RODNet: A Real-Time Radar Object Detection Network Cross-Supervised by Camera-Radar Fused Object 3D Localization

Yizhou Wang, *Student Member, IEEE*, Zhongyu Jiang, *Student Member, IEEE*, Yudong Li, Jenq-Neng Hwang, *Fellow, IEEE*, Guanbin Xing, and Hui Liu, *Fellow, IEEE*

***Abstract*—Various autonomous or assisted driving strategies have been facilitated through the accurate and reliable perception of the environment around a vehicle. Among the commonly used sensors, radar has usually been considered as a robust and cost-effective solution even in adverse driving scenarios, e.g., weak/strong lighting or bad weather. Instead of considering fusing the unreliable information from all available sensors, perception from pure radar data becomes a valuable alternative that is worth exploring. In this paper, we propose a deep radar object detection network, named RODNet, which is cross-supervised by a camera-radar fused algorithm without laborious annotation efforts, to effectively detect objects from the radio frequency (RF) images in real-time. First, the raw signals captured by millimeter-wave radars are transformed to RF images in range-azimuth coordinates. Second, our proposed RODNet takes a snippet of RF images as the input to predict the likelihood of objects in the radar field of view (FoV). Two customized modules are also added to handle multi-chirp information and object relative motion. The proposed RODNet is cross-supervised by a novel 3D localization of detected objects using a camera-radar fusion (CRF) strategy in the training stage. Due to no existing public dataset available for our task, we create a new dataset, named CRUW,[1] which contains synchronized RGB and RF image sequences in various driving scenarios. With intensive experiments, our proposed cross-supervised RODNet achieves 86% average precision and 88% average recall of object detection performance, which shows the robustness in various driving conditions.**

***Index Terms*—Radar object detection, deep CNN, autonomous driving, advanced driver assistance system, cross-modal supervision, M-Net, temporal deformable convolution, temporal inception CNN, radar object annotation.**

Manuscript received September 1, 2020; revised December 14, 2020; accepted February 6, 2021. Date of publication February 11, 2021; date of current version June 3, 2021. This work was supported by CMMB Vision – UWECE Center on Satellite Multimedia and Connected Vehicles. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Philipp Heidenreich. *(Corresponding author: Yizhou Wang.)*

Yizhou Wang, Zhongyu Jiang, Yudong Li, Jenq-Neng Hwang, and Guanbin Xing are with the Department of Electrical, and Computer Engineering, University of Washington, Seattle 98105, WA USA (e-mail: ywang26@uw.edu; zyjiang@uw.edu; yudonl@uw.edu; hwang@uw.edu; gxing@uw.edu).

Hui Liu is with the Department of Electrical, and Computer Engineering, University of Washington, Seattle 98105, WA USA, and also with the Silkwave Holdings Limited, Hong Kong (e-mail: huiliu@uw.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSTSP.2021.3058895, provided by the authors.

Digital Object Identifier 10.1109/JSTSP.2021.3058895

[1]The dataset and code are available at https://www.cruwdataset.org/

I. Introduction

RADAR is a common sensor for driving assistance these days, since it is effective in most driving scenarios, including different weather and lighting conditions, resulting in its robustness compared with other sensors, e.g., camera and LiDAR. Many autonomous or assisted driving solutions focus on sensor fusion to improve the accuracy and reliability of the perception results, where radar is mostly used as a complement for cameras or LiDARs. It is mainly because most fusion approaches only take advantage of the more robust localization information in the radar signals, while the rich semantic information hasn't been well exploited. Thus, in this paper, we manage to extract the semantic features of the radio frequency (RF) images by addressing a radar object detection task solely based on radar signals.

Object detection, which is aimed to detect different objects with their classes and locations, has been a crucial task for many applications. Recently, many effective image-based object detectors have been proposed [1]–[4] and widely used in autonomous or assisted driving systems [5]–[7]. Although cameras can give us better semantic understandings of visual scenes, it is not a robust sensor under adverse conditions, such as weak/strong lighting or bad weather, which lead to little/high exposure or blur/occluded images. On the other hand, LiDAR is an alternative sensor that can be used for accurate object detection and localization from its point cloud data. After the pioneer research on feature extraction from point cloud [8], [9], subsequent object detection from LiDAR point cloud has been addressed [10]–[12]. However, these methods require relatively dense LiDAR point cloud for detailed semantic information, not to mention its high equipment and computational costs.

Radar, on the other hand, is relatively more reliable in most harsh environments. Frequency modulated continuous wave (FMCW) radar, which operates in the millimeter-wave (MMW) band (30–300 GHz) that is lower than visible light, has the following properties: 1) MMW has great capability to penetrate through fog, smoke, and dust; 2) The huge bandwidth and high working frequency give FMCW radar great range detection ability. Typically, there are two kinds of data representations for the FMCW radar, i.e., RF image and radar points. The RF images are generated from the raw radar signals using a series of fast Fourier transforms (FFTs), and the radar points are then

1932-4553 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.

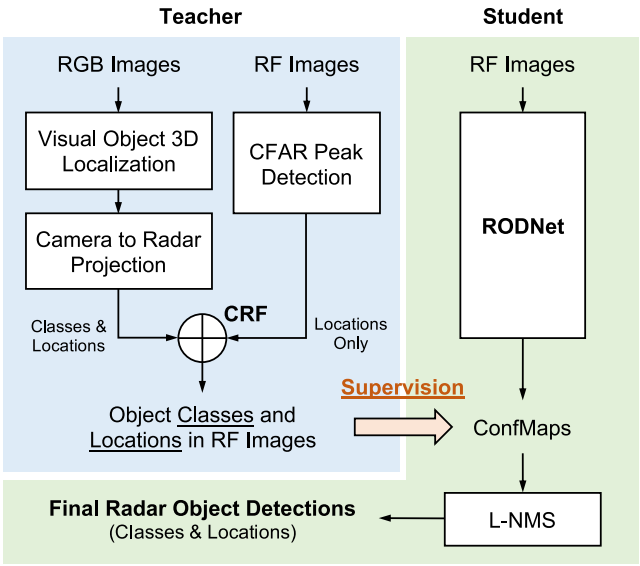


Fig. 1. The proposed cross-modal supervision pipeline for radar object detection. Teacher's pipeline fuses the results from both RGB and RF images to obtain the object classes and locations in RF images. Student's pipeline utilizes **only** RF images as the input to predict the corresponding ConfMaps under the supervision from the teacher. The L-NMS as post-processing is followed to calculate the final radar object detection results.

derived from these frequency images through a peak detection algorithm [13]. Although the radar points can be directly used like the input of the methods designed for the LiDAR point cloud [14], [15], these radar points are usually much sparser (less than 5 points on a nearby car) than the point cloud from a LiDAR [16], [17], so that the information is not enough to accomplish the object detection task. Whereas, the RF images can maintain the rich Doppler and object motion information so as to have the capability of understanding the semantic meaning of a certain object.

In this paper, we propose a radar object detection method, cross-supervised by a camera-radar fusion algorithm in the training stage, that can accurately detect objects purely based on the radar signals. More specifically, we propose a novel radar object detection pipeline, which consists of two parts: a teacher and a student. The teacher's pipeline estimates object classes and 3D locations in the field of view (FoV) by systematically fusing the information from a reliable camera-radar sensor fusion algorithm. The student's pipeline includes a radar object detection network (RODNet) that only takes RF image sequences as the input and estimates object confidence maps (ConfMaps, discussed in Section IV-C). From the ConfMaps, we can further infer the object classes and locations in the radar's range-azimuth coordinates through our post-processing method, called location-based non-maximum suppression (L-NMS, discussed in Section IV-C). The RODNet in the student's pipeline is trained by the annotations systematically labeled by the teacher pipeline without the laborious and unreliable human labeling efforts. The aforementioned proposed pipeline is shown in Fig. 1. As for the network architectures of the RODNet, we implement a 3D convolution neural network (3D CNN) based on an hourglass (HG) architecture with skip connections [18] for feature extraction from RF images. Several customized modules are designed in order to take advantage of the special properties of the RF image sequences. First, chirp information in each radar frame, which contains the detailed object features, is considered. Thus, a chirp merging module (M-Net) is proposed to combine the chirp-level features into the frame-level features. Second, since the radar reflection patterns are varying with time due to the relative motion between radar and objects, the classical 3D convolution cannot effectively extract temporal features. Thus, a novel convolution operation, called temporal deformable convolution (TDC), is proposed to handle the temporal evolution of the features in RF image sequences.

We train and evaluate the RODNet using our self-collected dataset, called Camera-Radar of the University of Washington (CRUW), which contains various driving scenarios of about 400 K synchronized camera-radar frames. As mentioned above, instead of using radar points as the data format, we choose RF images, which inherently contain detailed motion and surface texture information of objects. To assess the quantitative performance of our proposed RODNet, without the definition of bounding box widely used in image-based object detection, we further introduce an evaluation method to evaluate the radar object detection performance in RF images. With intensive experiments, our RODNet can achieve about 86% AP and 88% AR for object detection performance solely based on RF images in various driving scenarios, regardless of whether objects are visible or not in the camera's FoV.

