Robust Small Object Detection on the Water Surface through Fusion of Camera and Millimeter Wave Radar

Yuwei Cheng[1,2], Hu Xu[2,3], Yimin Liu[1]
[1]Tsinghua University, [2]ORCA-Uboat, [3]Northwestern Polytechnical University
chengyw18@mails.tsinghua.edu.cn, xuhu@mail.nwpu.edu.cn, yiminliu@tsinghua.edu.cn

Abstract

In recent years, unmanned surface vehicles (USVs) have been experiencing growth in various applications. With the expansion of USVs' application scenes from the typical marine areas to inland waters, new challenges arise for the object detection task, which is an essential part of the perception system of USVs. In our work, we focus on a relatively unexplored task for USVs in inland waters: small object detection on water surfaces, which is of vital importance for safe autonomous navigation and USVs' certain missions such as floating waste cleaning. Considering the limitations of vision-based object detection, we propose a novel radar-vision fusion based method for robust small object detection on water surfaces. By using a novel representation format of millimeter wave radar point clouds and applying a deep-level multi-scale fusion of RGB images and radar data, the proposed method can efficiently utilize the characteristics of radar data and improve the accuracy and robustness for small object detection on water surfaces. We test the method on the real-world floating bottle dataset that we collected and released. The result shows that, our method improves the average detection accuracy significantly compared to the vision-based methods and achieves state-of-the-art performance. Besides, the proposed method performs robustly when single sensor degrades.

1. Introduction

In recent years, unmanned surface vehicles (USVs) have attracted increasing attention and are gradually used for various autonomous activities on water surfaces such as oceanographic research [7], transportation [41], water quality monitoring [19], floating waste removal [32, 36, 1], etc.

Similar to autonomous vehicles on the road, to enable safe navigation and efficient autonomous operation, accurate and robust environmental perception is of vital importance for USVs. Small object detection on water surfaces is an important task for USVs environmental perception. It can be applied to USVs for avoiding small obstacles like buoys and reefs, and plays an important role in USVs' certain missions such as autonomous floating waste detection and cleaning. Vision can provide rich semantic information and is widely used for object detection of USVs. However, unlike autonomous vehicles on the road, there are three main challenges for vision-based small objects detection on water surfaces:

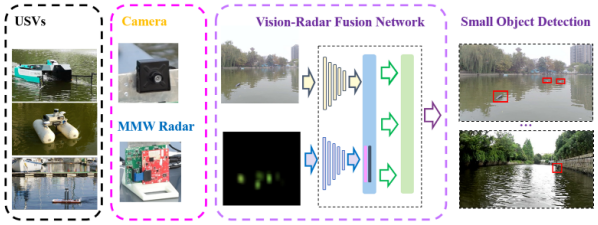


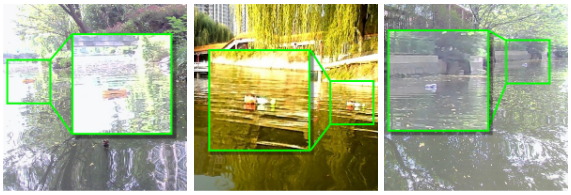
Figure 1. Overview of the proposed method for small object detection for USVs. Our method utilizes the fusion of RGB images and MMW radar data for small object detection of USVs, which can be applied for USVs' certain missions and safe navigation.

- Light reflection on the water surface. As shown in Figure 2(a), the strong light reflection on the water surface can cause high illumination and overexposed image. The small objects like the floating bottles can be shaded by water halos or fused with the background due to overexposure.

- Surrounding scene reflection interference. As shown in Figure 2(b), in some cases such like the small object detection in inland waters, the reflection of the constructions and vegetation on banks increase the complexity of separating the target from the background.

- A short detection range. A long detection range can significantly improve the safety of navigation and the working efficiency of USVs. However, as the size of the target is small, when the target is far from the camera, the number of occupied pixels of the target in RGB images becomes much less as shown in Figure 2(c).

With the increasing demands towards environmental perception for autonomous vehicles, in addition to the vision-based system, object detection based on other sensors like



(a) (b) (c)

Figure 2. Challenges in detecting small objects: Strong light reflection interference. Surrounding scene reflection interference. Small size and a short detection range.

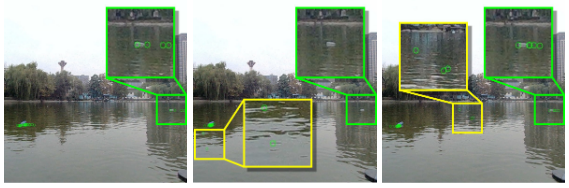


Figure 3. The figures show 3 successive frames of radar point clouds projected onto images. The point clouds of the targets and the clutter points are framed by green boxes and yellow boxes respectively. As can be seen, the point clouds of the small size floating bottles are unstable and hard for humans to identify. Besides, the water clutter can disturb the detection system.

millimeter wave (MMW) radar has shown great value in autonomous driving [49]. Compared to the vision-based system, MMW radar is more robust to lighting conditions and provides the possibility of seeing a long distance [24]. Despite this, for the real-world applications of small object detection on water surfaces based on MMW radar, as shown in Figure 3, difficulties remain to be overcome:

- Weak echoes from non-metallic targets. The radar cross-section (RCS) indicates how detectable a target is by radar. Usually, a target of large size and made of metal material has a larger RCS and are more detectable. The non-metallic small object has a lower RCS and its radar reflection is weak, which significantly increases the difficulties in detection.

