burdens the annotators

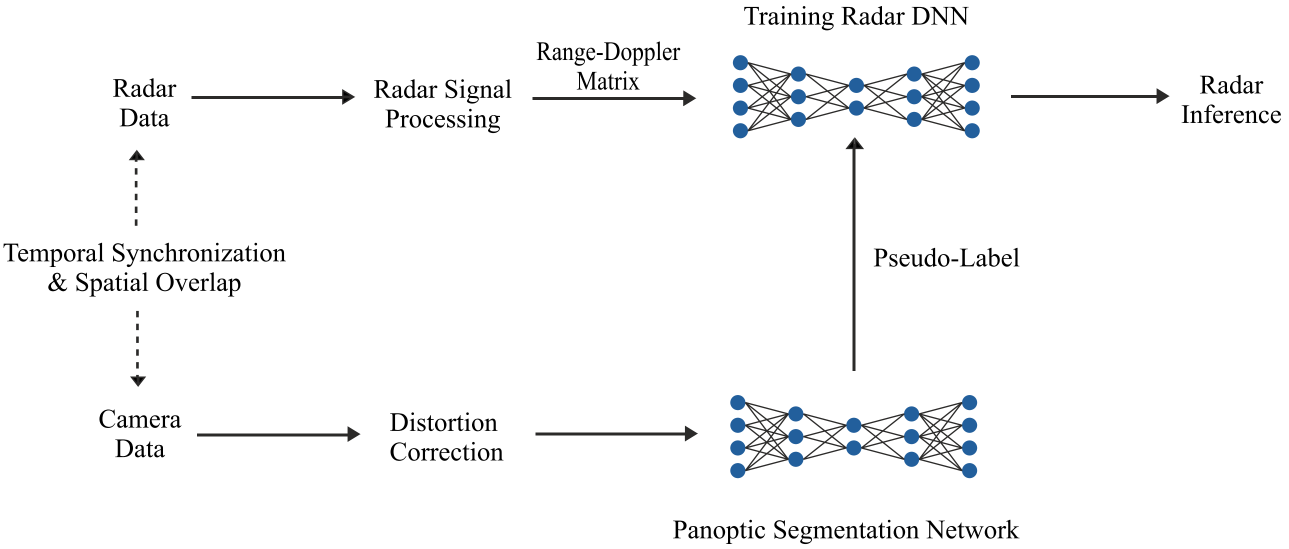


Fig. 1. Cross-modal supervision-based workflow. After distortion correction, the segmented camera images are provided as pseudo labels for the training of Radar DNN.

and, at the same time, affects the labeling quality if ghost targets are not correctly recognized. In addition, the radar cube data representing objects in range, Doppler, and angle (azimuth or even elevation) dimensions is not simply understandable for human beings. Considering this representation for objects and the complexity of radar cube data, consensus regarding labeling in radar cube is still not available. Although [24] proposes a semi-automatic annotation approach that relies on camera images to label radar data, its application is limited to single vehicle or person measurement scenarios. Moreover, after the generation of automatic annotation, the annotators must frame to frame check the annotation result, which is still time consuming and labor intensive.

Considering these issues, we propose a fully automatic annotation method that is entirely immune to the radar's ghost target issue and can annotate not only objects at the instance level but also stuff at the semantic level. The workflow of this annotation method and the training are illustrated in Fig. 1. The following conditions should be kept in mind: the input for annotation should be the images from a camera that is time synchronized and spatially has a common detection area with radar. First, the image distortion is corrected; then, a panoptic segmentation network, such as panoptic FPN [8], EfficientPS [25], can be applied to segment undistorted images. In our work, we adopt panoptic FPN and utilize Cityscapes dataset to fine-tune it further. The segmented image with radar customized adaption is employed as a pseudo label for the training of Radar DNN. The panoptic segmentation result for instance class like different kinds of vehicles is satisfactory, while for stuff like *free space* and *nonfree space* is sometimes not completely right. To fix this issue, manual labeling part of images from RADIal and then fine-tuning should be a good solution. However, as proof of concept, it is totally acceptable to directly use the output from panoptic FPN as supervision signal for radar in our work.

For radar data, the ADC signal first undergoes several signal processing operations, such as window function and FFT along fast-time and slow-time axes. The generated range-Doppler matrix is delivered to a radar DNN to achieve a radar-tailored panoptic segmentation. Since the panoptic segmentation can simultaneously be decomposed into object detection and segmentation of stuff, a multitask loss function entailing detection and segmentation is necessary. The details regarding radar-tailored panoptic segmentation, radar DNN, and multitask loss are elaborated in Sections IV-C, IV-D, and V, respectively.

B. Camera-Radar System Concept

Based on the cross-modal supervision-based training methodology in Section IV-A, we further design a novel camera-radar system concept illustrated in Fig. 2. This camera-radar system can achieve automatic DNN training with radar data under camera's supervision signal and adaptive camera-radar fusion at the image plane. To ensure that the system works normally, several requirements must be satisfied. In principle, the requirements are the same for the cross-modal training methodology in Fig. 1, that is, both sensors should spatially have a common field of view (FoV) and should be well time synchronized.

Generally, there are two phases in this system. The first phase is basically the training phase depicted in Fig. 2(a). In this stage, the camera system provides the ground truth for radar training. Since current camera-based segmentation algorithms are well developed, networks like [8] or [25] can be adopted in this system to achieve panoptic segmentation. The segmented image is utilized as ground truth for the DNN training with radar data. Considering that the camera system cannot work robustly under bad weather or worse illumination conditions, the training data should mainly be collected under good working conditions for the camera. And in this stage, the prediction results from the camera are much more reliable than those from the radar; thus, we mainly trust the camera. After training the radar with enough data, and its prediction result is on par with the camera, it comes to the second phase, that is, (b) and (c), when the system will decide to trust only one or both sensors based on the given situation. In this phase, the system is actually in the

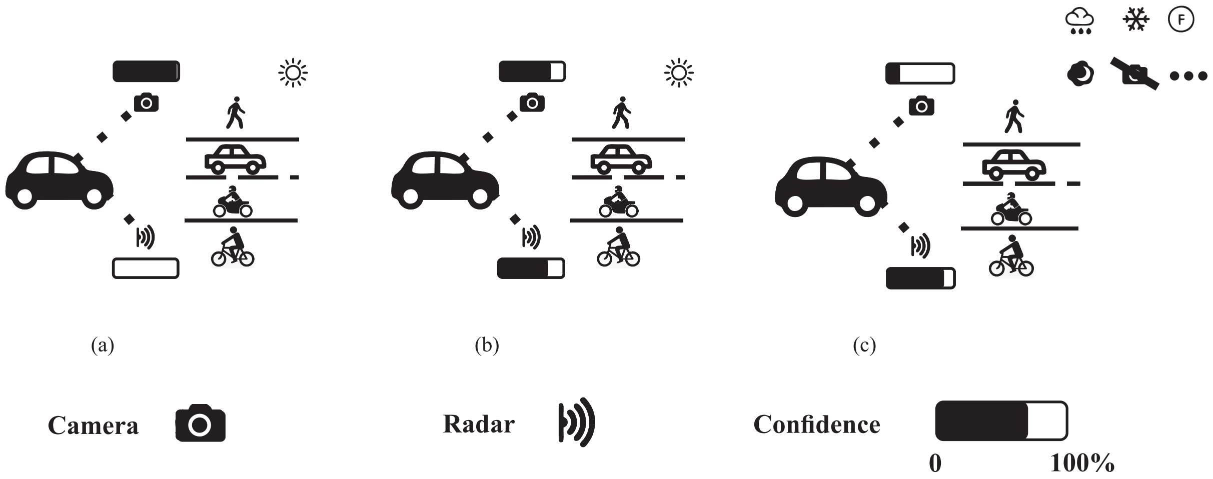


Fig. 2. The camera-radar system concept. In training phase (a), in the camera suitable working condition, the system mainly trusts the camera and trains DNN with the radar data under the camera's supervision signal. In fusion phase (b) and (c), the system adaptively fuses two modality data depending on the given situation, such as weather or illumination conditions.