Overall, our main contributions are the following:

- A novel and robust radar object detection network called RODNet for robust object detection in various driving scenarios, which can be used for autonomous or assisted driving without camera or LiDAR information.
- Customized modules, i.e., M-Net and temporal deformable convolution (TDC), are introduced to effectively take advantage of the special properties of RF images.
- A camera-radar fusion (CRF) supervision framework for training the RODNet, taking advantage of a monocular camera based object detection and 3D localization method facilitated with statistical detection inference of radar RF images.
- A new dataset, named CRUW, containing synchronized and calibrated camera-radar frames, is collected and can serve as a valuable dataset for camera/radar cross-modal research.
- A new evaluation method for RF image based radar object detection task is proposed and justified for its effectiveness.

The rest of this paper is organized as follows. Related works for camera and radar data learning are reviewed in Section II. The introduction on our proposed RODNet with customized modules is explained in Section III. The proposed CRF cross-modal supervision framework, obtaining reliable radar object annotations, is addressed in Section IV. In Section V, we present our self-collected CRUW dataset used for our training and evaluation. Then, the evaluation method and the experiments are shown in Section VI. Finally, we conclude our work in Section VII.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.

II. Related Works

A. Learning of Vision Data

Image-based object detection [1]–[4], [19]–[22], which is fundamental and crucial for many computer vision applications, is aimed to detect every object with its class and precise bounding box location from RGB images. Given the object detection results, most tracking algorithms focus on exploiting the associations between the detected object bounding boxes in consecutive frames, the so-called tracking-by-detection framework [23]–[28]. Among them, the TrackletNet Tracker (TNT) [26] is an effective and robust tracker to perform multiple object tracking (MOT) of the detected objects with a static or moving camera. Once the same objects among several consecutive frames are associated, the missing and erroneous detections can be recovered or corrected, resulting in better subsequent 3D localization performance. Thus, we implement this tracking technique into the vision part of our framework.

Object 3D localization has attracted great interest in autonomous and assisted driving communities [5], [6], [29]–[31]. One idea is to localize vehicles by estimating their 3D structures using a CNN, e.g., 3D bounding boxes [5] and 3D keypoints [6], [31], [32]. Then, a pre-defined 3D vehicle model is used to estimate the deformations, resulting in accurate 3D keypoints as well as the vehicle location. Another idea [29], [30], however, tries to develop a real-time monocular structure-from-motion (SfM) system, taking into account different kinds of cues, including SfM cues (3D points and ground plane) and object cues (bounding boxes and detection scores). Although these works achieve favorable performance in object 3D localization, they only work for the vehicles since only the rigid-body vehicle structure is assumed. To overcome this limitation, an accurate and robust object 3D localization system, based on the detected and tracked 2D bounding boxes of objects, is proposed in [7], which can work for most common moving objects in the road scenes, such as cars, pedestrians, and cyclists. Thus, this monocular camera based 3D localization system is adopted, fused with radar localization information, as our systematic camera annotation method to provide the ground truth for RODNet training.

B. Learning of Radar Data

Significant research in radar object classification has demonstrated its feasibility as a good alternative when cameras fail to provide good sensing performance [33]–[39]. With handcrafted feature extraction, Heuel, *et al.* [33] classify radar objects using a support vector machine (SVM) to distinguish cars and pedestrians. Moreover, Angelov *et al.* [34] use a neural network to extract features from the short-time Fourier transform (STFT) heatmap of radar signals. However, the above methods only focus on the *classification* tasks, which assume only one object has been appropriately identified in the scene. Recently, a radar object detection method is proposed in [40], which combines a statistical constant false alarm rate (CFAR) [13] detection algorithm with a CNN-based VGG-16 classifier [41]. All of the above approaches are not applicable to the complex driving scenarios with noisy background reflections, e.g., trees, buildings, traffic signs, etc., and could easily give many *false positives*. Besides, the laborious human annotations on the radar RF images required by these methods are usually impossible to obtain.

Recently, the concept of cross-modal learning has been proposed in the machine learning community [42]–[45]. This concept is trying to transfer or fuse the information between two different signal modalities in order to help train the neural networks. Specifically, RF-Pose [46] introduces the cross-modal supervision idea into wireless signals to achieve human pose estimation based on WiFi range radio signals, using a computer vision technique, i.e., OpenPose [47], to systematically generate annotations of human body keypoints from the camera. However, radar object detection is more challenging: 1) Feature extraction for object detection (especially for classification) is more difficult than human joint detection which could just classify different joints by their relative locations without considering object motion and texture information; 2) The typical FMCW radars on the vehicles have much less resolution than the WiFi array sensors used in RF-Pose. As for autonomous or assisted driving applications, Major *et al.* [48] propose an automotive radar based vehicle detection method using LiDAR information for cross-modal learning. However, our work is different from theirs: 1) They only consider vehicles as the target object class, while we detect pedestrians, cyclists, and cars; 2) The scenarios in their dataset are mostly highways without noisy obstacles, which is easier for radar object detection, while we are dealing with much more diverse driving scenarios. Palffy *et al.* [49] propose a radar based, single-frame multi-class object detection method. However, they only consider the data from a single radar frame, which does not involve the object motion information.

C. Datasets

Datasets are important to validate the algorithms, especially for the deep learning based methods. Since the first complete autonomous driving dataset, named KITTI [50], is published, larger and more advanced datasets are now available [16], [51], [52]. However, due to the hardware compatibility and less developed radar perception techniques, most datasets do not incorporate radar signals as a part of their sensor systems. Among the available radar datasets, nuScenes [16] and Astyx HiRes2019 [53] consider radar with good calibration and synchronization with other sensors. But their radar data format is based on sparse radar points that do not contain the useful Doppler and surface texture information of objects. While Oxford Radar RobotCar Dataset [54] contains dense radar point clouds, it does not provide any object annotations. After extensive research on the available datasets, we cannot find a suitable one that includes large-scale radar data in RF image format with labeled ground truth. Therefore, we collect our own CRUW dataset which will be introduced in Section V.

III. Radar Object Detection

In this section, the student's pipeline of our radar object detection is addressed. First, the raw radar signals are pre-transformed

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.

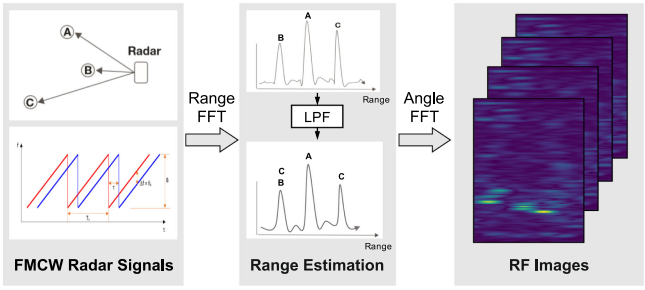


Fig. 2. The workflow of the RF image generation from the raw radar signals.

to RF images to be compatible with the image-based convolution neural networks (CNNs). After that, some special properties and challenges of RF images are analyzed. Second, the proposed RODNet is introduced with various functional components. Third, two customized modules are added to the RODNet to handle the above challenges. Finally, a post-processing method, called location-based non-maximum suppression (L-NMS), is adopted to recover ConfMaps for the final detections.

A. Radar Signal Processing and Properties

In this work, we use a common radar data representation, named radio frequency (RF) image, to represent our radar signal reflections. RF images are in radar range-azimuth coordinates and can be described as a bird's-eye view (BEV) representation, where the x-axis denotes azimuth (angle) and the y-axis denotes range (distance). For an FMCW radar, it transmits continuous chirps and receives the reflected echoes from the obstacles. After the echoes are received and pre-processed, we implement the fast Fourier transform (FFT) on the samples to estimate the range of the reflections. A low-pass filter (LPF) is then utilized to remove the high-frequency noise across all chirps in each frame at the rate of 30 FPS. After the LPF, we conduct a second FFT on the samples along different receiver antennas to estimate the azimuth angle of the reflections and obtain the final RF images. This RF image generation workflow is shown in Fig. 2. After being transformed into RF images, the radar data become a similar format as image sequences, which can thus be directly processed by an image-based CNN.

Moreover, radio frequency data have the following special properties to be handled for object detection task.

- **Rich motion information.** According to the Doppler principle of the radio signal, rich motion information is included. The objects' speed and its law of variation over time is dependent on their surface texture information, size and shape details, etc. For example, the motion information of a non-rigid body, like a pedestrian, is usually widely distributed, while for a rigid body, like a car, it should be more consistent due to the Doppler effect. In order to better utilize the temporal information, we need to consider multiple consecutive radar frames, instead of one single frame, as the system input.
- **Inconsistent resolution.** Radar usually has high-resolution in range but low-resolution in azimuth angle due to the limitation of radar hardware specifications, like the number of antennas and the distances among them.
- **Complex numbers.** Radio signals are usually represented as complex numbers containing frequency and phase information. This kind of data is unusual to be modeled by a typical CNN architecture.

According to the above properties, the proposed radar object detection method needs to have the following capabilities: 1) Extract temporal information; 2) Handle multiple spatial scales; 3) Be able to deal with complex number data. These capabilities will be discussed in the following sections.

B. RODNet Architecture

There are three major functional components adopted in constructing the network architecture of the RODNet, as shown in Fig. 3(a)–(c), which is implemented based on a 3D CNN with an autoencoder structure. More specifically, our RODNet starts with a naïve version of a 3D CNN autoencoder network, shown in Fig. 3(a). Then, built upon an hourglass based autoencoder [18], shown in Fig. 3(b), where skip connections are added to transmit the features directly from bottom layers to top layers. We further add the temporal inception convolution layers to extract different lengths of temporal features from the input RF image sequence, inspired by the spatial inception convolution layer proposed in [55], shown in Fig. 3(c).

