Journal of
Marine Science
and Engineering

Article

A Millimeter-Wave Radar-Aided Vision Detection Method for Water Surface Small Object Detection

Jiannan Zhu [1,2], Yixin Yang [1,*] and Yuwei Cheng [2,3]

[1] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710068, China; jacknyzhu@orca-tech.com.cn
[2] ORCA-Uboat, Xi'an 710004, China
[3] Department of Electronic Engineering, Tsinghua University, Beijing 100080, China; chengyw@orca-tech.com.cn
* Correspondence: yxyang@nwpu.edu.cn

Abstract: Unmanned surface vehicles (USVs) have wide applications in marine inspection and monitoring, terrain mapping, and water surface cleaning. Accurate and robust environment perception ability is essential for achieving autonomy in USVs. Small object detection on water surfaces is an important environment perception task, typically achieved by visual detection using cameras. However, existing vision-based small object detection methods suffer from performance degradation in complex water surface environments. Therefore, in this paper, we propose a millimeter-wave (mmWave) radar-aided vision detection method that enables automatic data association and fusion between mmWave radar point clouds and images. Through testing on real-world data, the proposed method demonstrates significant performance improvement over vision-based object detection methods without introducing more computational costs, making it suitable for real-time application on USVs. Furthermore, the image–radar data association model in the proposed method can serve as a plug-and-play module for other object detection methods.

Keywords: unmanned surface vehicle; object detection; visual–radar fusion

check for updates

Citation: Zhu, J.; Yang, Y.; Cheng, Y. A Millimeter-Wave Radar-Aided Vision Detection Method for Water Surface Small Object Detection. *J. Mar. Sci. Eng.* **2023**, *11*, 1794. https://doi.org/10.3390/jmse11091794

Academic Editor: Mohamed Benbouzid

Received: 23 August 2023
Accepted: 27 August 2023
Published: 14 September 2023

1. Introduction

In recent years, unmanned surface vehicles (USVs) have been gradually used in various fields, such as autonomous surface transportation [1], water quality testing [2], autonomous surface cleaning [3], etc. To ensure that USVs complete their tasks safely and intelligently, an excellent and robust perception system is essential. Among all the perception tasks, object detection plays an important role in both safe navigation and special task completion, and small object detection causes the most challenges, for example, the small reefs and other small obstacles that may affect USVs or small floating wastes that a cleaning USV needs to collect.

Recent development in computer vision makes vision-based object detection one of the most cost-effective solutions for the detection system of USVs. However, for vision-based small object detection on water surfaces, many can be missed and falsely detected due to the water surface environments. For vision-based small object detection on water surfaces, on the one hand, as the sky and water surfaces occupy the most area of the image, the reflection of sunlight may cause overexposure. The small objects can be shaded by the halo or fused with the background, which can cause miss detection. Besides, the reflection of objects in the surrounding environments also disturbs the detection system and causes false detection. In addition to the camera, LiDAR is also widely used for object detection as it can provide precise location and shape information of the objects [4]. However, for the small object detection on water surfaces, for LiDAR with a low number of beams, the possibility of LiDAR beams falling on small objects is low and the objects might be unstable in sequential frames. In addition, dense fog is easy to appear on the water surface, which can disturb the propagation of LiDAR and lead to more clutter points [5].

With the development of integrated circuits, the low-cost single-chip 77 GHz millimeter-wave (mmWave) radar is gradually used in autonomous vehicles and mobile robots recently. The mmWave radar can provide measurements of the range, azimuth, and Doppler velocity of the objects. Besides, benefiting from the inherent propagation characteristics of 77 GHz electromagnetic wave, the mmWave radar shows better robustness to harsh weather conditions and lighting conditions compared to camera and LiDAR [6] and can be used during all types of weather and all day. Despite this, there are still some challenges in using mmWave radar for small object detection on water surfaces. The angular resolution of mmWave radar point clouds is relatively low and the points of the objects are usually more sparse [7]. Furthermore, the semantic information of mmWave radar point clouds is often insufficient, making it difficult to accurately discern the types of targets.