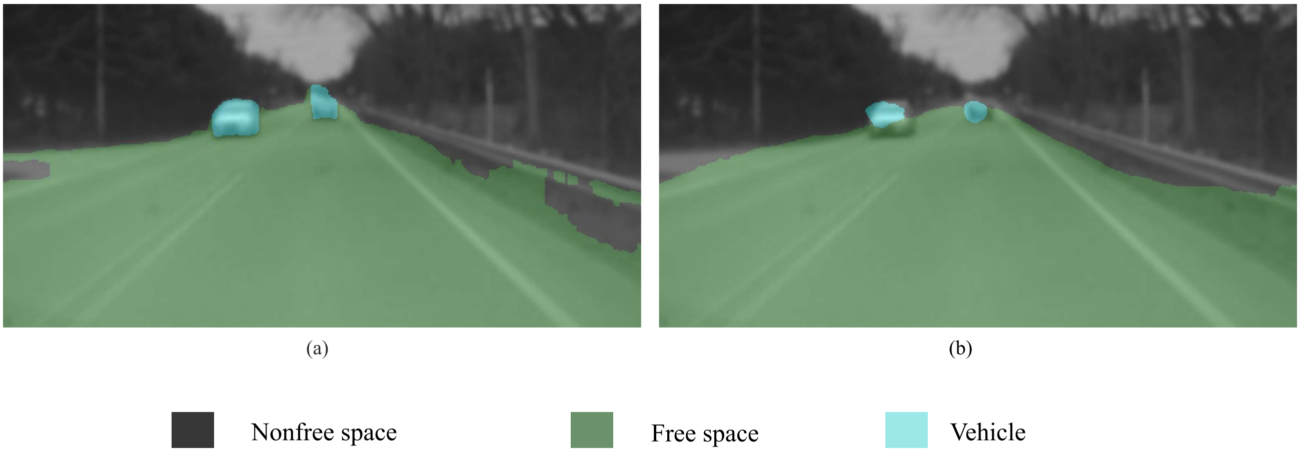


Fig. 3. Semantic segmentation results. (a) is from the fine-tuned panoptic FPN as the ground truth, and (b) is the prediction result from a model trained with radar range-Doppler matrix. The camera image is taken directly from the RADIal dataset and has a limited resolution.

fusion working mode. For (b), under camera-suitable working conditions, results from the camera and radar are considered. In contrast, for (c), under adverse weather conditions like fog, rain, or snow, or when the camera works abnormally, the system trusts the radar more than the camera.

In this work, we mainly focus on the phase (a) that leverages vision supervision to train the radar DNN. As a proof of concept, we simplify the problem by assuming the confidence of the vision supervision as 1. Considering the generally precise image segmentation from our fine-tuned model, this simplification is totally acceptable. The fusion phase shown in Fig. 2(b) and (c) is not intensively handled in this work, but it can be regarded as promising direction for future work.

C. Semantic Segmentation With Radar Range-Doppler Matrix

In this section, we first describe the trial to conduct semantic segmentation on the image plane with radar range-Doppler matrix since it can be viewed as a prerequisite for panoptic segmentation. Although the trial succeeded, based on the analysis, we find it less meaningful to directly apply pixel-to-pixel panoptic segmentation like the image to radar data. As a solution, we propose a novel adaption to handle moving objects.

In this trial, sole semantic level annotations are adopted. Specifically, there are three classes *free space*, *nonfree space*, and *vehicle* assigned to all the image pixels.

One example of this trial is presented in Fig. 3, and we can observe two phenomena. First, since the segmentation is basically correct, the DNN is capable of understanding the environment through the range-Doppler matrix. Second, the segmentation results vary significantly among different classes. For *free space* and *nonfree space*, the segmentation is fairly precise, while it is relatively worse for *vehicle*. The reason for this discrepancy is apparent that even for a HD radar, its angular FoV and resolution are not on par with a common camera system, and the performance gap is tremendous, especially in the elevation direction. For example, in the elevation direction, this HD radar has 10° FoV with 1° resolution, while the camera covers 90° FoV with approximately 0.17° resolution. Therefore, a model trained with radar data cannot precisely segment small objects in the camera image.

When we plan to extend the radar-based environment understanding from semantic level to instance or even to panoptic level, the naive method that directly segments pixel by pixel in the camera image is theoretically feasible. However, based on the analysis of Fig. 3, the naive method is not optimal since radar cannot accurately perceive the geometry of small objects like the camera. Thus, pixel-by-pixel segmentation in the camera image achieves extremely frustrating results. At the same time, it is not a fair and rational evaluation for a model trained with radar data, considering the perceptual discrepancy between these two modalities.

Therefore, we propose a novel adaption applied to road users to achieve radar-tailored panoptic segmentation. Our inspiration comes from the core idea in [15] that an object is modeled as a keypoint (its center point of the bounding box) during object detection. Not like other detectors such as Faster RCNN [26], YOLOv3 [16] that represent the object through an axis-aligned, tightly encompassing bounding box, [15] represents objects by a single point at their bounding box center. Object detection is divided into the estimation of the keypoint and regression of other properties, such as size, orientation, 3D location, and even pose, from the features at the keypoint.

In our work, a DNN is trained to predict the object center position and depth in the camera image plane. Since a moving object is denoted by a single point, we circumvent the estimation of the object's precise shape and size, which are beyond the radar's perceptual capability. Compared with the camera, the radar has an essential advantage: it can accurately detect the distance to objects. Accurate distance information of objects is critical when generating the 3D surrounding map of the ego vehicle, which can be applied to various crucial tasks for AD, such as collision avoidance, trajectory planning. With this in mind, the distance of moving objects (namely, the depth for the image plane) as an extra feature is also regressed from this DNN. In addition to detecting moving objects, the DNN simultaneously classifies and segments the surroundings into two classes in the image plane, namely *free space* and *nonfree space*. Through combing the semantic level information from surroundings and instance level information from objects, a radar-tailored panoptic segmentation is achieved by the DNN, which is elaborated in Section IV-D.

D. DNN Structure

The structure of the DNN is depicted in Fig. 4. Basically, the U-Net structure is adopted and can be divided into three parts: an encoder and detection and segmentation branches in the decoder.

The range-Doppler matrix comprises originally complex numbers with 16 channels owing to 16 receivers, and we split the complex number into real and imaginary parts to form the input size $256 \times 512 \times 32$. In the encoder part, we follow [5] to use a dilated convolution at the beginning to encode the input with a larger reception field because Doppler-division multiplexing is utilized, and the Doppler signal of one object appears 12 times along the Doppler axis, which can be clearly observed in Fig. 4. Concretely, the dilated convolution has a 12×1 kernel size, with dilated rate 16×1. This specification is determined by Doppler-division multiplexing, which is elaborated in Section II. The two convolutional layers in the Conv. Block have 3×3 kernel size. For the input range-Doppler matrix, the angular information is encoded in the channel dimension. Thus, a self-attention layer along the channel dimension [27] is included in the Conv. Block to enhance the cross-channel correlation. The output from the encoder comprises the extracted features from the range-Doppler matrix, and it is delivered to the detection and segmentation branches in the decoder. In the detection branch, the outputs are the keypoint prediction, offset, and depth heatmaps shown in Fig. 4. Since the keypoint heatmap is four times smaller than the original input size, an offset heatmap is needed to estimate the position offsets of a keypoint between the keypoint heatmap and the original size image in the row and column directions. In addition, the depth heatmap is responsible for estimating keypoint depth to the image plane. Besides the detection branch, the decoder has a segmentation branch that predicts the *free space* in the image plane.