The input of our network is a snippet of RF images $\boldsymbol{R}$ with dimension (C_{RF}, T, n, H, W), where C_{RF} is the number of channels in each complex-numbered RF images, referring [46], where the real and imaginary values are treated as two different channels in one RF image, i.e., $C_{RF} = 2$; T is the number of RF image frames in the snippet; n is the number of chirps in each frame; H and W are the height and width of the RF images, respectively.

After passing through the network, ConfMaps $\hat{\boldsymbol{D}}$ with dimension (C_{cls}, T, H, W) are predicted, where C_{cls} is the number of object classes. Note that RODNet predicts separate ConfMaps for each individual object class of radar RF images. With systematically derived binary annotations using the teacher's pipeline described in Section IV, we can train our RODNet using binary cross entropy loss,

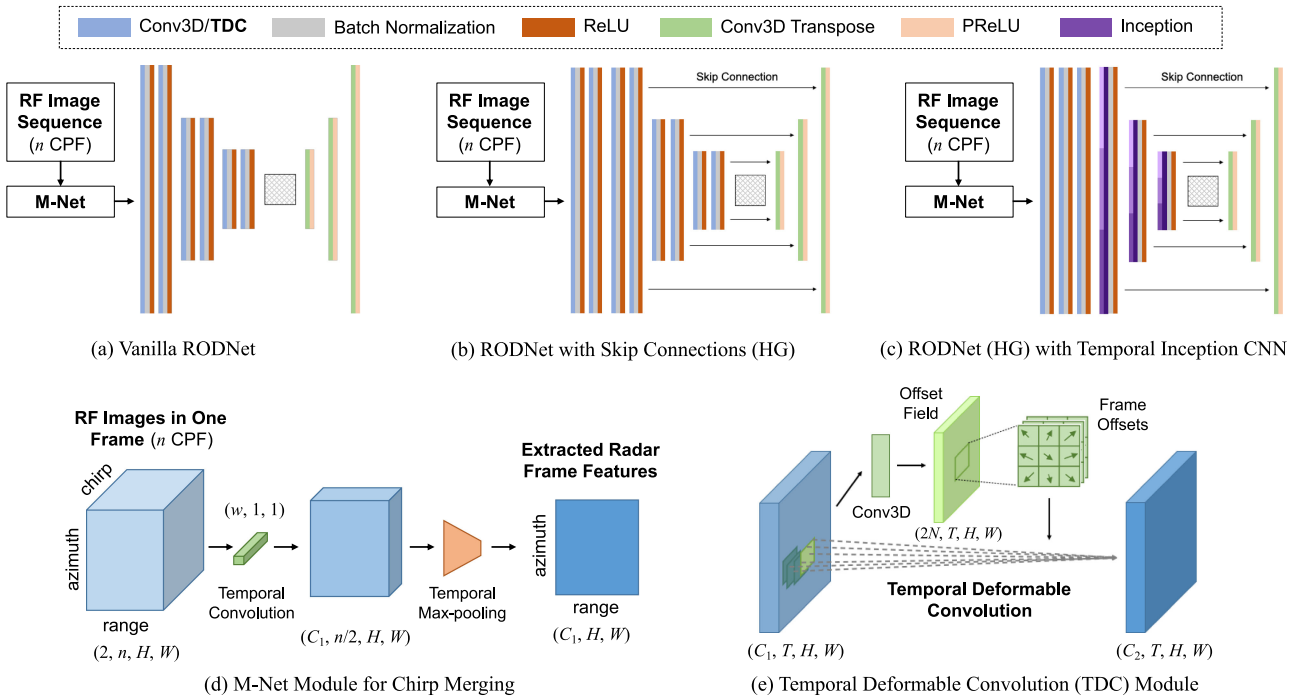
$$\ell = -\sum_{cls}\sum_{i,j} \boldsymbol{D}_{i,j}^{cls} \log \hat{\boldsymbol{D}}_{i,j}^{cls} + \left(1 - \boldsymbol{D}_{i,j}^{cls}\right) \log\left(1 - \hat{\boldsymbol{D}}_{i,j}^{cls}\right). \tag{1}$$

Here, $\boldsymbol{D}$ represents the ConfMaps generated from CRF annotations, $\hat{\boldsymbol{D}}$ represents the predicted ConfMaps, (i, j) represents the pixel indices, and cls is the class label.

C. M-Net Module

Besides the temporal features across different frames in each RF snippet, all the information from different chirps contribute to features for radar object detection. In order to better integrate this dynamic information from different chirps, we propose a

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.



(a) Vanilla RODNet (b) RODNet with Skip Connections (HG) (c) RODNet (HG) with Temporal Inception CNN

(d) M-Net Module for Chirp Merging (e) Temporal Deformable Convolution (TDC) Module

Fig. 3. The architecture and modules of our proposed RODNet. Three different components of the RODNet are all implemented based on the 3D CNN/TDC and autoencoder network, as shown in (a), (b), and (c). The input of the RODNet is RF images with n chirps per frame (CPF). When $n = 1$, we select only one chirp's data randomly to feed into the RODNet, while for $n > 1$, the M-Net module is implemented to merge the data from different chirps in this frame. The M-Net module, described in (d), which takes as input one frame with multiple chirps of radar data and outputs this frame's merged features. Moreover, the temporal deformable convolution (TDC) module in (e) is introduced to handle the radar object dynamic motion within the input RF image sequence.

customized module, called M-Net, before the RF snippets are sent into the RODNet. As shown in Fig. 3(d), the RF images of one frame with n chirps are sent into M-Net with a dimension of (C_{RF}, n, H, W), where $C_{RF} = 2$. First, a temporal convolution is applied to extract temporal features among the n chirps. This M-Net CNN operation performs like a Doppler compensated FFT to extract dynamic motion features but can be trained end-to-end in the deep learning architecture. Then, to merge the features from n chirps into one, the temporal max-pooling layer is applied. Finally, the output of M-Net is the extracted radar frame features with a dimension of (C_1, H, W), where C_1 is the number of filters for the temporal convolution. After M-Net is applied to each radar frame, the extracted features from all the frames in the input snippet are concatenated along time and sent as the input to the subsequent RODNet, as shown in Fig. 3(a)–(c).

D. Temporal Deformable Convolution

As mentioned in Section III-B, the input of the RODNet is a snippet of RF images after features are merged by the M-Net. Thus, during this period, locations of the objects in the radar range-azimuth coordinates may be shifted due to object relative motion, which means the patterns in RF images may change their locations within the snippet. However, the classical 3D convolution can only capture the static features within a regular cuboid, therefore it is not the best feature extractor for the RF snippets with object's relative motion.

Recently, Dai *et al.* [56] propose a new convolution network, named deformable convolution network (DCN), for image-based object detection to handle the deformed objects within the images. In the deformable convolution, the original convolution grid is deformable in the sense that each grid point is shifted by a learnable offset, and the convolution is operated on these shifted grid points.

Inspired by DCN, we generalize the deformed kernel to the 3D CNN and name this novel operator as temporal deformable convolution (TDC). Use the 3D CNN with kernel size of $(3, 3, 3)$ and dilation 1 as an example, the regular receptive field $\mathcal{R}$ can be defined as

$$\mathcal{R} = \{(-1, -1, -1), (-1, 0, 0), \ldots, (0, 1, 1), (1, 1, 1)\}. \quad (2)$$

For each location $\mathbf{p}_0$ on the output feature map $\mathbf{y}$, the classical 3D convolution can be described as

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n), \quad (3)$$

where $\mathbf{w}$ is the convolution kernel weight, $\mathbf{x}$ is the input feature map, and $\mathbf{p}_n$ enumerates the locations in $\mathcal{R}$.

In order to handle the object dynamic motion in temporal domain, we propose TDC by adding an additional offset field $\{\Delta\mathbf{p}_n\}_{n=1}^{N}$, where $N = |\mathcal{R}|$ is the size of the receptive field. So that Eq. 3 becomes

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n). \quad (4)$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.

Note that the offset field $\Delta\mathbf{p}_n$ is only deformed within each temporal location, i.e., the receptive location of a certain frame will not be deformed to other frames, so that the temporal domain of the offset field is always zero. To make it easy for implementation, the offset vectors are defined as 2D vectors so that the overall offset field has a dimension of $(2N, T, H, W)$. An illustration of our proposed TDC is shown in Fig. 3(e).

Similar to [56], since the offset field $\Delta\mathbf{p}_n$ is typically fractional, Eq. 4 is implemented via bilinear interpolation as

$$\mathbf{x}(\mathbf{p}) = \sum_{\mathbf{q}} G(\mathbf{q}, \mathbf{p}) \cdot \mathbf{x}(\mathbf{q}), \tag{5}$$

where $\mathbf{p} = \mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n$ is the fractional location; $\mathbf{q}$ enumerates all integer locations in the 3D feature map $\mathbf{x}$; and G is the bilinear interpolation kernel which is also two dimensional in the spatial domain. The back-propagation formulation of TDC is similar to that discussed in [56] except adding the temporal domain, and is described in the supplementary document.