- Interference caused by the water surface clutter. To detect the floating bottles on the water surface, the radar is usually equipped at a relatively low height. A lower equipment height makes the radar more easily affected by the water wave and causes falsely detected targets.

- Lack of semantic information. Compared to RGB images, radar provides very little semantic information. Therefore, it is challenging to classify the targets using radar data.

It can be seen that, for small object detection on water surfaces, the performance achieved through a single sensor has bottlenecks. Recently, the nuScenes dataset [4] for object detection and tracking in autonomous driving has been published. The dataset contains images and MMW radar point clouds data, and significantly promotes the researches on deep-level radar-vision fusion based object detection in autonomous driving. However, for small object detection on water surfaces, the characteristics of vision and radar data have changed, which poses new problems. To our knowledge, object detection based on deep-level fusion of images and radar in scenes of water surfaces is a relatively unexplored area. To increase the robustness of object detection on water surfaces and fully utilize the MMW radar point clouds data, in our work, we explore using radar data effectively and propose a novel method, which is based on the deep fusion of radar point clouds and RGB images for robust small object detection on water surfaces. Evaluating on the dataset collected in the real-world scene, our model achieves 90.05% average detection accuracy and outperforms the YOLOv4 [2] baseline (78.46% average accuracy) significantly. In addition, the result of robustness evaluation shows that our model still keeps a good performance when a single sensor degrades.

To summarize, this paper mainly contributes to the following aspects:

1. A first-of-its-kind radar-vision fusion based method that can be applied to small object detection for USVs. Compared to conventional methods, our method can significantly improve the detection performance.

2. A novel approach for the deep-level fusion of MMW radar point clouds and RGB images. By putting forward a novel representation format of radar point clouds and a model that combines different attention mechanisms, the proposed method achieves state-of-the-art accuracy and shows good robustness on detecting small object on water surfaces.

3. A real-time object detection system for USVs with extensive evaluations on the real-world dataset of floating bottles. Besides, we release our code as well as a radar-vision dataset for small object detection on water surfaces to benefit the multi-modal fusion object detection research community.

2. Related Work

2.1. Object Detection for USVs

Vision-based methods are commonly used for object detection on sea surfaces for marine USVs [50]. The public Singapore Marine Dataset [25] and the benchmark [21] built on it have especially supported researches for vision-based maritime object detection [34, 15, 29]. Besides, methods based on the fusion of images and Lidar data are proposed to increase the accuracy and robustness for object detection of marine USVs [35, 44]. For marine USVs, large objects like ferries and ships are the most common targets for detection.

Recently, USVs in inland waters have gained more attention due to its potential application value, for example, the Roboat project [41, 42] which aims at autonomous transportation in urban waterways using USVs. The narrow inland water environments raise higher requirements and pose new challenges for object detection of USVs. The surface reflection, high illumination, and wave interference make it more difficult to detect small objects like small stones, fountain devices and floating bottles that usually may appear in inland waters. For USVs' safe navigation in inland waters, *Hammedi et al.* [11] evaluated common vision-based algorithms on their inland object detection dataset which contains categories of the riverside, vessel, etc. However, no specific small objects are concluded in their dataset. To our knowledge, small object detection for USVs is still a relatively unexplored area.

2.2. Radar-Vision Fusion based Object Detection

In high-level autonomous driving, to improve detection accuracy, robustness and real-time performance, methods based on the fusion of sensors have been widely used for object detection. While the vision system provides abundant semantic information but can be easily affected by adverse conditions, MMW radar can provide location and velocity information of the target robustly under harsh weather conditions. Therefore, the fusion of vision and radar is widely used for object detection in autonomous driving. Early radar-vision fusion is mainly based on object-level fusion. The object-level outputs from the independent radar and image detection pipeline are fused by data association methods such as nearest-neighbor algorithm (NN) and joint probabilistic data association (JPDA). *Wang et al.* [43] achieved on-road vehicle detection and tracking by identifying the vehicle inside the region of interest (ROI) of the monocular image provided by radar detection. Object-level fusion loosely couples vision and radar information. In this case, the robustness of the detection system can be ensured as when one sensor fails, the other one can still work. However, the object-level fusion can bring information loss and cannot make full use of the information from two sensors.

With the development of deep learning, increasing attention has been paid to the deep level radar-vision fusion (data-level and feature-level). Radar point clouds are the final output of the typical MMW radar signal processing pipeline as well as a kind of data that is easy to obtain. Therefore, for deep level radar-vision fusion, most works are based on radar point clouds. Recently, some works [23, 16, 13, 45, 5, 22, 20, 6] explore using feature-level fusion of images and radar for object detection in autonomous vehicle. For feature-level fusion, it is essential to extract features from the irregular and sparse MMW radar point clouds. [20] transformed radar point clouds into BEV images and used CNN for feature extraction. [23, 16, 13, 45, 5, 6] projected radar point clouds to RGB images plane as radar sparse images and then extract features. For the fusion of RGB images and radar data, [23, 13, 5, 20] directly fused the features extracted from the two modalities. [16, 45, 6, 22] improved the performance of fusion by introducing the attention mechanism. Due to the lack of semantic information in radar data, for object detection based on deep level radar-vision fusion, radar data are usually used as the supplementary information to images. However, for small object detection on water surfaces, the robustness of vision information decreases a lot. Therefore, it is worth digging into making full use of the robustness of MMW radar data and better utilize information provided by radar data to improve the performance of object detection based on sensor fusion.