V. Training Details

The network was trained on a workstation equipped with a Xeon W-1390P CPU, a Nvidia RTX A5000 GPU and 64 GB RAM. We set 4 as the batch size, 1.5×10^{-4} as the initial training rate and select the Adam optimizer with a step decaying learning rate schedule.

Since it is multitask learning, the total loss function can generally be divided into segmentation and detection parts.

Segmentation Task: The *free-space segmentation* task is viewed as a common pixel-level binary classification. A 1×1 2D convolutional layer as the last layer in the segmentation branch outputs a 2D feature map. After a sigmoid activation function, the predicted *free space* probability of each pixel in the camera image plane is available. Assuming that $\boldsymbol{E}$ is a training example, $y \in \{0,1\}^{256\times 512}$ is the ground truth for each pixel, and $\widehat{y} \in [0,1]^{256\times 512}$ is the prediction outcome. To learn this task, the *free-space segmentation* loss L_{seg} is formulated:

$$L_{\text{seg}} = \alpha_{\text{bce}} \cdot L_{\text{bce}} + \alpha_{\text{dice}} \cdot L_{\text{dice}}, \tag{4}$$

where L_{bce} and L_{dice} are the binary cross entropy and dice loss, α_{bce} and α_{dice} are their weights, respectively. We set α_{bce} and α_{dice} as 0.5. L_{bce} and L_{dice} can be expressed in (5) and (6):

$$L_{\text{bce}} = -\sum_{\boldsymbol{E}} y \cdot log(\widehat{y}) \tag{5}$$

$$L_{\text{dice}} = 1 - \frac{2 \cdot \sum_{\boldsymbol{E}} y \cdot \widehat{y}}{\sum_{\boldsymbol{E}} y + \sum_{\boldsymbol{E}} \widehat{y}} \tag{6}$$

Detection Task: In Fig. 4, the detection branch receives the encoded features and generates the keypoint prediction, offset, and depth heatmaps. The detection loss L_{det} is formulated in (7), where α_{hm}, α_{off}, and α_{dep} are their heatmap weights, respectively. We define $\alpha_{\text{hm}} = 0.4$, $\alpha_{\text{off}} = 0.4$, and $\alpha_{\text{dep}} = 0.2$. The loss for the keypoint prediction heatmap in (8) is a logistic regression using focal loss [28] with $\alpha = 2$ and $\beta = 4$. The penalty is alleviated at the pixels, in which the keypoints locations $\widetilde{\boldsymbol{p}}$ are

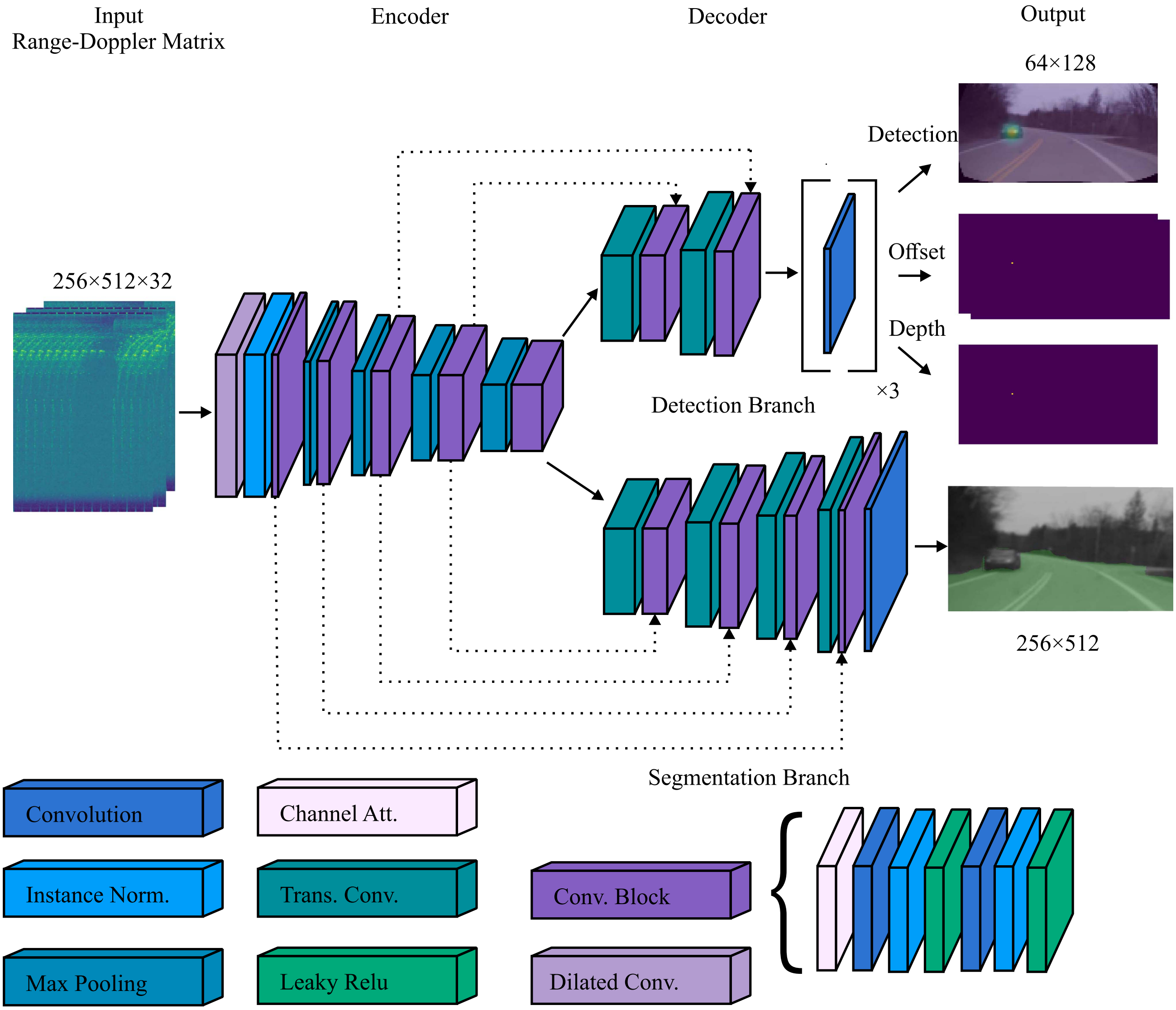


Fig. 4. The DNN structure consists of an encoder and detection and segmentation branches. ×3 in the detection branch means three different convolutional layers generate the keypoint prediction, offset, and depth heatmaps, respectively. Since there are offsets in the row and column directions, offset heatmap has two channels. The keypoint prediction and depth heatmaps have only one channel.