E. Post-Processing by Location-Based NMS

After predicting ConfMaps from a given RF snippet, a post-processing step is still required to obtain the final detections. Here, we adopt the idea of non-maximum suppression (NMS), which is frequently used in image-based object detection to remove the redundant bounding boxes from the detection results. Traditionally, NMS uses intersection over union (IoU) as the criterion to determine if a bounding box should be removed due to its too much overlapping with the detection candidate of the highest confidence. However, there is no bounding box definition in our RF images nor the resulting output ConfMaps. Thus, inspired by object keypoint similarity (OKS) defined for human pose evaluation in the COCO dataset [57], we define a new metric, called object location similarity (OLS) that is similar to the role of IoU, to describe the correlation between two detections considering their distance, classes and scale information on ConfMaps. More specifically,

$$\text{OLS} = \exp\left\{\frac{-d^2}{2(s\kappa_{cls})^2}\right\}, \tag{6}$$

where d is the distance (in meters) between the two points in an RF image; s is the object distance from the radar sensor, representing object scale information; and κ_{cls} is a per-class constant that represents the error tolerance for class cls, which can be determined by the object average size of the corresponding class. We empirically determine κ_{cls} to make OLS distributed reasonably between 0 and 1. Here, we try to interpret OLS as a Gaussian distribution, where distance d acts as the bias and $(s\kappa_{cls})^2$ acts as the variance. Therefore, OLS is a metric of similarity, which also considers object sizes and distances, so that more reasonable than other traditional distance metrics, such as Euclidean distance, Mahalanobis distance, etc. This OLS metric is also used to *match detections and ground truth* for evaluation purpose, mentioned in Section VI-A.

After OLS is defined, we propose a location-based NMS (L-NMS) for post-processing. An example of the L-NMS is shown in Fig. 4, and the procedure can be summarized as follows:

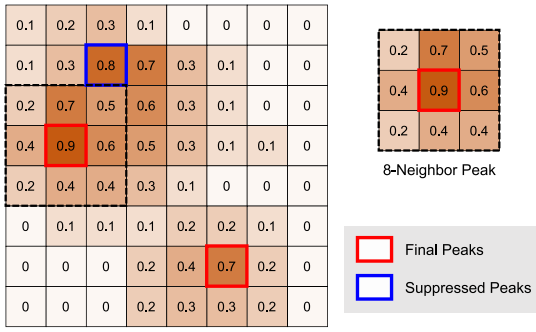


Fig. 4. Example for L-NMS on a ConfMap. The numbers represent the confidence scores predicted by the RODNet. The 8-neighbor peaks are first detected and some peaks are then suppressed if they are nearby some other peaks with higher confidence.

1) Get all the 8-neighbor peaks in all C_{cls} channels in ConfMaps within the 3×3 window as a peak set $P = \{p_n\}_{n=1}^{N}$.
2) Pick the peak $p^* \in P$ with the highest confidence, put it to the final peak set P^* and remove it from set P. Calculate OLS with each of the rest peaks p_i $(p_i \neq p^*)$.
3) If OLS between p^* and p_i is greater than a threshold, remove p_i from the peak set.
4) Repeat Steps 2 and 3 until the peak set becomes empty.

Moreover, during the inference stage, we can send overlapped RF snippets into the network, which provides different ConfMaps predictions for a single radar frame. Then, we average these different ConfMaps together to obtain the final ConfMaps results. This scheme can improve the system's robustness and can be considered as a performance-speed trade-off, which will be further discussed in Section VI-C.

IV. Cross-Modal Supervision

In this section, the teacher's pipeline that provides supervision for the RODNet is described. First, the annotation methods by camera-only and camera-radar fusion are introduced to obtain the accurate object classes and 3D locations. Then, these annotations are applied to ConfMaps as ground truth for training the network end-to-end.

A. Camera-Only (CO) Supervision

Object detection and 3D localization have been explored by researchers in computer vision community for many years. Here, we use Mask R-CNN [3] as our image-based object detector, that can provide object classes, bounding boxes, as well as their instance masks. While, object 3D localization is more challenging, due to the loss of the depth information during the camera projection from 3D world into 2D images. To recover the 3D information from 2D images, we take advantage of a recent work on an effective and robust system for visual object 3D localization based on a monocular camera [7]. The proposed system takes a CNN inferred depth map as the input, incorporating adaptive ground plane estimation and multi-object tracking results, to effectively estimate object classes and 3D

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.

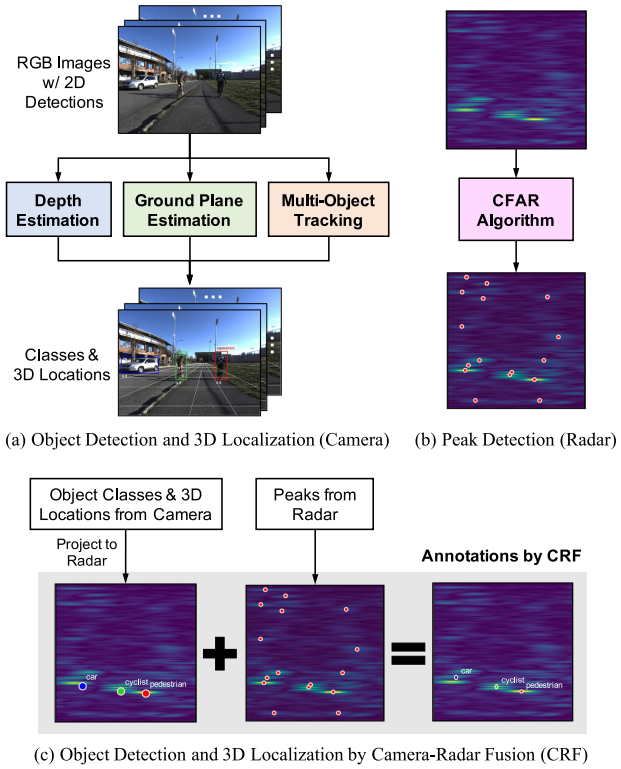


Fig. 5. Three teacher's pipelines for cross-model supervision. (a) Camera-only method that provides object classes and 3D locations; (b) Radar-only method that only provides peak locations without object class; (c) Camera-radar fusion method that provides object classes and more accurate 3D locations.

locations relative to the camera. The main strength of this camera-only system is that it can robustly estimate objects' 3D locations purely from monocular videos, resulting in a very low requirement on the visual data used for annotation, i.e., one single camera is good enough to obtain the annotation results. The rational behind this capability can be described in three folds: 1) Depth estimation used in the system is self-supervised by the stereo image pairs in training, so as to provide the absolute scale information missing in the monocular camera systems; 2) Adaptive ground plane estimation, based on sparse detected object feet points and dense semantic segmented ground points, is proposed to handle inaccurate depth within each frame; 3) A multi-object tracking technique is incorporated to address the inter-frame bias issue and temporally smooth the object 3D trajectories. A simplified illustration of the monocular camera object 3D localization system proposed in [7] is shown in Fig. 5(a). Stereo cameras can also be used for object 3D localization, however, high computational cost and sensitivity to camera setup configurations (e.g., baseline distance) result in the limitation of the stereo object 3D localization system.

However, as also observed in [58], the above camera-only system may not be accurate enough after transforming to the radar's range-azimuth coordinates because: 1) The systematic bias in the camera-radar sensor system that the peaks in the RF images may not be consistent with the 3D geometric center of the object; 2) Cameras' performance can be easily affected by lighting or weather conditions. Since we do have the radar information available, camera-radar cross calibration and supervision should be used. Therefore, an even more accurate self-annotation method, based on camera-radar fusion, is required for training the RODNet.

B. Camera-Radar Fusion (CRF) Supervision

An intuitive way of improving the above camera-only annotation is by taking advantage of radar, which has a plausible capability of range estimation without any systematic bias. Here, we adopt the Constant False Alarm Rate (CFAR) detection algorithm [13], which is commonly used in signal processing to detect peaks in the RF image. As shown in Fig. 5(b), the CFAR algorithm can detect a number of peaks in the RF image, denoted as red dots. However, these detected peaks cannot be directly used as supervision because 1) CFAR algorithm cannot provide the object classes for each detection; 2) CFAR algorithm usually gives a large number of false positive detections. Thus, an object localization method by camera-radar fusion strategy is needed to address these issues.

Fig. 5(c) illustrates the camera-radar fusion (CRF) pipeline, where the classes and 3D locations of the detected objects from the camera are first passed through a transformation to project the detections from 3D camera coordinates to radar range-azimuth coordinates. The transformation can be formulated as

$$\begin{aligned} \rho_c &= \sqrt{(x^c - x_{or})^2 + (z^c - z_{or})^2}, \\ \theta_c &= \tan^{-1}\left(\frac{x^c - x_{or}}{z^c - z_{or}}\right), \end{aligned} \tag{7}$$

where (ρ_c, θ_c) denotes the projected location in radar range-azimuth coordinates; (x^c, z^c) denotes the object location in the camera BEV coordinates; and (x_{or}, z_{or}) denotes the location of radar origin in the camera BEV coordinates, aligned from the sensor system calibration. The peak detections from the CFAR algorithm are also involved in the same radar range-azimuth coordinates. Finally, the fusion algorithm is applied to estimate the final annotations on the input RF image.