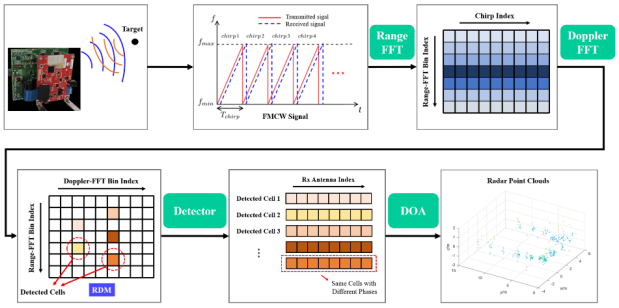


Figure 4. The FMCW radar signal processing chain.

3. Our Approach

3.1. MMW Radar Pipeline

Radar Point Clouds Generation. The MMW radar system transmits frequency-modulated continuous wave (FMCW) and captures the reflected wave. As shown in Figure 4, the sampled beat signals are first transferred to range-Doppler matrix (RDM) via range FFT and Doppler FFT. Then, in the detector processing block, the cells with stronger energy in the RDM are detected. The most common detector in the conventional FMCW signal processing chain is constant false alarm rate (CFAR) detector, which determines the detection threshold according to the surrounding noise level and a scaling factor called a threshold factor. Finally, for each detected cell, direction of arrival (DOA) estimation is performed by utilizing the echo signals of multiple Rx antennas. Thus, we obtain the so-called point clouds that are composed of a number of detected objects with different positions. The radar point cloud can be represented as a set of points and each point can be represented as (x, y, z, v, p) where x,y,z denote the XYZ coordinate data of radar point clouds, v denotes the Doppler velocity, and p denotes the energy of the point.

Radar Point Cloud Projection. The RGB image is a 2-dimension (2D) vertical plane while the radar data is

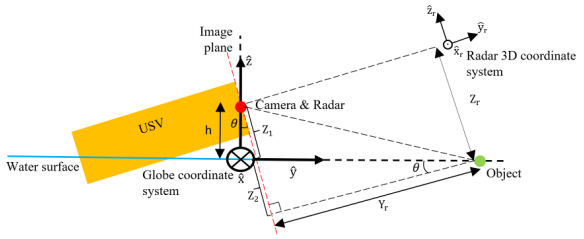


Figure 5. The position compensation projection method of calculating radar distance in z_r Z-axis with the camera height h, camera pitch angle θ, and radar Y-axis distance y.

(a) (b) (c)

Figure 6. Results of different projection methods. The radar point clouds (the green points) are projected onto the images. (a) shows the results of direct perspective projection method. (b) shows the results of fixing height projection method. (c) shows the results of our position compensation projection method. It can be seen that, our method performs better in projection accuracy than the other two methods.

situated in a 3-dimension (3D) coordinate system. In order to eliminate the differences of data formats between two modalities and simplify the fusion learning process [5], we transform the radar point clouds in 3D coordinate into image-plane data in 2D coordinate through projection. However, there are two main challenges for radar point cloud projection of USVs on water surfaces. Firstly, unlike Lidar point clouds, the MMW radar point cloud is inaccurate in Z coordinate [16]. Besides, unlike on the road, the camera's view angle always changes when the USV sails on the water. Therefore, inspired by the fixed height perspective projection method [5], we propose a new position compensation projection method to tackle the problems. As the changes of the height of camera causes relatively little influence on the Z coordinate of point clouds compared to the changes of camera's view angle, we assume that the height of camera is approximately unchanged. As shown in 5, given a fixed height of camera and the pitch angle from IMU, we can compute a new value z_r of the point in Z coordinate using

$$z_r = z_1 + z_2 = \frac{h}{\cos\theta} + y * \tan\theta, \tag{1}$$

where h is the height of camera, θ is the pitch angle. The projection result is shown in Figure 6. Compared to the perspective projection method and the fixed height projection method, our position compensation projection method performs better on water surfaces.

Radar Point Density Map. Radar point projections are usually transformed into binary radar point map [5]. In order to make better use of radar data, we propose a new radar input format: radar point density map (RPDM) inspired by the ground truth generation method used in crowd density counting task [46]. Each radar point is projected onto the image plane to generate the RPDM. RPDM $\in \mathbb{R}^{3\times H_0\times W_0}$. If there is a radar point projected at pixel $\boldsymbol{u}_i$ in RPDM, we represent it as a delta function $\delta(\boldsymbol{u}-\boldsymbol{u}_i)$. Hence an RPDM with N radar points can be represented as a function

$$\boldsymbol{F}(\boldsymbol{u}) = \sum_{i=1}^{N} \delta(\boldsymbol{u}-\boldsymbol{u}_i) * G_{\sigma_0}(\boldsymbol{u}) \cdot (r_i, v_i, p_i)^T, \tag{2}$$

where G_{σ_0} is the Gaussian kernel with variance σ_0, $r_i = \sqrt{{x_i}^2 + {y_i}^2 + {z_i}^2)}$ denotes the range and v_i, p_i denote the Doppler velocity and energy of the ith radar point respectively.

An illustration of the RPDM is given in Figure 7. Transforming raw radar point clouds onto RPDM, our radar input contains not only spatial distribution information of radar point clouds but also Doppler velocity and energy of radar point clouds. On the other hand, with density distribution characterization, RPDM has more abundant gradient characterization than binary radar point map [16] and can be learned by convolutional neural network (CNN) more effectively.