not located. Symbol $\widetilde{\boldsymbol{p}}$ as labeled keypoint position is defined in (1). N means the number of keypoints (objects) in example $\boldsymbol{E}$. Since the heatmap size is four times smaller than the original image size, we use the offset heatmap to regress the row and column offset $\widehat{\boldsymbol{\Delta}} \in \mathcal{R}^{64\times 128\times 2}$. Formula (9) presents the offset loss L_{off}, and L_{dep} is a L1 loss. In the offset and depth heatmaps, only the data at keypoint location $\widehat{\boldsymbol{p}}$ is utilized, and other data are neglected.

$$L_{\text{det}} = \alpha_{\text{hm}} \cdot L_{\text{hm}} + \alpha_{\text{off}} \cdot L_{\text{off}} + \alpha_{\text{dep}} \cdot L_{\text{dep}} \tag{7}$$

$$L_{\text{hm}} = -\frac{1}{N}\sum_{rc}\begin{cases}(1-\widehat{v}_{rc})^{\alpha}\cdot\log(\widehat{v}_{rc}) \text{ if } v=1\\(1-v_{rc})^{\beta}\cdot(\widehat{v}_{rc})^{\alpha}\cdot\log(1-\widehat{v}_{rc}) & \text{otherwise}\end{cases} \tag{8}$$

$$L_{\text{off}} = \frac{1}{N}\sum_{\boldsymbol{p}}\left|\widehat{\boldsymbol{\Delta}}_{\widetilde{\boldsymbol{p}}} - \left(\frac{\boldsymbol{p}}{4} - \widetilde{\boldsymbol{p}}\right)\right| \tag{9}$$

Multitask: Combining the previous two tasks, we obtain the final end-to-end L_{total}:

$$L_{\text{total}} = \lambda_{\text{seg}} \cdot L_{\text{seg}} + \lambda_{\text{det}} \cdot L_{\text{det}}. \tag{10}$$

The entire network, including an encoder and segmentation and detection branches, is trained through the minimization of L_{total}. We set $\lambda_{\text{seg}} = 0.7$, and $\lambda_{\text{det}} = 0.3$.

VI. Experimental Results

The results from the proposed network are compared to other contributions in the radar community. Considering our training tasks, the comparison consists of two aspects, namely *free-space segmentation* and *object detection*.

TABLE I
FREE-SPACE SEGMENTATION COMPARISONS

Test index	Method	Input	No. frame(s)	Evaluation plane	mIoU(%) ↓		
					Overall	Easy	Hard
I	Ours	Range-Doppler matrix	Single	Image	**93.2**	**93.8**	**91.7**
II	Ours w/o offset heatmap	Range-Doppler matrix	Single	Image	92.1	92.7	90.2
III	Ours w/o depth heatmap	Range-Doppler matrix	Single	Image	90.3	90.9	88.3
IV	Ours w/o det. branch	Range-Doppler matrix	Single	Image	89.4	90.0	87.5
V		Seg. img.		Polar BEV	85.5	86.1	83.4
VI	Ours	Range-Doppler matrix	Single	Polar BEV	80.4	81.6	76.7
VII	FFT-RadNet [5]	Range-Doppler matrix	Single	Polar BEV (seg. img.)	75.0	76.1	73.0
VIII	FFT-RadNet [5]	Range-Doppler matrix	Single	Polar BEV	74.0	74.6	72.3
IX	Grid map [10]	Point cloud	Multiple	Polar BEV	17.6	17.7	16.6

[1] We keep the original dataset division for easy and hard cases from the RADIal. In hard cases, the radar is perturbed by factors such as interference from other radars, serious side-lobes. This two cases division can also be found in the comparison of object detection and depth estimation. [2] In the third column, "Seg. img." means the segmentation results from panoptic FPN. [3] In the fourth column, the number of frames indicates whether the method utilizes single or multiple accumulated measurements to predict. [4] In the fifth column, "Polar BEV" represents that the evaluation happens in the polar bird's eye view plane, and these labels are provided by RADIal. In contrast, "Image" means that the prediction is evaluated in the image plane with the ground truth results from panoptic FPN (namely "Seg. img."). Furthermore, "Polar BEV (seg. img)" denotes that the ground truth is obtained through the conversion of segmented images from panoptic FPN to the same label format as RADIal, and the evaluation is also in the polar bird's eye view plane.

A. *Free-Space Segmentation*

To have a fair comparison among different methods, we follow [5] to restrict the area for mean Intersection-over-Union (IoU) within $[0\,m, 50\,m]$ range since over $50\,\mathrm{m}$ the lidar signal from the road is too weak to be recognized exactly for annotation.

The results are presented in Table I. The test I shows the segmentation result from radar evaluated with the panoptic FPN generated ground truth. Note that the segmentation and evaluation occur in the image plane. The fact that mean IoU in all three categories are over 90% indicates that our DNN, with the help of image-based training methodology, can segment the free space precisely. Furthermore, we conduct an ablation analysis (tests II, III and IV) to investigate the network. Obviously, the offset and depth affect the segmentation performance, and the depth heatmap is more significant. In the test IV, the metrics decrease distinctly for the structure without detection branch. Based on the ablation analysis, we can find that the detection branch facilitates the performance of the segmentation branch since it provides prior information regarding the detected vehicles.

As the evaluation from [5] was based on BEV in the polar coordinate system (which we call polar BEV for short), which is different from the image plane, relying on the camera's intrinsic and extrinsic parameters, we convert segmented images from the image plane to the polar BEV. Assuming that $\boldsymbol{M}_{\text{Img}}$ is the *free-space segmentation* in the image plane and $\boldsymbol{M}_{\text{Polar}}$ is the counterpart in polar BEV, the process of converting one location (ρ_i, θ_i) in the polar coordinate system to an image pixel (m_i, n_i) can be divided into two steps in (11). First, $f_{\text{Polar}}^{\text{Cartesian}}$ fulfills the conversion from polar to Cartesian, and then the Cartesian location is transformed to the pixel position by $f_{\text{Cartesian}}^{\text{Img}}$, which can be determined with camera calibration. Furthermore, based on the derived relationship in (11), $\boldsymbol{M}_{\text{Polar}}(\rho_i, \theta_i)$ should have the same value as $\boldsymbol{M}_{\text{Img}}(m_i, n_i)$, which we denote in (12).

$$(m_i, n_i) = f_{\text{Cartesian}}^{\text{Img}}(f_{\text{Polar}}^{\text{Cartesian}}(\rho_i, \theta_i)) \tag{11}$$

$$\boldsymbol{M}_{\text{Polar}}(\rho_i, \theta_i) = \boldsymbol{M}_{\text{Img}}(m_i, n_i) \tag{12}$$

In the test V, we find that although our label is generally good, it still has some discrepancies compared with the annotation provided by the RADIal. Note that the segmented images were generated from panoptic FPN after fine-tuning with Cityscapes, and we did not label any image from RADIal. For instance, referring to Fig. 3, except for the guard rail at the right bottom of the image, the other area is correctly segmented. The imperfection of the label largely comes from the domain shift between the Cityscapes (source domain) and RADIal dataset (target domain). One solution to this issue can be manual labeling some portion of the data and then further training, which could be considered as future work.

For the test VI, we convert the result from the DNN from the image plane to polar BEV and evaluated it with annotations from RADIal. Despite the flaw in the ground truth, the mean IoU significantly surpasses FFT-RadNet [5] in all three categories. Furthermore, we can infer that if we could improve our labeling quality with manual labeling, the performance of the DNN would be further boosted.