After the coordinates between camera and radar are aligned, a probabilistic CRF algorithm is further developed to achieve a more reliable and systematic annotation performance. The basic idea of this algorithm is to generate two probability maps for camera and radar locations separately, and then fuse them by element-wise product. The probability map for camera locations with object class cls is generated by

$$\begin{aligned} \mathcal{P}^c_{(cls)}(\mathbf{x}) = \max_i \Bigg\{ \mathcal{N}\Bigg(& \frac{1}{2\pi\sqrt{|\mathbf{\Sigma}^c_{i(cls)}|}} \exp \\ & \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}^c_i)^\top (\mathbf{\Sigma}^c_{i(cls)})^{-1} (\mathbf{x} - \boldsymbol{\mu}^c_i) \right\} \Bigg) \Bigg\}, \\ \boldsymbol{\mu}^c_i = \begin{bmatrix} \rho^c_i \\ \theta^c_i \end{bmatrix}, & \mathbf{\Sigma}^c_{i(cls)} = \begin{bmatrix} \left(d_i s_{(cls)}/c_i\right)^2 & 0 \\ 0 & \delta_{(cls)} \end{bmatrix}. \end{aligned} \tag{8}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.

Here, d_i is the object depth, $s_{(cls)}$ is the scale constant, c_i is the depth confidence, and $\delta_{(cls)}$ is the typical azimuth error for camera localization. $\mathcal{N}(\cdot)$ represents the normalization operation for each object's probability map. Similarly, the probability map for radar locations is generated by

$$\mathcal{P}^r(\mathbf{x}) = \max_j \left\{ \mathcal{N}\left(\frac{1}{2\pi\sqrt{|\mathbf{\Sigma}_j^r|}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j^r)^\top (\mathbf{\Sigma}_j^r)^{-1}(\mathbf{x}-\boldsymbol{\mu}_j^r) \right\} \right) \right\},$$

$$\boldsymbol{\mu}_j^r = \begin{bmatrix} \rho_j^r \\ \theta_j^r \end{bmatrix}, \mathbf{\Sigma}_j^r = \begin{bmatrix} \delta_j^r & 0 \\ 0 & \epsilon(\theta_j^r) \end{bmatrix}. \quad (9)$$

Here, δ_j^r is the radar's range resolution, and $\epsilon(\cdot)$ is the radar's azimuth resolution. Then, an element-wise product is used to obtain the fused probability map for each class,

$$P_{(cls)}^{CRF}(\mathbf{x}) = P_{(cls)}^{c}(\mathbf{x}) * P^r(\mathbf{x}). \quad (10)$$

Finally, the fused annotations are derived from the fused probability maps P^{CRF} by peak detection.

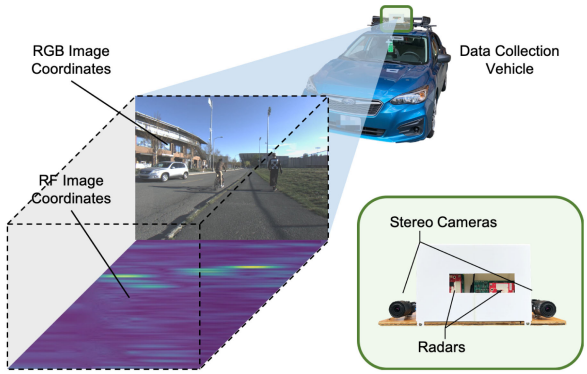
The object 3D localization accuracy of both CO and CRF annotation is discussed later in Section VI-B.

C. ConfMap Generation

After objects are accurately localized in the radar range-azimuth coordinates, we need to transform the results into a proper representation that is compatible with our RODNet. Considering the idea in [47] that defines the human joint heatmap to represent joint locations, we define the confidence map (ConfMap) in range-azimuth coordinates to represent object locations. One set of ConfMaps has multiple channels, where each channel represents one specific class label, i.e., car, pedestrian, and cyclist. The value at the pixel in the cls-th channel represents the probability of an object with class cls occurring at that range-azimuth location. Here, we use Gaussian distributions to set the ConfMap values around the object locations, whose mean is the object location, and the variance is related to the object class and scale information.

V. CRUW Dataset

Going through some existing datasets for autonomous driving discussed in Section II-C, the 3D radar point format is commonly used. While, it does not contain the discriminating object motion and surface texture information that is needed for our radar object detection task. In order to efficiently train and evaluate our RODNet using radar data, we collect a new dataset, named Camera-Radar of the University of Washington (CRUW), which uses the format of RF images for the radar data, as mentioned in Section III-A. Our sensor platform contains a pair of stereo cameras [59] and two perpendicular 77 GHz FMCW MMW radar antenna arrays [60]. The sensors, assembled and mounted together as shown in Fig. 6(a), are well-calibrated and synchronized. Some configurations of our sensor platform in shown in Table I. Even though our final cross-modal supervision requires just one monocular camera, the stereo cameras are setup to provide ground truth of depth for performance validation of the CRF supervision.



(a) Illustration for CRUW dataset collection

(b) Different scenarios in CRUW dataset

Fig. 6. The sensor platform and driving scenarios for our CRUW dataset.

TABLE I
SENSOR CONFIGURATIONS FOR CRUW DATASET

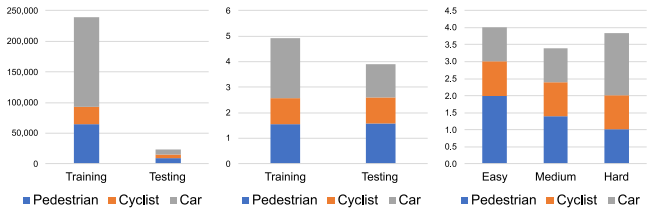
Camera	Value	Radar	Value
Frame rate	30 FPS	Frame rate	30 FPS
Pixels (H×W)	1440×1080	Frequency	77 GHz
Resolution	1.6 MegaPixels	# of transmitters	2
Field of View	93.6°	# of receivers	4
Stereo Baseline	0.35 m	# of chirps per frame	255
		Range resolution	0.23 m
		Azimuth resolution	∼15°

TABLE II
DRIVING SCENARIOS STATISTICS FOR CRUW DATASET

Scenarios	# of Seqs	# of Frames	Vision-Hard %
Parking Lot	124	106K	15%
Campus Road	112	94K	11%
City Street	216	175K	6%
Highway	12	20K	0%
Overall	464	396K	9%

The CRUW dataset contains 3.5 hours with 30 FPS (about 400 K frames) of camera-radar data in different driving scenarios, including campus road, city street, highway, and parking lot. Some sample scenarios are shown in Fig. 6(b). The data are

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.



(a) # of objects in total (b) # of objects per frame (c) # of objects per frame in the testing set

Fig. 7. Illustration for our CRUW dataset object distribution in the radar's FoV (0-25 m, ±90°).

collected in two different views, i.e., driver front view and driver side view, to ensure that our method is applicable to different perspective views for autonomous or assisted driving. Besides, we also collect several vision-hard sequences of poor image quality, i.e., weak/strong lighting, blur, etc. These data are only used for testing to illustrate that our method can still be reliable when vision techniques most likely fail.

The data distribution of CRUW is shown in Fig. 7. The object statistics in (a)–(c) only consider the objects within the radar's field of view (FoV), i.e., 0-25 m, ±90°, based on the current hardware capability. There are about 260 K objects in CRUW dataset in total, including 92% for training and 8% for testing. The average number of objects in each frame is similar between training and testing data. As for the testing set, we split it into three difficulty levels, i.e., easy, medium, and hard, to evaluate the performance under different scenarios. The criteria for this split include the number of objects, clear/noisy background, normal/extreme lighting, and object relative motion. The four different driving scenarios included in the CRUW dataset are shown in Table II with the number of sequences, frames and vision-hard percentages. From each scenario, we randomly select several complete sequences as testing sequences, which are not used for training. Thus, the training and testing sequences are captured at different locations and different time. For the ground truth needed for evaluation purposes, we annotate 10% of the visible and 100% of the vision-hard data. The annotations are operated on the RF images by labeling the object classes and locations according to the corresponding RGB and RF images.

VI. Experiments

A. Evaluation Metrics

To evaluate the performance, we utilize our proposed object location similarity (OLS) (see Eq. 6) in Section III-E, replacing the role of IoU widely used in image-based object detection, to determine how well a detection result can be matched with a ground truth. During the evaluation, we first calculate OLS between each detection result and ground truth in every frame. Then, we use different thresholds from 0.5 to 0.9 with a step of 0.05, for OLS and calculate the average precision (AP) and average recall (AR) for different OLS thresholds, which represent different localization error tolerance for the detection results. Here, we use AP and AR to represent the average values among all different OLS thresholds from 0.5 to 0.9, and use AP^{OLS} and AR^{OLS} to represent the values at a certain OLS threshold. Overall, we use AP and AR as our main evaluation metrics for the radar object detection task.

B. Radar Object Detection Results

We train our RODNet using the training data with CRF annotations in the CRUW dataset. For testing, we perform inference and evaluation on the human-annotated data. The quantitative results are shown in Table III. We compare our RODNet results with the following baselines that also use radar-only inputs: 1) A decision tree using some handcrafted features from radar data [40]; 2) CFAR detection is first implemented and a radar object classification network with ResNet backbone [34] is appended; 3) Similar with 2), a radar object classification network with VGG-16 backbone based on CFAR detections mentioned in [40]. Among all the three competing methods, the AR performance for [34], [40] is relatively stable in all three different test sets, but their APs vary a lot. Especially, the APs drop from around 80% to 10% for easy to hard testing sets. This is caused by a large number of false positives detected by the traditional CFAR algorithm, which would significantly decrease the precision.