3.2. Fusion Model Architecture

A strong robust small object detection model is based on the complementary interactions of camera and MMW radar, whose roles are adaptively adjusted or re-weighted according to self/environmental dynamics. Based on the requirement of robust small object detection on the water surface and the characteristics of MMW radar data, we propose a Radar-Image spatiotemporal fusion network (RISFNet) to fuse adjacent frames radar data with single frame RGB image under different scales. Considering the instability of radar towards weak reflection objects, we make use of adjacent frames of radar points as inputs of radar backbone. Besides, inspired by most one-stage detection networks [30], we generate image and radar feature maps of different sizes to fuse them in order that the detection model can detect objects of different sizes.

As seen in Figure 7, the RISFNet model mainly consists of three blocks: backbone, feature fusion block, and feature pyramid networks (FPN) [17]. For the backbone block shown in Figure 7(a), we select two backbones to extract features from images and RPDMs separately. The fusion block (as shown in Figure 7(b)) utilizes temporal position encoding as well as self-attention block to fuse multi-frame radar data and adopts global attention module to fuse multi-scale radar and image features. Finally, the fusion features are fed into FPN prediction block shown in Figure

7(c) to predict detection results under three scales. Next, we will introduce more details about the important modules in RISFNet model.

Backbone. RGB images and RPDMs have different characteristics, and RGB images contain richer information. Therefore, using different backbone networks for radar and image feature extraction can improve the efficiency of the model. Compared with the complex weighty image backbone network, the radar backbone network we select is light and is suitable for extracting features from RPDMs. As shown in Figure 7(a), for the image backbone network, we adopt the same backbone architecture named CSPdarknet53 as used in YOLOv4 [2]. The CSPdarknet53 network extracts image features of three different sizes. For radar feature extraction, we use the light VGG-13 backbone network [33] and the network transforms different frames of radar backbone inputs into radar features with the same size of image features. The input sizes of image and RPDM are both $416 \times 416 \times 3$. The final sizes of extracted image and radar features are $512\times13\times13$, $512\times26\times26$, $256\times52\times52$.

Temporal Position Encoding. As the radar point clouds of small targets in current frame have the characteristics of instability and glitter, and the water clutter has random distribution in different frames, we adopt the temporal position encoding and fuse the past frames of RPDM to enhance RPDM in current frame. However, there are spatial position errors between the past frames of RPDM and the RGB image in current moment. An earlier radar frame has greater errors. Thus, referring to position encoding used in natural language processing tasks [40], we adopt a similar position encoding method to add temporal information of radar data. Then, the feature map of the t_kth frame radar data $\mathbf{F}_{t_k}$ with temporal encoding is computed as:

$$\mathbf{F}_{t_k} = \mathbf{F}_{t_k} \cdot \sin\big((n+k)/n\big), \tag{3}$$

where n is the total number of radar frames, t_k is the temporal order position of radar frames, $k \in [0, -n+1]$.

Self-Attention Block. The concept of self-attention was originally designed for natural language processing and image transformation task [9]. Similar to a self-filtering process that autonomously sieves informative features, self-attention block lets individual sensor branches adapt themselves first and is usually used as a promising way to control the information flow and enable model adaptation [6]. As we all know, radar data contains points of real target and clutter points. The clutter points lead to false object information and can cause errors in detection results. In this case, we need to enhance real target points and weaken clutter points before fuse radar data with RGB images. Besides, self-attention block is also used to learn the radar points' relationship of surroundings With several independent multi-layer perceptron (MLP) blocks, radar feature maps of different frames are separately processed into $\mathbf{F}'_{t_k} \in \mathbb{R}^{1\times H\times W}$, and then all radar feature maps in different frames are merged into a fusion radar feature $\mathbf{F}_{\text{radar}}$ by concatenation operation:

$$\mathbf{F}'_{t_k} = \mathbf{c}\big(\mathbf{F}_{t_k} + MLP_k\left(\mathbf{F}_{t_k}\right)\big) \tag{4}$$

$$\mathbf{F}_{\text{radar}} = \text{cat}\big(\mathbf{F}'_{\text{t}_{-\text{n}+1}}, \mathbf{F}'_{\text{t}_{-\text{n}+2}}, \cdots, \mathbf{F}'_{\text{t}_0}\big), \tag{5}$$

where $\mathbf{F}_{t_k} \in \mathbb{R}^{C\times H\times W}$ denotes the feature map of the t_kth frame radar data, C, H, W denote the channel, height, and width of the feature map respectively (the values of C, H, W are different under different feature scales), and MLP_k is the independent multi-layer perceptron for $\mathbf{F}_{t_k}$, $\mathbf{c} \in \mathbb{R}^{C/n\times1\times1}$ denotes convolution module to reduce channel before merge, cat is concatenation operation.