In addition, we train and evaluate FFT-RadNet with different ground truths for *free-space segmentation* task. The process of making labels in the polar coordinate system from our segmented images is the same as in (11) and (12). Analyzing the tests VII and VIII, FFT-RadNet trained with the ground truths that are converted from our segmented images has a little improvement over that trained with RADIal provided. When generating the free space mask in polar coordinate system, authors from [5] first adopted the same way that utilized segmented images. However, they obtained the bounding box information of the vehicle from lidar and then subtracted the corresponding area in the free space mask. Evidently, factors such as the synchronization between lidar and camera, can affect the labeling quality. In light of the slight improvement in the test VII, their labeling quality is generally acceptable within the $50\,\mathrm{m}$ range. Furthermore, comparing

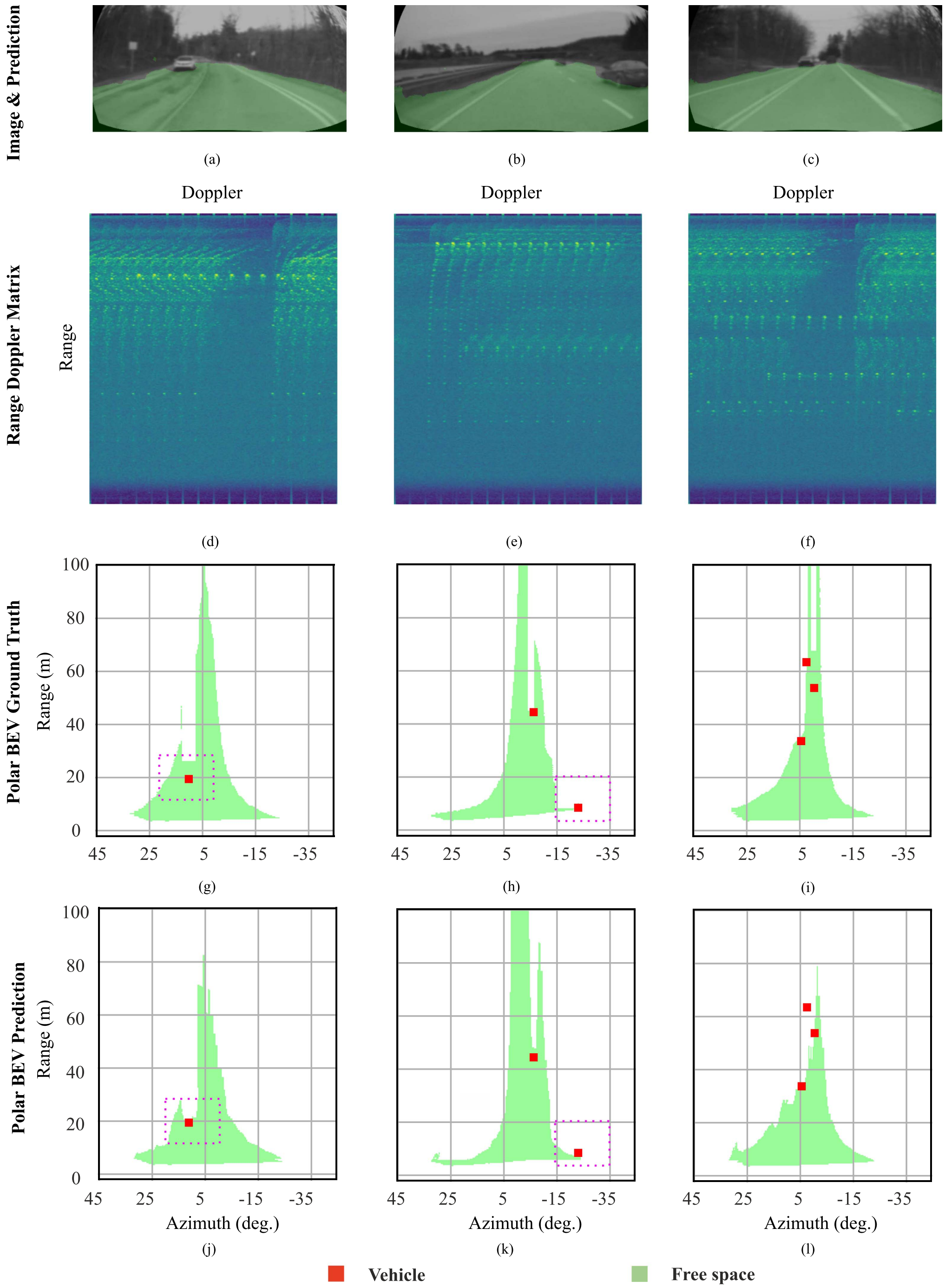
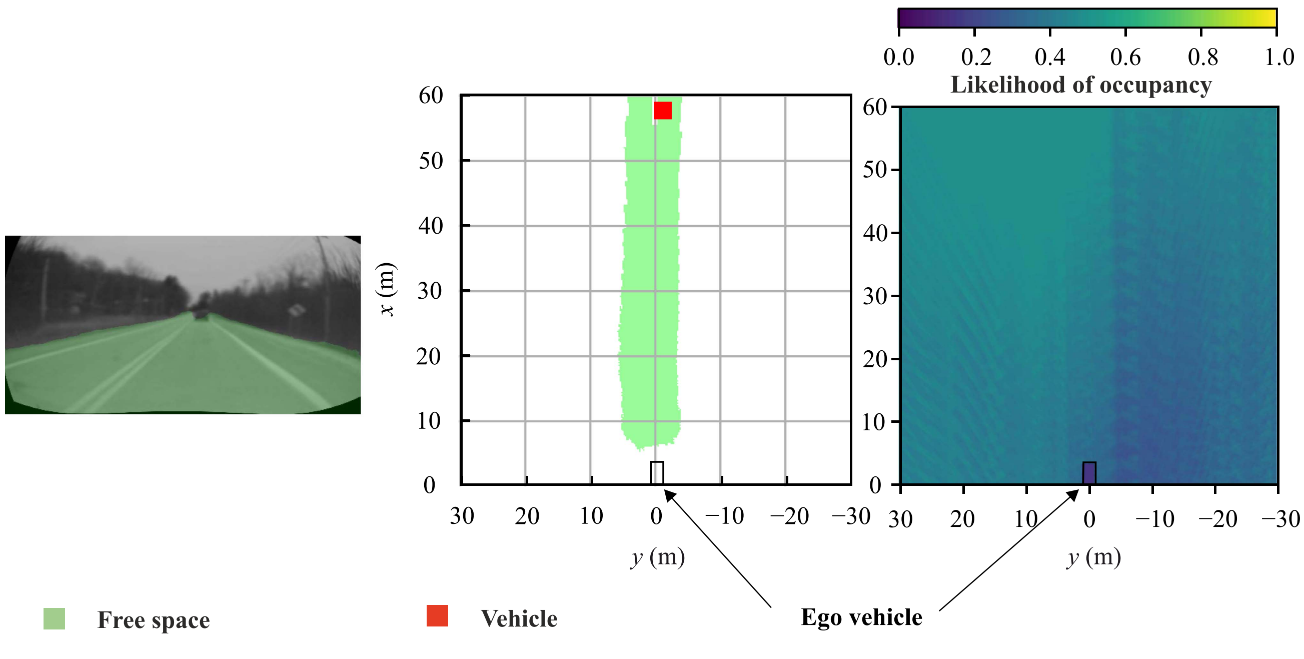


Fig. 5. Three qualitative results for segmentation, which have one, two, and three vehicles, respectively. (a)–(c) show undistorted images with overlapped prediction from the model. The corresponding range-Doppler matrices are in (d)–(f). (g)–(i) are ground truth segmentation in the polar coordinate system from RADIal, and (j)–(l) show the model predicted counterparts. In the magenta dashed line box, the prediction is even more precise than the ground truth.

the results in the tests VI and VII, although complete removal of the influence from different networks is impossible, we find that the training with the radar data in the camera perspective for *free-space segmentation* does not have a significantly detrimental effect. The reason is that in a short range such as within $50\ \mathrm{m}$ in this work, the projection of a precisely segmented image to the BEV is rather precise considering that the camera's perspective favors a better resolution of neighboring objects, and the projection error is relatively minor in the vicinity of the camera. This point can be supported by the magenta dashed line boxes in Fig. 5(j)–(k). Apart from these, the availability of precisely segmented ground truths is more convenient in the image plane than the BEV. Thus, training the DNN in the image plane and then projecting to the BEV is more advantageous.