Comparing with the above baseline methods, our RODNet outperforms significantly on both AP and AR metrics, achieving the best performance of 85.98% AP and 87.86% AR, especially the sustained performance on the medium and hard testing sets, which shows the robustness to noisy scenarios. Note that the results of RODNet shown in Table III include all the components proposed for RODNet, i.e., CRF supervision, M-Net, TDC, and temporal inception CNN.

Some qualitative results are shown in Fig. 8, where we can find that the RODNet can accurately localize and classify multiple objects in different scenarios. The examples in Fig. 8 consist of RGB and RF image pairs as well as RODNet detection results under different driving scenarios and conditions, including parking lot, campus road, and city street, with different lighting conditions. Some other examples to show the special strengths of our RODNet are shown in Section VI-D.

To illustrate our teacher's pipeline is qualified for this cross-supervision task, we evaluate the object 3D localization performance for both CO and CRF annotations in Table V. Besides, we also compare the performance between CO/CRF supervision and our RODNet on both visible (V) and vision-hard (VH) data. The results are shown in Fig. 9 with respect to different OLS thresholds. From Fig. 9, the performance of the vision-based method drops significantly given a tighter OLS threshold, while our RODNet shows its superiority and robustness on its localization performance. Moreover, the RODNet can still maintain the performance on vision-fail data where the vision-based methods have a hard time maintaining the performance.

C. Ablation Studies

In this section, we analyze the performance-speed trade-off of several components of the RODNet and dive into some details to show how our method can accomplish this radar object detection task very well.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.

TABLE III
RADAR OBJECT DETECTION PERFORMANCE EVALUATED ON CRUW DATASET

Methods	Overall		Easy		Medium		Hard	
	AP	AR	AP	AR	AP	AR	AP	AR
Decision Tree [41]	4.70	44.26	6.21	47.81	4.63	43.92	3.21	37.02
CFAR+ResNet [35]	40.49	60.56	78.92	85.26	11.00	33.02	6.84	36.65
CFAR+VGG-16 [41]	40.73	72.88	85.24	88.97	47.21	62.09	10.97	45.03
RODNet (Ours)	**85.98**	**87.86**	**96.97**	**98.02**	**76.11**	**78.57**	**67.28**	**72.60**

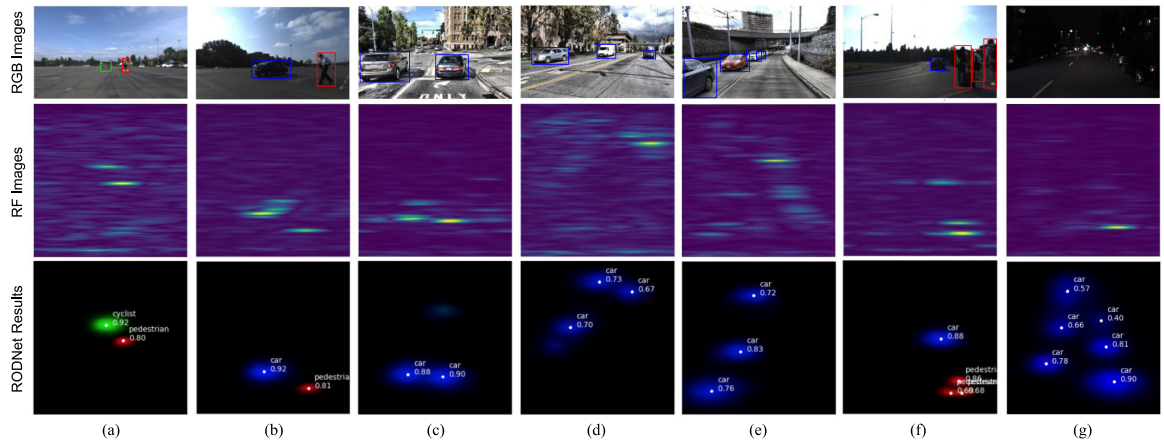


Fig. 8. Examples of detection results from our RODNet. The first row shows the RGB images and the second row shows the corresponding RF images. The ConfMaps predicted by the RODNet are shown in the third row, where the white dots represent the final detections after post-processing. Different colors represent different detected object classes (red: pedestrian; green: cyclist; blue: car). Various driving scenarios are shown, i.e., clear parking lot, crowded city street, and strong/weak lighting condition. More qualitative results are presented in the supplementary materials.

TABLE IV
ABLATION STUDIES ON THE PERFORMANCE IMPROVEMENT BY SEVERAL FUNCTIONAL COMPONENTS IN THE RODNET

Backbone	Supervision	M-Net	TDC	Inception	AP	$AP^{0.5}$	$AP^{0.7}$	$AP^{0.9}$	AR	$AR^{0.5}$	$AR^{0.7}$	$AR^{0.9}$
Vanilla	CO				52.62	78.21	54.66	18.92	63.95	84.13	68.76	30.71
	CRF				74.29	78.42	76.06	64.58	77.85	80.05	78.93	71.72
	CRF	✓			78.36	82.73	81.03	65.82	81.54	84.51	83.39	73.53
	CRF	✓	✓		79.86	84.08	82.37	66.74	82.85	86.06	84.43	73.93
Hourglass	CO				73.86	80.34	74.94	61.16	79.87	83.94	80.73	71.39
	CO			✓	77.75	82.88	79.93	61.88	81.11	85.13	82.78	68.63
	CRF				81.10	84.71	83.08	70.21	84.26	86.54	85.42	77.44
	CRF	✓			83.37	87.51	86.04	71.11	85.64	88.55	87.19	77.37
	CRF			✓	83.76	87.99	86.00	70.88	85.62	88.79	87.37	76.26
	CRF	✓	✓		84.38	88.69	85.73	73.31	86.97	89.67	88.14	79.59
	CRF	✓	✓	✓	**85.98**	**88.77**	**87.78**	**76.34**	**87.86**	**89.93**	**89.02**	**81.26**

TABLE V
THE MEAN LOCALIZATION ERROR (STANDARD DEVIATION) OF CO/CRF ANNOTATIONS ON CRUW DATASET (IN METERS)

Supervision	Pedestrian	Cyclist	Car
CO	0.69 ($\pm$0.77)	0.87 ($\pm$0.89)	1.57 ($\pm$1.12)
CRF	0.67 ($\pm$0.55)	0.82 ($\pm$0.59)	1.26 ($\pm$0.64)

1) Performance-Speed Trade-off of the RODNet: Since our RODNet starts with an M-Net chirp-merged input to the vanilla autoencoder with 3D convolution layers, further added with skip connections in hourglass (HG) structure and inception convolution layers, and eventually with TDC incorporation, it is important to know the performance influence of each functional module, as well as the computational complexities involved.

First of all, APs under different OLS thresholds are evaluated in Table IV. Here, we use different combinations of the backbones, supervision, and other modules in the RODNet.[2] From the results in the table, the following conclusions can be reached: 1) The performance of the HG backbone is better than the vanilla backbone by around 5%. 2) Training the RODNet using CRF supervision can improve the performance by about 8%. 3) The customized modules, i.e., M-Net and TDC, along with temporal inception, can also each improve the detection performance by approximately $1\% - 2\%$, respectively.

Moreover, *real-time* implementation is very important for autonomous or assisted driving applications. As mentioned in Section III-E, we use different overlapping lengths of RF frames

[2]Some experimental results shown in Table IV are also mentioned in our previous work [58].

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.

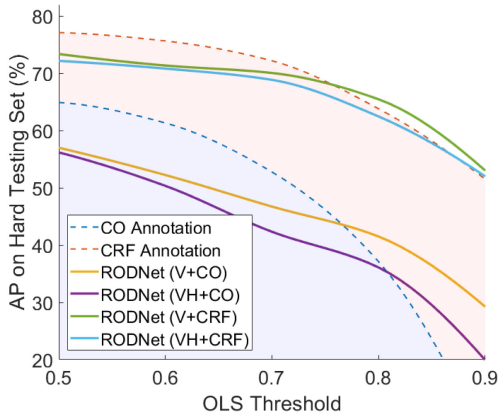


Fig. 9. Performance of vision-based and our RODNet on hard testing set with different OLS thresholds, representing localization error tolerance. (CO: camera-only; CRF: camera-radar fusion; V: visible data; VH: vision-hard data.).