Global Attention Block. The end goal of attention block is to realize adaption through complementary sensor interactions. Although classical fusion algorithm (e.g., Bayesian filtering or fixed-lag smoothers) can realize such an adaptation by incorporating physical models into the algorithm design, they perform hard in complex nonlinear feature space and require better design. Multilayer global attention network observes all sensor channels and better exploit complementary sensor behaviors, which can improve the robustness of fusion model [47]. Therefore, in contrast to concatenate image feature $\mathbf{F}_{\text{image}} \in \mathbb{R}^{C\times H\times W}$ and radar feature $\mathbf{F}_{\text{radar}} \in \mathbb{R}^{C\times H\times W}$ into a "big" vector directly, we adopt global channel attention block [12] to endow a multimodal fusion object detection model with the ability to adapt to uncertain environment. When the camera or radar fails and the model gets poor sensor data, the global attention block will adjust camera or radar fusion to reduce the decline of the model performance. As shown in Figure 7(e), we use a shared MLP block to generate fusion features $\mathbf{F}_{\text{fusion}}$ from image features $\mathbf{F}_{\text{image}}$ and radar features $\mathbf{F}_{\text{radar}}$. In short, the global channel attention fusion is computed as:

$$\begin{aligned}\mathbf{F}_{\text{fusion}} = \sigma\Big(&\mathbf{W_1}\tau\Big(\mathbf{W_0}\big(\text{MaxPool}\left(\mathbf{F}_{\text{image}}\right)\big)\Big)\\ &+\mathbf{W_1}\tau\Big(\mathbf{W_0}\big(\text{MaxPool}\left(\mathbf{F}_{\text{radar}}\right)\big)\Big)\Big),\end{aligned}$$

where σ denotes the sigmoid function and τ denotes the ReLU function. MLP weights $\mathbf{W_0} \in \mathbb{R}^{C/16\times1\times1}$ and $\mathbf{W_1} \in \mathbb{R}^{C\times1\times1}$ are shared for both image and radar inputs.

4. Experiments

4.1. Dataset

Floating waste cleaning is one of the most popular applications for USVs and plastic wastes like floating bottles are the common targets for cleaning USVs' detection system. The plastic bottles have small size and low RCS, which can

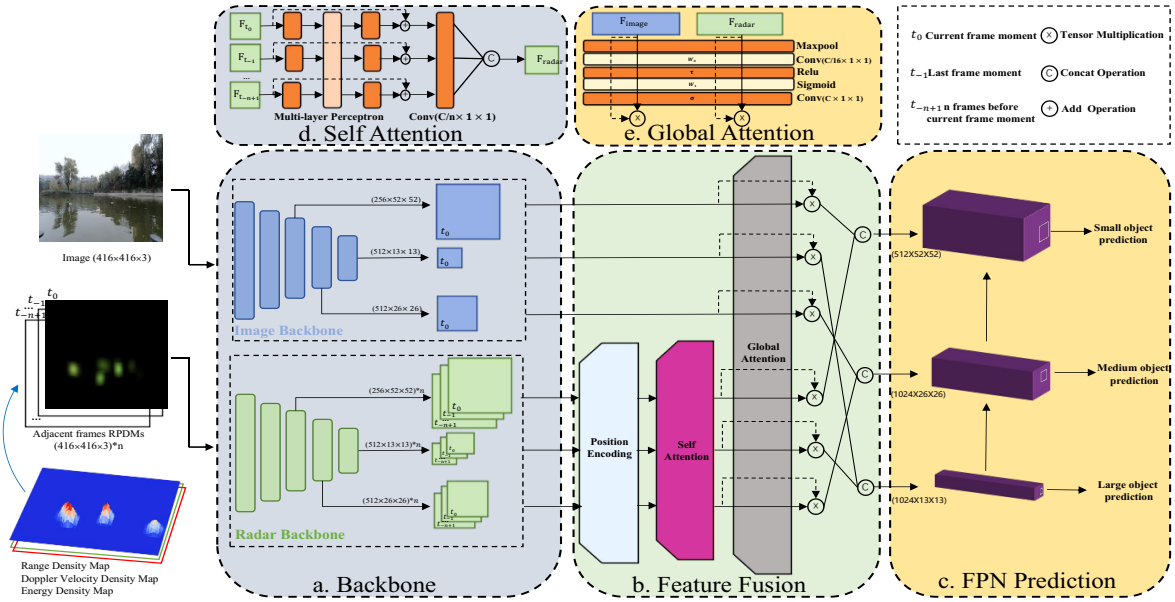


Figure 7. The RISFNet model architecture.

better present the challenges in small object detection on water surfaces. Therefore, we choose the floating bottle as the detection target to test our model.

The dataset we use for training and evaluation are collected in the real-world inland waters including rivers and lakes. One camera collects 1280×720 RGB images at 15 Hz. The IMU collects pose information at 10 Hz. For the MMW radar, we use a Texas Instrument 77 Ghz FMCW radar AWR1843. The maximum range of radar is set as 30 m and the range resolution is 0.04 m. The maximum velocity of radar is set as 4.10 m/s and the velocity resolution is 0.03 m/s. The radar frame rate is also 10 Hz. Data from different sensors are well synchronized by using the recorded timestamps.

We gathered 12000 frames of synchronized images and radar data in total. To avoid diminishing return and overfitting of the model caused by the high-similarity successive frames, we firstly down-sampled the data and finally selected 1895 frames of radar data and RGB images. The data are annotated manually by using the LabelImg tool [39] and are verified repeatedly to ensure the annotation quality. There are 3164 labeled objects in total. According to commonly used definition in Coco dataset [18], the objects that occupy area smaller than $< 32 \times 32$ pixels are regarded as small objects. In our dataset, there are 1946 small objects in total.

4.2. Implementation and Details

The dataset is divided into the training set and the test set according to the ratio of 4:1. During training, multi-scale data augmentation methods such as image resizing, image placing, and image left-right flipping are used for our training images and RPDMs, and we also randomly adjust the hue saturation value of images.

In our experiment, we adopt past three radar frames data to generate radar backbone input RPDM. For RPDM generation, we set the Gaussian kernel size as 101×101 square, the variance σ_0 is 30. In order to keep feature scales consistent between modalities, we scale each modality by its mean and standard deviation calculated over the training set. We use the same loss function as that in YOLOv4 [2], which contains location CIoU loss [48], confidence loss and classification loss.