For the test IX, first, we utilize the method in [10] to generate an occupancy grid map (OGM) from the radar detection. For the



(a) (b) (c)

Fig. 6. Comparison with the grid map method. (a) and (b) show the prediction result of our model in the image plane and BEV respectively. (c) is the corresponding occupancy grid map of this measurement. Due to the limited Doppler unambiguous range, static detection is falsely classified as dynamic. Consequently, most area in the occupancy grid map is unknown or free status (less than 0.5 likelihood), and its free-space segmentation result is fairly worse.

OGM, ego-motion information is required; otherwise, neither the differentiation between moving and static detection nor the accumulation of static detection from different moments will succeed. However, ego-motion is not available for all validation measurements; thus, we obtain the OGM from only two traces in the validation dataset. Despite this, using partial measurement does not collide with the conclusion. Once the OGM is obtained, we select the likelihood p_{ogm}. If the likelihood for one cell in OGM is smaller than p_{ogm}, this cell is considered as *free space*, otherwise as *nonfree space*. Although using a segmentation network like [10] should be a better solution than setting a threshold, and the segmentation performance can have some extent improvement, the result is still far behind the aforementioned methods, which utilize a single range-Doppler matrix as input.

Fig. 5 presents three examples of *free-space segmentation*. In the first row, the predicted semantic masks from our DNN are overlaid on the undistorted images. The segmentation performance is rather satisfactory since the majority of free space area, including details like the challenging road boundary, is precisely segmented. Furthermore, to compare with the ground truth provided by RADIal in the third row, the predicted masks are converted to the BEV in the polar coordinate system in the last row. One intriguing point is that our prediction in the dashed line box is even more accurate than its counterpart in the ground truth. At the same time, it proves that the projection of an exactly segmented image in the BEV within $50\,\mathrm{m}$ range is rather precise.

Compared with the grid map-based method [10], our approach has four advantages. First, our method is capable of correctly segmenting the area occupied by the moving road user like (j)–(l) in Fig. 5 and in Fig. 6(a), while the grid map-based counterpart has to abandon the moving detection because it causes discrepancy during the accumulation. Furthermore, it is more robust in facing situations like clutter or Doppler ambiguity. For instance, in Fig. 6(c), when the vehicle drives faster than its radar's Doppler unambiguity velocity, plenty of detection from static objects is falsely classified as dynamic. It causes the serious consequence that the OGM can barely offer useful information as the likelihood of occupancy in the most area is less than 0.5, i.e., at an unknown or free status, which contradicts the fact. Third, getting rid of accumulation processing, our method's prediction is solely based on a single radar measurement. It means the prediction can respond faster to changing surroundings, which is crucial for various AD scenarios. Furthermore, our single-measurement method does not degrade the segmentation for details like the road boundary, which are usually challenging for grid map-based counterpart. Last but not least, the intensive computation for angle estimation and the CFAR algorithm is skipped in our approach; thus, it is more efficient than the grid map-based approach.

B. Object Detection

To evaluate the object detection, we adopt the average precision (AP) and average recall (AR) as the evaluation metrics. FFT-RadNet [5] is selected as the baseline since it is able to detect object and utilizes a range-Doppler matrix as input. As our detection output for an object was solely a point instead of a bounding box, the criterion for true positive (TP), false positive (FP), and false negative (FN), which is based on the IoU metric, has to be modified.

Basically, we define a predicted detection $\boldsymbol{d}_i(r_i, c_i)$ as TP, if $\boldsymbol{d}_i \in \boldsymbol{S}_{o_j}$ where $\boldsymbol{S}_{o_j}$ is the pixel area for object o_j. $\boldsymbol{S}_{o_j}$ is

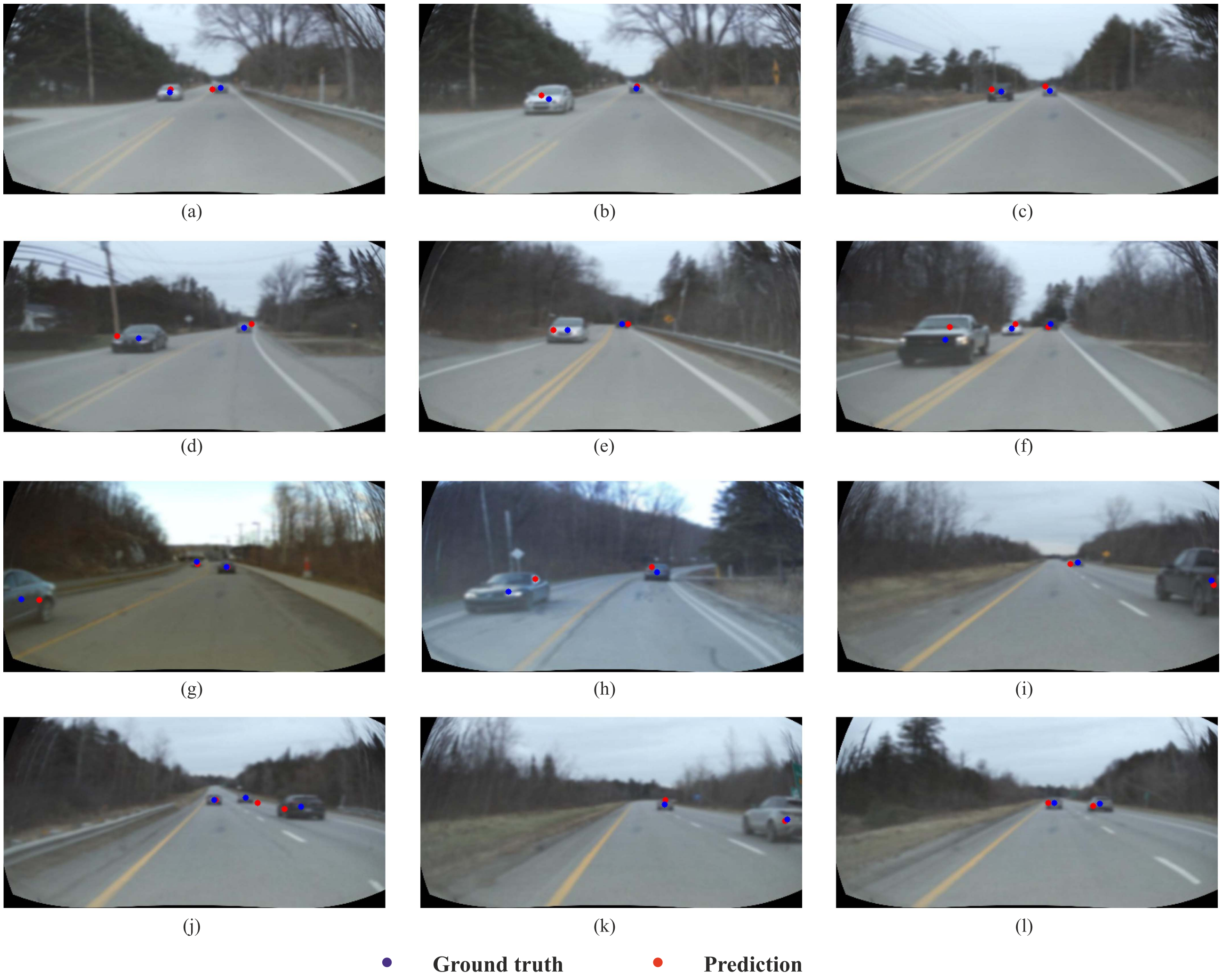


Fig. 7. Vehicle detection examples. The blue point denotes the ground truth for the vehicle. The red point is the predicted vehicle position. In these examples, the predicted vehicle positions are rather precise since they are close to or even overlapped with the ground truth positions.