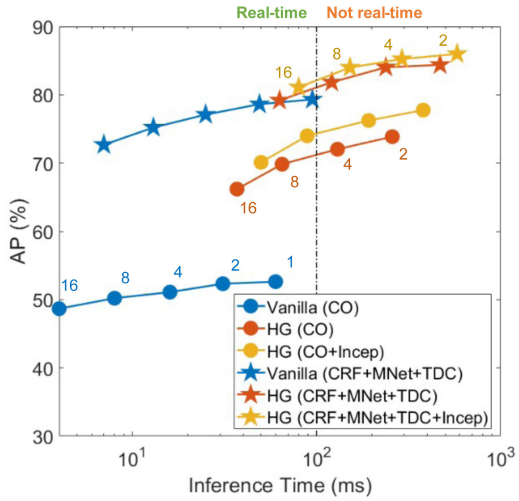


Fig. 10. Performance-speed trade-off for real-time implementation. Here, the inference time of less than 100 ms is used as the real-time criterion. We use the snippet length of 16, and the numbers besides the markers are the steps between each two overlapped RF snippets.

during the inference. With more overlapped frames, more robust detection results from the RODNet can be achieved, however, the inference time will also increase. The training and inference of the RODNet models are run on an NVIDIA Quadro GV100 GPU, and the time consumed is reported in Fig. 10. Here, we show the AP of three building architectures (Vanilla, HG, and HG with temporal inception) for the RODNet, and use 100 ms as a reasonable real-time threshold. The results illustrate that the RODNet with relatively simpler vanilla backbone can achieve real-time and finish the prediction within 100 ms. As for the HG backbone, it steps across the real-time threshold when the overlapping length increases. Moreover, HG without temporal inception layers is slightly faster than HG with all network components.

2) RF Snippet Length: Because our RODNet takes RF image snippets as the input, we would like to know how long the radar signals are required to obtain good detection results. Thus, we try different lengths of the snippets and evaluate their AP on our test set. The results are shown in Fig. 12(a). The experiments are operated on the backbone of HG (without temporal inception) with CRF, M-Net and TDC. From the figure, AP is low with short input snippets because of insufficient radar information from a short temporal period. The detection AP of RODNet reaches the highest at length of 16. While the snippet length continually increases, the AP starts to drop, due to the snippet being too long to extract efficient features for radar object detection.

3) Number of Chirps: The number of chirps sent into the M-Net is another important parameter for our method. A large number of chirps means more data are fed into the RODNet, resulting in higher input dimensions and time complexity. The experiments are also operated on the backbone of HG (without temporal inception) with CRF, M-Net and TDC. As shown in Fig. 12(b), the performance boosts with the number of chirps increases. However, it turns flat when the number of chirps is greater than 8. Therefore, we choose 8 to be the number of input chirps. Referring to the overall number of chirps 255 per $\frac{1}{30}$-second RF frame used in the CRUW dataset, we use 8 chirps (about 3% of the radar data) to achieve favorable detection performance, which shows the efficiency of our method.

4) Extracted Feature Visualization: After the RODNet is well-trained, we would like to analyze the features learned from the radar RF images. In Fig. 13, we show two different kinds of feature maps, i.e., the features after the first convolution layer and the features before the last layer of the RODNet. These feature maps are generated by cropping some randomly chosen objects from the original feature maps and average over all channels into one. From the visualization, we notice that the feature maps are similar in the beginning, and become significantly discriminative toward the end of the RODNet. Note that the visualized features are pixel-wise averaged within each object class to better represent the general class-level features.

D. Strengths Comparing With Visual Object Detection

Some examples to illustrate the RODNet's advantages are shown in Fig. 11. First, the RODNet can maintain similar detection performance in different driving conditions, even during the night, as shown in the first example. Moreover, the RODNet can handle some occlusion cases when the camera can easily fail. In the second example, two pedestrians are nearly fully occluded in the image, but our RODNet can still separately detect both of them. This is because they are separated in the range of the radar point of view, where 3D information is revealed by the radar. Third, RF images have a wider FoV than RGB images so that they can see more information. As shown in the third example, only a small part of the car visible in the camera view, which can hardly be detected from the camera side, but the RODNet can successfully detect it. Besides, RODNet can detect objects that are not detected by image-based methods due to strong lighting conditions. The fourth example shows that a pedestrian and a cyclist are detected by RODNet, but can hardly observed in the corresponding RGB image. Last but not least, our RODNet is able to distinguish noisy obstacles from objects after trained by CRF supervision. In the fifth and the last example, the noises,

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.

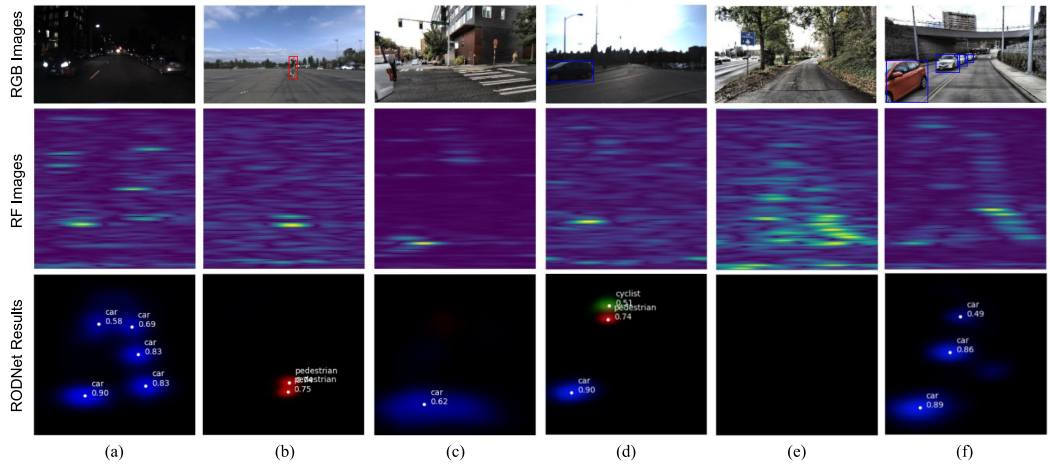
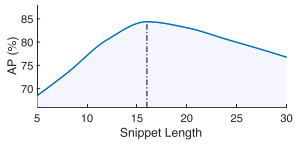
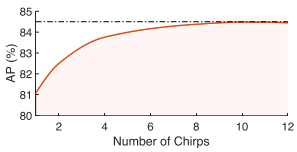


Fig. 11. Examples illustrate the strengths of our RODNet. Conditions from left to right: (a) difficult for vision-based methods during the night; (b) visually occluded objects can be separately detected; (c) visually truncated vehicle can be detected; (d) pedestrian and cyclist fail to be detected under strong lighting condition; (e) & (f) good detection results with a noisy background.



(a) AP for different snippet lengths



(b) AP for different numbers of chirps

Fig. 12. Results of different RF snippet length and number of chirps.

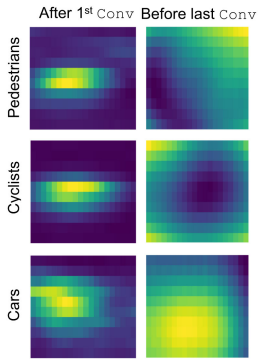


Fig. 13. Feature visualization of different object classes.

e.g., trees, walls, traffic signs and poles are suppressed in our RODNet results.

VII. Conclusion

Object detection is crucial in autonomous driving and many other areas. Computer vision community has been focusing on this topic for decades and come up with many good solutions. However, vision-based detection schemes are still suffering from many severe lighting and weather conditions. This paper proposed a brand-new and novel object detection method purely from radar information, which can be more robust than vision in adverse conditions. The proposed RODNet can accurately and robustly detect objects, based on fully systematic cross-modal supervision scheme from an effective camera-radar fusion algorithm, in various autonomous and assisted driving scenarios even during the night or bad weather, which can potentially improve the role of radar in autonomous and assisted driving applications.

Acknowledgment

The authors would also like to thank the colleagues and students in Information Processing Lab at UWECE for their help and assistance on the dataset collection, processing, and annotation works.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[2] 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.

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

[4] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.

[5] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7074–7082.

[6] J. A. Ansari, S. Sharma, A. Majumdar, J. K. Murthy, and K. M. Krishna, "The earth ain't flat: Monocular reconstruction of vehicles on steep and graded roads from a moving camera," in *Proc. IEEE/RSJ Int. Conf. Intell./ Robots Syst.*, 2018, pp. 8404–8410.

[7] Y. Wang, Y.-T. Huang, and J.-N. Hwang, "Monocular visual object 3D localization in road scenes," in *Proc. 27th ACM Int. Conf. Multimedia.* 2019, pp. 917–925.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.

[8] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[9] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5099–5108.

[10] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3D object detection from RGB-D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 918–927.

[11] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 770–779.

[12] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020, p. 2, doi: 10.1109/TPAMI.2020.2977026.

[13] M. A. Richards, *Fundamentals of Radar Signal Processing*. New York, NY, USA: Tata McGraw-Hill Educ., 2005.

[14] O. Schumann, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic segmentation on radar point clouds," in *Proc. 21st IEEE Int. Conf. Inf. Fusion*, 2018, pp. 2179–2186.

[15] N. Scheiner *et al.*, "Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2068–2077.

[16] H. Caesar *et al.*, "Nuscenes: A multimodal dataset for autonomous driving," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11621–11631.

[17] D. Feng *et al.*, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, 2020, pp. 1–20, doi: 10.1109/TITS.2020.2972974.

[18] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.

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

[20] W. Liu *et al.*, "Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[21] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.

[22] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6569–6578.

[23] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 941–951.

[24] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5188–5197.

[25] Z. Tang and J.-N. Hwang, "MOANA: An online learned adaptive appearance model for robust multiple object tracking in 3D," *IEEE Access*, vol. 7, pp. 31934–31945, 2019, doi: 10.1109/ACCESS.2019.2903121.