In the training, we use the model CSPDarknet53 pre-trained from VOC datasets [10] for image backbone. Our implement is based on PyTorch and trained on 4 Nvidia GTX 1070 GPUs with initial learning rate set to be $1e^{-3}$ and batch size set to be 4. The network is trained for 100 epochs using the ADAM optimizer [14] with weight decay of 5×10^{-4} and the mini-batch StepLR descent algorithm with step-size = 1, gamma = 0.9. During testing, the average running speed of our RISFNet model in embedded device Nvidia Jeston TX2 is about 6 frames per second (FPS). As the speed of USVs is much lower than autonomous vehicles, our model can meet the real-time requirement of object detection on water surfaces.

5. Quantitative Evaluation

5.1. Comparison with Single Modality

To verify the improvement in detection accuracy using the fusion of two modalities, we compared our method with

Table 1. Results on the real-world dataset using our method and methods based on single modality.

Modality	Method	AP^{35}	AP^{50}
Image	Faster-RCNN [31]	77.35%	57.58%
	YOLOv4 [2]	78.46%	57.04%
	EfficientDet [37]	78.62%	58.52%
	FCOS [38]	68.71%	58.56%
Radar	Danzer et al.[8]	25.44%	18.81%
	VoteNet [26]	36.98%	20.06%
Image & Radar	**RISFNet (ours)**	**90.05%**	**75.09%**

* AP^{35} and AP^{50} denote the average precision with the IoU threshold at 35% and 50% respectively.

Table 2. Results on the real-world dataset using our method and other radar-vision fusion models.

Method	AP^{35}	AP^{50}
CRF-Net [23]	79.63%	57.74%
Li et al. [16]	85.28%	64.64%
RISFNet (ours)	**90.05%**	**75.09%**

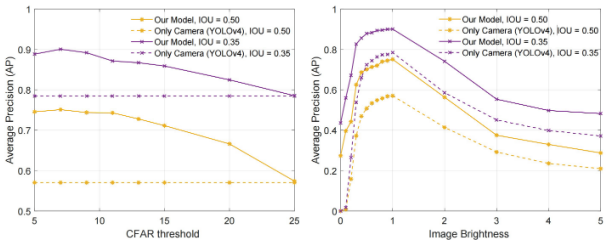
methods using single modality on the real-world dataset. As shown in Table 1, we compare our RISFNet with 4 methods based on RGB images and 2 methods based on radar point clouds. The training and test set used in all of the baseline methods and our method are the same. As for the training settings for the baseline methods, we use the recommended training settings with little optimization. The result shows that, compared to the methods that are based on a single sensor, there is a significant improvement in the performance of small object detection using fusion of vision and radar data.

5.2. Comparison with other fusion models.

To verify the improvement in detecting small objects using our method compared to other radar-vision fusion based methods used in autonomous driving, we test the performance of the methods presented in [23] and [16] on our dataset using the public codes. The result in our dataset is shown in Table 2. Besides, with reference to the recent work, we also test our method on the nuScenes dataset [3]. We compare our method with [16] using the same mini-dataset implement as used in their work. The mean average precision (mAP) of [16] 24.3%, and the mAP of our method is 28.25%. The result indicated that our approach also performs well under the real-world scenes other than in inland waters, such as autonomous driving on road.

(a) threshold=6 (b) threshold=11 (c) threshold=20

Figure 8. The figures show the radar point clouds of a same frame under different CFAR thresholds projected onto the images.



(a) (b)

Figure 9. (a) The figure shows the average precision (*AP*) under different radar CFAR detector thresholds. The *AP* of our fusion model is higher than *AP* of the model that only relies on visual information despite the radar degradation. (b) The figure shows the *AP* of our model and the vision baseline under different image brightness. Our model achieves a higher accuracy and shows better robustness when image degrades.

5.3. Robustness Analysis

The robustness of the method is essential to the deep fusion of multiple sensors. It is expected that when one sensor degrades or even becomes completely unusable, the performance of the fusion model should be better than using a single sensor. Therefore, we test how the model performs under the conditions that radar or image degrades respectively. In our experiment, we still use the model trained on normal dataset.

Radar Degradation. For radar point clouds, parameters of the detector used in radar signal processing pipeline are vital. Usually, the threshold of the detector is adjusted to meet the requirements of different tasks and application scenes. A higher threshold usually leads to sparser point clouds of valid targets. If the threshold is too high, the target will not be detected, which means that the radar data contribute little in sensor fusion. On the contrary, if the threshold is too low, there will be more water clutter points, which lead to interference for the detection system. Therefore, the robustness analysis is carried out by changing the threshold of radar CFAR detector (as shown in Figure 8). The result is shown in Figure 9(a). It can be seen that, when radar data degrades (less valid target points or more clutter points), our model still out-performs the model that only relies on visual information.

Image Degradation. For RGB images, we mainly consider the influence caused by the changes of lighting condi-

Table 3. The detection accuracy of the model using different input formats of radar data.