TABLE II
OBJECT DETECTION COMPARISON

Model	Overall		Easy		Hard	
	AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)
Ours	**96.90**	**83.49**	**98.61**	**91.98**	92.79	**66.41**
FFT-RadNet	96.84	82.18	98.49	91.69	**92.93**	64.82

described as

$$\boldsymbol{S}_{o_j} = \{(r,c)|p(r,c) > p_{\text{th}}, 0 < r <= R \text{ and } 0 < c <= C\}, \tag{13}$$

where R and C are the number of rows and columns of the image, and p_{th} is the threshold likelihood for a pixel classified as part of the object. Similar to the criteria with the bounding box, if $\boldsymbol{d}_i \notin \boldsymbol{S}_{o_j}$, then $\boldsymbol{d}_i$ is FP and for object o_j is FN.

The comparison is listed in Table II. The results in the table show that, in general, our method surpasses its counterpart. Considering AR, the improvement is fairly notable, especially for hard measurements. In contrast, both networks have almost the same performance at AP. Since our output is directly in the image plane, fusion between the radar and the camera can be easily achieved. In this case, a higher AR should be more important for the safety of AD because a road user is more likely to be detected. The qualitative evaluation of vehicle detection is showed in Fig.7. The ground truth and predicted detection are overlapped on the image to give readers an intuitive understanding. Due to the perspective and objects' limited physical size, they occupy fewer pixels than the free space. Furthermore, they usually have a faster and more complicated relative motion to the radar than the static free space. All these make *object detection* a harder task than *free-space segmentation*. Despite these challenges, our method can still precisely detect vehicles from a great distance.

However, while analyzing the *object detection* results, we find that several factors, such as the quality of the camera calibration and the synchronization between two sensors, can impact the detection performance. The details are elaborated in Chapter VIII.

In addition, our network achieves a range estimation for every predicted keypoint. The evaluation results are listed in Table III.

TABLE III
MEAN RANGE ESTIMATION RESULTS

Model	Mean Range Estimation Error (m)		
	Overall	Easy	Hard
Ours	0.45	0.40	0.48

Considering that the following factors can influence the accuracy of the range estimation, the results are rather good. First, it is strongly determined by the radar's range resolution. In this work, the radar has a range resolution of $0.2\,\mathrm{m}$. Furthermore, a vehicle has a certain size and thus occupies a cluster of pixels in the image. However, owing to the adaption to the radar, a vehicle is simplified as a point at the center of its bounding box. The ground truth range for the vehicle is provided from the dataset and is yielded from the lidar coordinate system, while our estimation is in the radar coordinate system. These discrepancies deteriorate the estimation result. In addition, issues such as imprecise synchronization between the radar and camera, imprecise camera calibration, or transient changing of the relative orientation between the two sensors due to the road or vehicle's motion affect the estimation.

VII. Complexity Analysis

We analyze the complexity of the proposed model and its temporal performance. The proposed model has approximately $7.7\,\mathrm{M}$ parameters and requires 358 GFLOPS for one inference. Temporally, it needs $68\,\mathrm{ms}$ to process one sample. Since the working cycle of a current automotive radar system is around $60\,\mathrm{ms}$ and no common optimization techniques are utilized here, we expect that it should be easy to put our model in a real-time application.

VIII. Discussion

To sum up, our cross-modal supervised-based training methodology and model provide promising results. The improvement in *free-space segmentation* is notable. For *object detection*, compared with previous works, the performance is improved, albeit slightly for precision. We further analyze the failure case in *object detection* and find some possible reasons, which are given in Section VIII-A. Then, we talked bout the possible feature work based on our method in Section VIII-B.

A. Problem Analysis

In Fig. 8, we reproject the vehicle position from 3D to the image plane using intrinsic and extrinsic parameters of the provided camera, and the reprojected point is denoted in magenta. The 3D position of the vehicle is provided by RADIal through labeling in the lidar data. Theoretically, the reprojected point should be overlapped with the blue ground truth point. Nonetheless, an apparent discrepancy can be observed. This discrepancy could be caused by one or multiple reasons, such as worse synchronization between the radar and the camera and imprecise intrinsic or extrinsic camera calibration. The prediction result will probably be inaccurate when training the networks with ground truth like this. Based on our observation, this case could sometimes happen when an abrupt road change exists.

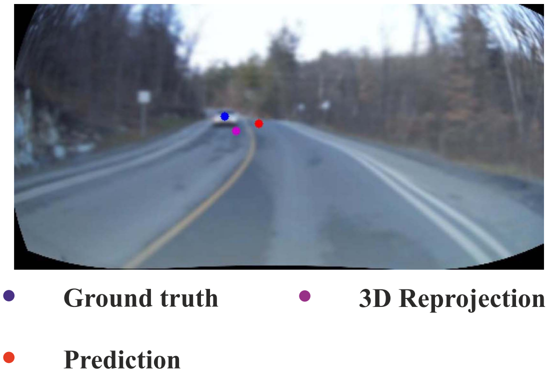


Fig. 8. An example of the object detection failure. We use intrinsic and extrinsic parameters of the provided camera to reproject the 3D location of the vehicle to the image and obtain this 3D reprojection point. There is an apparent discrepancy between the ground truth and the 3D reprojection.

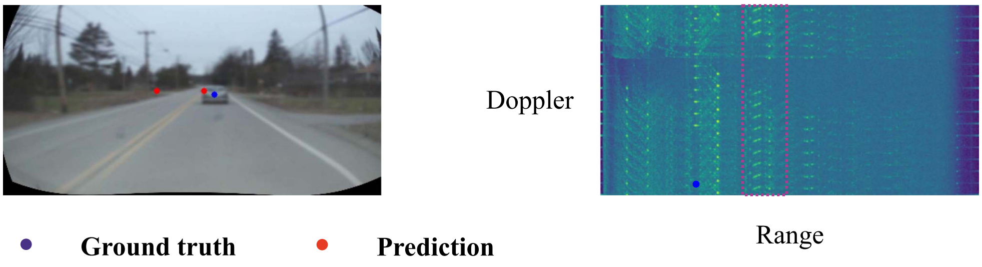
B. Future Work

As future work, in addition to the designed camera-radar system concept in Section IV-B, the new fusion paradigm that fuses the radar and camera in the image plane can be regarded as a promising direction for future work, since this work proves that a DNN can understand the surroundings based on the radar data and transform the result to the camera perspective.