[26] G. Wang, Y. Wang, H. Zhang, R. Gu, and J.-N. Hwang, "Exploit the connectivity: Multi-object tracking with trackletnet," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 482–490.

[27] J. Cai, Y. Wang, H. Zhang, H.-M. Hsu, C. Ma, and J.-N. Hwang, "IA-MOT: Instance-aware multi-object tracking with motion consistency," 2020, *arXiv:2006.13458*.

[28] H.-M. Hsu, Y. Wang, and J.-N. Hwang, "Traffic-aware multi-camera tracking of vehicles based on reid and camera link model," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 964–972.

[29] S. Song and M. Chandraker, "Robust scale estimation in real-time monocular SFM for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1566–1573.

[30] S. Song and M. Chandraker, "Joint SFM and detection cues for monocular 3 D localization in road scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3734–3742.

[31] J. K. Murthy, G. S. Krishna, F. Chhaya, and K. M. Krishna, "Reconstructing vehicles from a single image: Shape priors for road scene understanding," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 724–731.

[32] H. Kim, B. Liu, and H. Myung, "Road-feature extraction using point cloud and 3D lidar sensor for vehicle localization," in *Proc. 14th IEEE Int. Conf. Ubiquitous Robots Ambient Intell.*, 2017, pp. 891–892.

[33] S. Heuel and H. Rohling, "Two-stage pedestrian classification in automotive radar systems," in *Proc. 12th Int. Radar Symp.*, Sep. 2011, pp. 477–484.

[34] A. Angelov, A. Robertson, R. Murray-Smith, and F. Fioranelli, "Practical classification of different moving targets using automotive radar and deep neural networks," *IET Radar, Sonar Navigation*, vol. 12, no. 10, pp. 1082–1089, 2018.

[35] S. Capobianco, L. Facheris, F. Cuccoli, and S. Marinai, "Vehicle classification based on convolutional networks applied to FMCW radar signals," in *Proc. Italy Conf. Traffic Police*., 2017, pp. 115–128.

[36] J. Kwon and N. Kwak, "Human detection by neural networks using a low-cost short-range doppler radar sensor," in *Proc. IEEE Radar Conf.*, 2017, pp. 0755–0760.

[37] P. Cao, W. Xia, M. Ye, J. Zhang, and J. Zhou, "RADAR-ID: Human identification based on radar micro-doppler signatures using deep convolutional neural networks," *IET Radar, Sonar Navigation*, vol. 12, no. 7, pp. 729–734, Mar. 2018.

[38] R. Pérez, F. Schubert, R. Rasshofer, and E. Biebl, "Single-frame vulnerable road users classification with a 77 GHz FMCW radar sensor and a convolutional neural network," in *Proc. 19th IEEE Int. Radar Symp.*, 2018, pp. 1–10.

[39] K. Patel, K. Rambach, T. Visentin, D. Rusev, M. Pfeiffer, and B. Yang, "Deep learning-based object classification on automotive radar spectra," in *Proc. IEEE Radar Conf.*, 2019, pp. 1–6.

[40] X. Gao, G. Xing, S. Roy, and H. Liu, "Experiments with mmWave automotive radar test-bed," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, 2019, pp. 1–6.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Int. Conf. Learn. Representations*, 2015.

[42] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.

[43] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4534–4542.

[44] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via siamese convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 2460–2464.

[45] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," 2019, p. 1, doi: 10.1109/TPAMI.2020.2992393.

[46] M. Zhao *et al.*, "Through-wall human pose estimation using radio signals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7356–7365.

[47] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.

[48] B. Major *et al.*, "Vehicle detection with automotive radar using deep learning on range-Azimuth-doppler tensors," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 924–932.

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

[50] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[51] "Apollo scape dataset," 2018. [Online]. Available: http://apolloscape.auto/

[52] "Waymo open dataset: An autonomous driving dataset," 2019. [Online]. Available: https://www.waymo.com/open

[53] M. Meyer and G. Kuschk, "Automotive radar dataset for deep learning based 3D object detection," in *Proc. 16th IEEE Eur. Radar Conf.*, 2019, pp. 129–132.

[54] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 6433–6438.

[55] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[56] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.

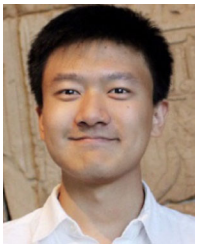
[57] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[58] Y. Wang, Z. Jiang, X. Gao, J.-N. Hwang, G. Xing, and H. Liu, "RODNET: Object detection under severe conditions using vision-radio cross-modal supervision," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 504–513.

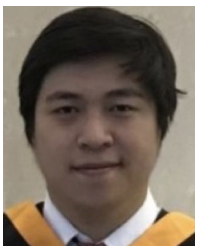
[59] "Flir systems," 2019. [Online]. Available: https://www.flir.com/

[60] "Texas instruments," 2019. [Online]. Available: http://www.ti.com/

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.



Yizhou Wang (Student Member, IEEE) received the M.S. degree in electrical engineering in 2018 from Columbia University, New York City, NY, USA, advised by Prof. Shih-Fu Chang. He is currently working toward the Ph.D. degree in electrical and computer engineering with the University of Washington, Seattle, WA, USA, advised by Prof. Jenq-Neng Hwang. His research interests include autonomous driving, computer vision, deep learning, and cross-modal learning.



Zhongyu Jiang (Student Member, IEEE) received the B.E. degree in computer science and technology from Tsinghua University, Beijing, China, in 2018 and the M.Sc. degree in computer science and system from the University of Washington, Tacoma, WA, USA, where he is currently working toward the Ph.D. degree in electrical and computer engineering.

Yudong Li is currently working toward the B.S. degree in informatics an undergraduate junior with Information School, the University of Washington, Seattle, WA, USA, and planning on graduation in 2022. He is currently a Research Assistant with Information Processing Lab, University of Washington, facilitating the works on machine learning related topics.

Jenq-Neng Hwang (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, CA, USA. In the summer of 1989, he joined the Department of Electrical and Computer Engineering, University of Washington, Seattle WA, USA, where since 1999, he has been a Full Professor. From 2003 to 2005 and from 2011 to 2015, he was the Associate Chair for research. From 2015 to 2020, he was also the Associate Chair for Global Affairs. He is the Founder and a Co-Director with Information Processing Lab, which was the recipient of CVPR AI City Challenges Awards in the past years. He has authored or coauthored more than 380 journals, conference papers, and book chapters in the areas of machine learning, multimedia signal processing, and multimedia system integration, and networking, including an authored textbook *Multimedia Networking: from Theory to Practice*, published by Cambridge University. He has close working relationship with the industry on multimedia signal processing and multimedia networking. He was the recipient of the 1995 IEEE Signal Processing Society's Best Journal Paper Award. He is a Founding Member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society and was the Society's representative to the IEEE Neural Network Council from 1996 to 2000. He is currently a Member of the Multimedia Technical Committee of the IEEE Communication Society and a Member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society. He was also an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON IMAGE PROCESSING and the *Signal Processing Magazine*. He is currently on the Editorial Board of the *ZTE Communications*, the *ETRI*, the *IJDMB* and the *JSPS* journals. He is currently the Program Co-Chair of the IEEE International Conference on Multimedia & Expo 2016 and was the Program Co-Chair of the International Conference on Acoustics, Speech, and Signal Processing 1998 and International Symposium on Circuits and Systems 2009.

Guanbin Xing received the B.S. and M.S. degrees in electrical engineering from Peking University, Beijing, China, in 1996 and 1999, respectively, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, WA, USA, in 2004. He was a Senior System Artitect with Wireless Communications and Digital Broadcasting industry having an experience of more than 15 years. In 2017, he was a Research Scientist with CMMB Vision, UW EE Center on Satellite Multimedia and Connected Vehicles working on the mmWave radar signal processing and machine learning based sensor fusion solutions for autonomous driving.

Hui Liu (Fellow, IEEE) received the B.S. degree in 1988 in electrical enginnering from Fudan University, Shanghai, China, and the Ph.D. degree in electrical engineering from the University of Texas at Austin, Austin, TX, USA, in 1995. He is currently the President and the Chief Technology Officer with Silkwave Holdings and an Affiliate Professor with the University of Washington, Seattle, WA, USA. He was a Full Professor/Associate Chair with the Department of Electrical Engineering, University of Washington and a Chair Professor/Associate Dean with the School of Electronic, Information & Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. He was one of the Principal Designers of the 3G TD-SCDMA mobile technologies, and the Founder of Adaptix which pioneered the development of OFDMA-based mobile broadband networks (mobile WiMAX and 4G LTE). He has authored or coauthored more than 90 journal articles, two textbooks, and more than 80 awarded patents. His research interests include broadband wireless networks, satellite communications, digital broadcasting, machine learning, and autonomous driving. He made contributions to global standards for broadband cellular and mobile broadcasting. He was the General Chairman for the 2005 Asilomar Conference on Signals, Systems, and Computers and the 2014 IEEE/CIC International Conference on Communications in China (ICCC14). He was the recipient of the 1997 National Science Foundation (NSF) CAREER Award, the Gold Prize Patent Award in China, three IEEE Best Conference Paper Awards, and the 2000 Office of Naval Research (ONR) Young Investigator Award.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 05,2022 at 06:14:03 UTC from IEEE Xplore. Restrictions apply.