Radar point cloud representations	AP^{35}	AP^{50}
RPDM (ours)	**90.05%**	**75.09%**
RPDM(only density map)	82.48%	63.93%
RPDM(only range density map)	88.80%	72.20%
RPDM(only velocity density map)	83.67%	64.01%
RPDM(only energy density map)	84.59%	66.85%
Point Clouds (PointNet [27])	87.12%	60.06%
Point Clouds (PointNet++ [28])	87.64%	69.55%
Radar sparse image [5]	87.12%	69.58%
Line shape radar image [16]	85.15%	66.48%

tions under real-world outdoor environments. The result is shown in Figure 9(b). When changing the brightness of input RGB images, the accuracy of the model decreases, but is still higher than the accuracy of the model using a only the camera.

Platform and Environment. For the real-world applications, we evaluate the robustness of the proposed method under two conditions, the increase of USV's speed and the water wave interference. For the increase of USV's speed, during our data acquirement, the max speed of our USV is 2 m/s. We simulate higher speed (4m/s) through downsampling radar data frame rate. The result of our method is 89.98% (AP^{35}). For the water wave interference, we test our model on wave scene data in our dataset separately and the result is 89.22% (AP^{35}).

5.4. Ablation Study

Input Radar Data Format. We evaluate how different input formats of radar point clouds influence the performance of the model. For the feature extraction method for directly using 3D point clouds, we use the PointNet [27] and PointNet++ [28]. It can be seen from Table 3 that the proposed RPDM can better represent the information of radar point clouds.

Model Architecture. First, for the backbone block, we test extracting features from radar data and images separately as well as using one backbone to extract features from concatenated data of two sensors. The result is shown in Table 4. It can be seen that extracting features from radar data and images separately is more effective. Besides, we evaluate the performance of the model using only a single frame of radar data. As shown in Table 4, the temporal position encoding and the self-attention block are effective for enhancing radar data. Finally, we evaluate the performance of the model without introducing the global attention module to test its influence. The result shows that, the global attention module contributes a little to the detection accuracy. However, when we evaluate the model on degraded sensor data, the AP^{35} of the model without global attention module is 87.34% while the AP^{35} of the model with the module is 90.05%, which shows that the global attention module can improve the robustness of the model.

Table 4. Results of ablation study on model architecture.

Ablation ways	AP^{35}	AP^{50}
RISFNet (ours)	**90.05%**	**75.09%**
Use only one backbone	82.81%	63.68%
Use a single frame radar data	88.34%	68.83%
Not use Position encoding	89.72%	72.24%
Not use Self attention	88.72%	71.38%
Not use Global attention	88.95%	70.40%

Figure 10. Detection results of our approach on test dataset: The blue boxes are groundtruth, the green boxes are the detection result with IOU threshold 0.5. Our approach shows good performance on small object detection in inland waters.

Visualization. The visualization of the detection result is shown in Figure 10. As can be seen, our approach shows strong robustness in challenging situations such as: rivers with waves (radar degradation situation) which make radar data contain cluster points from water as well as the bright or dark scenes (image degradation situation) under different weather and lighting conditions.

6. Conclusion

In this paper, we have studied a relatively unexplored task for USVs in inland waters: small object detection. We proposed a novel method for representing the radar point clouds efficiently as well as a new model for object detection based on radar-vision fusion. Our model utilizes a deep-level fusion of RGB images and multi-frame MMW radar data at multi-scale. In the experiment based on the real-world floating bottle detection dataset, our method not only achieves significant improvements in detection accuracy compared to the vision-based object detection methods but also shows good robustness when a single sensor degrades. The proposed method can be applied to autonomous driving and mobile robots for robust radar-vision fusion based object detection. In the future, we plan to further extend the water surface small object dataset we released. Sensors like Lidar will be added to support researches on object detection using fusions of various modalities and to further improve the accuracy and robustness of the small object detection system.

References

[1] Abir Akib, Faiza Tasnim, Disha Biswas, Maeesha Binte Hashem, Kristi Rahman, Arnab Bhattacharjee, and Shaikh Anowarul Fattah. Unmanned floating waste collecting robot. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pages 2645–2650. IEEE, 2019. 1

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2, 5, 6, 7

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 7

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2

[5] Simon Chadwick, Will Maddern, and Paul Newman. Distant vehicle detection using radar and vision. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8311–8317. IEEE, 2019. 3, 4, 8

[6] Shuo Chang, Yifan Zhang, Fan Zhang, Xiaotong Zhao, Sai Huang, Zhiyong Feng, and Zhiqing Wei. Spatial attention fusion for obstacle detection using mmwave radar and vision sensor. *Sensors*, 20(4):956, 2020. 3, 5

[7] Joseph Curcio, John Leonard, and Andrew Patrikalakis. Scout-a low cost autonomous surface platform for research in cooperative autonomy. In *Proceedings of OCEANS 2005 MTS/IEEE*, pages 725–729. IEEE, 2005. 1

[8] Andreas Danzer, Thomas Griebel, Martin Bach, and Klaus Dietmayer. 2d car detection in radar data with pointnets. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 61–66. IEEE, 2019. 7

[9] Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Longyue Wang, Shuming Shi, and Tong Zhang. Dynamic layer aggregation for neural machine translation with routing-by-agreement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 86–93, 2019. 5

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[11] Wided Hammedi, Metzli Ramirez-Martinez, Philippe Brunet, Sidi-Mohamed Senouci, and Mohamed Ayoub Messous. Deep learning-based real-time object detection in inland navigation. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2019. 3

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[13] Vijay John and Seiichi Mita. Rvnet: deep sensor fusion of monocular camera and radar for image-based obstacle detection in challenging environments. In *Pacific-Rim Symposium on Image and Video Technology*, pages 351–364. Springer, 2019. 3