This novel fusion paradigm can enhance the perception capability of radar systems and thus has the potential to solve the radar ghost target issue. For instance, in Fig. 9, apart from a correctly predicted vehicle, we can observe a ghost target in the vicinity of a pole. In the range-Doppler matrix (b), the blue point signifies the vehicle's unambiguous Doppler signal, and the signal within the dashed line box generates the ghost target. The ghost target can be easily recognized and eliminated if the prediction is fused with the camera image.

IX. Conclusion

In this work, we propose a novel network to achieve radar-tailored panoptic segmentation (i.e., *free-space segmentation* and *object detection*) with only range-Doppler matrix data in urban, rural, and highway scenarios. Our network outperforms previous work, and the improvement in *free-space segmentation* is especially significant. To get rid of labeling challenges for radar data, we utilize a cross-modal supervision method. Specifically, segmented camera images are adopted and adjusted as the ground truth for the training, which significantly alleviate the labeling expense. Relying on this cross-modal supervision approach, we design a camera-radar system concept that realizes automatic training of a DNN with the radar data. Based on this work, the novel fusion paradigm that fuses radar with a camera in the image plane, would be a good direction for future work.



(a) (b)

Fig. 9. Ghost target example. The blue point denotes the ground truth position of a vehicle in (a) and its unambiguous Doppler velocity in the range-Doppler matrix (b). The red point is the predicted vehicle position. The left red point is a ghost detection, which is probably caused by the pole, and its signal is marked in the magenta dashed line box in (b). The signal in (b) is fairly noisy, because some noise signals are at the left side of the blue point and the right side of the magenta dashed line box.

References

[1] L. M. Clements and K. M. Kockelman, "Economic effects of automated vehicles," *Transp. Res. Rec.*, vol. 2606, no. 1, pp. 106–114, 2017.

[2] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, 2021, Art. no. 2140. [Online]. Available: https://www.mdpi.com/1424-8220/21/6/2140

[3] N. Scheiner, O. Schumann, F. Kraus, N. Appenrodt, J. Dickmann, and B. Sick, "Off-the-shelf sensor vs. experimental radar-How much resolution is necessary in automotive radar classification?," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion*, 2020, pp. 1–8.

[4] Y. Jin, R. Prophet, A. Deligiannis, I. Weber, J.-C. Fuentes-Michel, and M. Vossiek, "Comparison of different approaches for identification of radar ghost detections in automotive scenarios," in *Proc. IEEE Radar Conf.*, 2021, pp. 1–6.

[5] J. Rebut, A. Ouaknine, W. Malik, and P. Pérez, "Raw high-definition radar for multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17021–17030.

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

[7] I. Orr, M. Cohen, and Z. Zalevsky, "High-resolution radar road segmentation using weakly supervised learning," *Nature Mach. Intell.*, vol. 3, no. 3, pp. 239–246, 2021.

[8] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6399–6408.

[9] M. A. Richards, "Detection fundamentals," in *Fundamentals of Radar Signal Processing*. New York, NY, USA: McGraw-Hill Education, 2014, pp. 295–345.

[10] R. Prophet, G. Li, C. Sturm, and M. Vossiek, "Semantic segmentation on automotive radar maps," in *Proc. IEEE Intelli. Veh. Symp. (IV)*, 2019, pp. 756–763.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[14] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[15] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019. [Online]. Available: https://arxiv.org/abs/1904.07850

[16] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018. [Online]. Available: https://arxiv.org/abs/1804.02767

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

[18] S. Chadwick and P. Newman, "Radar as a teacher: Weakly supervised vehicle detection using radar labels," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 222–228.

[19] M. Dimitrievski, I. Shopovska, D. V. Hamme, P. Veelaert, and W. Philips, "Weakly supervised deep learning method for vulnerable road user detection in FMCW radar," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst.*, 2020, pp. 1–8.

[20] C. Grimm, T. Fei, E. Warsitz, R. Farhoud, T. Breddermann, and R. Haeb-Umbach, "Warping of radar data into camera image for cross-modal supervision in automotive applications," *IEEE Trans. Veh. Technol.*, vol. 71, no. 9, pp. 9435–9449, Sep. 2022.

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

[22] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019. [Online]. Available: https://github.com/facebookresearch/detectron2

[23] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

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

[25] R. Mohan and A. Valada, "EfficientPS: Efficient panoptic segmentation," *Int. J. Comput. Vis.*, vol. 129, pp. 1551–1579, 2020.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[27] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018. [Online]. Available: https://arxiv.org/abs/1804.03999

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

Yi Jin (Graduate Student Member, IEEE) was born in Kunshan, China, in 1992. He received the M.Sc. degree in electromobility from the University of Stuttgart, Stuttgart, Germany, in 2019. In 2019, he joined the Institute of Microwaves and Photonics (LHFT), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, where he is currently pursuing a Ph.D. degree. His research focuses on (self)-supervised learning with automotive radar and sensor fusion.

Anastasios Deligiannis (Member, IEEE) received the Diploma (bachelor's and master's degrees equivalent) from the School of Electrical and Computer Engineering, University of Patras, Patras, Greece, in 2012, and the Ph.D. degree in radar signal processing from Loughborough University, Loughborough, U.K., in 2016. He was a Postdoctoral Research Associate in signal processing with Loughborough University till 2018. From 2018 to 2019, he was with Volvo Cars, as a Radar Performance Engineer for autonomous driving projects. Since July 2019, he has been with BMW Group, as a Radar Expert with autonomous driving and advanced driver assistance projects. His research interests include signal processing algorithms, automotive radar performance, sparse array design, convex optimization, beamformer development and game theoretic methods, within the radar network framework, and wireless communications.

Juan-Carlos Fuentes-Michel received the Ph.D. degree from the Clausthal University of Technology, Clausthal-Zellerfeld, Germany, in 2008. He is currently a Radar Technology Expert for driver assistance systems and autonomous driving with the BMW Group and the Author of several publications on wireless positioning and radar signal processing. During his career, he has contributed to the industrialization of different radar components with the automotive industry, from the initial concept phase until mass implementation.

Martin Vossiek (Fellow, IEEE) received the Ph.D. degree from Ruhr-Universität Bochum, Bochum, Germany, in 1996. In 1996, he joined Siemens Corporate Technology, Munich, Germany, where he was the Head of the Microwave Systems Group, from 2000 to 2003. Since 2003, he has been a Full Professor with Clausthal University, Clausthal-Zellerfeld, Germany. Since 2011, he has been the Chair of the Institute of Microwaves and Photonics (LHFT), Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. He has authored or coauthored more than 350 articles. His research has led to more than 100 granted patents. His research interests include radar, microwave systems, wave-based imaging, transponder, RF identification, communication, and wireless locating systems. He is a Member of the German National Academy of Science and Engineering (acatech) and German Research Foundation (DFG) Review Board. He is a Member of the IEEE Microwave Theory and Technology (MTT) Technical Committees MTT-24 Microwave/mm-wave Radar, Sensing, and Array Systems, Founding Chair of MTT-27 Connected and Autonomous Systems and MTT-29 Microwave Aerospace Systems. He is also with the Advisory Board of the IEEE CRFID Technical Committee on Motion Capture and Localization. He was the recipient of the numerous best paper prices and other awards and Microwave Application Award from the IEEE MTT Society (MTT-S) for Pioneering Research in Wireless Local Positioning Systems in 2019. Dr. Vossiek is a Member of organizing committees and technical program committees for many international conferences and he has served on the Review Boards for numerous technical journals. From 2013 to 2019, he was an Associate Editor for the IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES. Since October 2022, he has been an Associate Editor-in-Chief for IEEE TRANSACTIONS ON RADAR SYSTEMS.