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] Sung-Jun Lee, Myung-Il Roh, Hye-Won Lee, Ji-Sang Ha, Il-Guk Woo, et al. Image-based ship detection and classification for unmanned surface vehicle using real-time object detection neural networks. In *The 28th International Ocean and Polar Engineering Conference*. International Society of Offshore and Polar Engineers, 2018. 2

[16] Liang-qun Li and Yuan-liang Xie. A feature pyramid fusion detection algorithm based on radar and camera sensor. In *2020 15th IEEE International Conference on Signal Processing (ICSP)*, volume 1, pages 366–370. IEEE, 2020. 3, 4, 7, 8

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[19] Dario Madeo, Alessandro Pozzebon, Chiara Mocenni, and Duccio Bertoni. A low-cost unmanned surface vehicle for pervasive water quality monitoring. *IEEE Transactions on Instrumentation and Measurement*, 69(4):1433–1444, 2020. 1

[20] Michael Meyer and Georg Kuschk. Deep learning based 3d object detection for automotive radar and camera. In *2019 16th European Radar Conference (EuRAD)*, pages 133–136. IEEE, 2019. 3

[21] Sebastian Moosbauer, Daniel Konig, Jens Jakel, and Michael Teutsch. A benchmark for deep learning based object detection in maritime environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[22] Ramin Nabati and Hairong Qi. Rrpn: Radar region proposal network for object detection in autonomous vehicles. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3093–3097. IEEE, 2019. 3

[23] Felix Nobis, Maximilian Geisslinger, Markus Weber, Johannes Betz, and Markus Lienkamp. A deep learning-based radar and camera sensor fusion architecture for object detection. In *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pages 1–7. IEEE, 2019. 3, 7

[24] Sujeet Milind Patole, Murat Torlak, Dan Wang, and Murtaza Ali. Automotive radars: A review of signal processing techniques. *IEEE Signal Processing Magazine*, 34(2):22–35, 2017. 2

[25] Dilip K Prasad, Deepu Rajan, Lily Rachmawati, Eshan Rajabally, and Chai Quek. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: a survey. *IEEE Transactions on Intelligent Transportation Systems*, 18(8):1993–2016, 2017. 2

[26] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 7

[27] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 8

[28] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 8

[29] Dalei Qiao, Guangzhong Liu, Jun Zhang, Qiangyong Zhang, Gongxing Wu, and Feng Dong. M3c: Multimodel-and-multicue-based tracking by detection of surrounding vessels in maritime environment for usv. *Electronics*, 8(7):723, 2019. 2

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 4

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 7

[32] Niramon Ruangpayoongsak, Jakkrit Sumroengrit, and Monthian Leanglum. A floating waste scooper robot on water surface. In *2017 17th International Conference on Control, Automation and Systems (ICCAS)*, pages 1543–1548. IEEE, 2017. 1

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[34] Xueqiang Song, Peng Jiang, and He Zhu. Research on unmanned vessel surface object detection based on fusion of ssd and faster-rcnn. In *2019 Chinese Automation Congress (CAC)*, pages 3784–3788. IEEE, 2019. 2

[35] Leo Stanislas and Matthew Dunbabin. Multimodal sensor fusion for robust obstacle detection and classification in the maritime robotx challenge. *IEEE Journal of Oceanic Engineering*, 44(2):343–351, 2018. 2

[36] Jakkrit Sumroengrit and Niramon Ruangpayoongsak. Economic floating waste detectionfor surface cleaning robots. In *MATEC Web of Conferences*, volume 95, page 08001. EDP Sciences, 2017. 1

[37] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 7

[38] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9627–9636, 2019. 7

[39] Tzutalin. Labelimg. Free Software: MIT License, 2015. https://github.com/tzutalin/labelImg. 6

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 5

[41] Wei Wang, Banti Gheneti, Luis A Mateos, Fabio Duarte, Carlo Ratti, and Daniela Rus. Roboat: An autonomous surface vehicle for urban waterways. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6340–6347. IEEE, 2019. 1, 3

[42] Wei Wang, Tixiao Shan, Pietro Leoni, David Fernandez-Gutierrez, Drew Meyers, Carlo Ratti, and Daniela Rus. Roboat ii: A novel autonomous surface vessel for urban environments. *arXiv preprint arXiv:2007.10220*, 2020. 3

[43] Xiao Wang, Linhai Xu, Hongbin Sun, Jingmin Xin, and Nanning Zheng. On-road vehicle detection and tracking using mmw radar and monovision fusion. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):2075–2084, 2016. 3

[44] Yingying Wu, Huacheng Qin, Tao Liu, Hao Liu, and Zhiqiang Wei. A 3d object detection based on multi-modality sensors of usv. *Applied Sciences*, 9(3):535, 2019. 2

[45] Ritu Yadav, Axel Vierling, and Karsten Berns. Radar+ rgb fusion for robust object detection in autonomous vehicle. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1986–1990. IEEE, 2020. 3

[46] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 4

[47] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3186–3195, 2020. 5

[48] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12993–13000, 2020. 6

[49] Taohua Zhou, Mengmeng Yang, Kun Jiang, Henry Wong, and Diange Yang. Mmw radar-based technologies in autonomous driving: A review. *Sensors*, 20(24):7283, 2020. 2

[50] Zhiguo Zhou, Siyu Yu, and Kaiyuan Liu. A real-time algorithm for visual detection of high-speed unmanned surface vehicle based on deep learning. In *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, pages 1–5. IEEE, 2019. 2