Therefore, for small object detection on water surfaces, vision and mmWave radar data complement each other effectively, and fusion of vision and radar can improve the detection performance. Compared to other levels of fusion, decision-level fusion has greater robustness and adaptability, and the fused results are also more interpretable. However, there are two challenges in the decision-level fusion of camera and radar in USVs scenes:

- Extrinsic Calibration. To perform decision-level fusion, the spatial relationship between the mmWave radar and camera needs to be found, which is referred to as extrinsic calibration. Due to the characteristics of glittery and sparsity of mmWave radar point clouds, extrinsic calibration between mmWave radar and cameras typically requires specific markers, and the calibration process is usually complex. Current extrinsic calibration is mainly conducted offline with human assistance. However, the positions of sensors on the platform may change due to vibrations, shocks, or structural deformations of USVs, leading to some degree of variation in the extrinsic parameter between the mmWave radar and the camera.
- Data association. Traditional methods tend to manually craft various distance metrics to represent the similarities between vision and mmWave radar data. However, these manually crafted metrics are not adaptable when the data from different sensors degrade, and setting the parameters is also challenging.

In this paper, we propose a water surfaces small object detection method based on the decision-level fusion of vision and mmWave radar data. Compared to traditional methods, the proposed method has the following advantages: (1) With an initial offline calibrated extrinsic parameter, the proposed method is adapted to changes in extrinsic parameters to some degree during USVs' online operation; (2) The method has lower computational complexity and can run in real time on embedded systems; (3) The method achieves a higher detection accuracy in the water surface small object detection task.

The contribution of this paper mainly lies in the following aspects:

- We propose a new mmWave radar-aided visual small object detection method.
- We propose a new image–radar association model based on the metric learning model, which can achieve a robust association of mmWave radar data and images with inaccurate extrinsic parameters to some degree.
- We test the proposed method on real-world data, and the results show that our method achieves significantly better performance than current vision detection methods.

The detailed composition of this paper is listed as follows. In Section 2, we discuss the related works, including object detection on water surfaces and the visual–radar fusion-based detection method. In Section 3, we introduce the proposed mmWave radar-aided visual small object detection method in detail. Section 4 gives the results of experiments based on real-world data. Finally, Section 5 concludes this paper.

2. Related Works

2.1. Object Detection on Water Surfaces

Attention from researchers has been paid to object detection on water surfaces. Hammedi et al. [8] proposed a relevant dataset for inland water navigation that contains categories of riverside, vessel, person, etc. Moosbauer et al. [9] proposed a benchmark for object detection in maritime environments based on the Singapore Marine Dataset [10] to support relevant research. Vision-based detection methods are the ones that are mainly used for water surface object detection. For example, the method proposed in [11] is based on MobileNet for feature extraction and SSD for fast multi-scale detection to achieve real-time marine object detection of high-speed USVs. Zhang et al. [12] proposed a method for marine object detection and tracking based on improved YOLOv3 and used their method on a real USV experiment platform. The authors of [13] fused DenseNet in YOLOv3 for robust detection of marine objects under various weather conditions.

The vision-based methods for object detection on water surfaces are easily disturbed by weather and lighting conditions. Besides, the methods mainly aim at detection and cannot provide relative location information of the object. Therefore, methods based on the fusion of LiDAR data and images are proposed to improve detection accuracy and support object localization. Wu et al. [14] proposed a 3D object detection method based on the fusion of image and LiDAR point cloud for USVs in marine environments. They used a two-stage network which contains the proposal generation network and the deep fusion detection network. Cardillo et al. [15] analyzed the detection performance of radars with different frequency bands for USVs obstacle avoidance tasks, providing a valuable reference for the perception applications of mmWave radar in USVs. Im et al. [16] conducted object detection and tracking in USVs using frequency-modulated continuous wave (FMCW) radar with improved density-based spatial clustering of applications with noise (DBSCAN). Ha et al. [17] achieved autonomous obstacle avoidance tasks of USVs based the on marine radar. Stanislas et al. [18] utilized the fusion of LiDAR point clouds, camera, and 2D sparse radar point clouds for robust detection and classification in marine environments. The fusion-based methods can provide location information of the object in addition to object detection.

Current water surface object detection research mainly aims at maritime object detection. The objects are mostly vessels and other objects which are relatively big. However, for USVs, there are many other small objects that may cause dangers, such as small fountain nozzles, or are the searching targets of USVs, such as floating wastes. Besides, the Lidar that can be applied to complex water body environments is relatively expensive.

2.2. Visual–Radar Fusion Detection

Using solely visual information for object detection is susceptible to the influence of factors such as weather conditions, lighting, and object motion, which can result in detection errors and unreliability. In contrast, mmWave radar offers robust localization and velocity information for objects even in adverse weather conditions. Consequently, the fusion of visual and radar modalities, known as camera–radar fusion detection, has garnered increasing attention in the field of computer vision in recent years. Various fusion methods have been proposed to combine the strengths of camera and radar modalities and achieve improved detection performance in diverse scenarios. Based on the fusion stage within the network, the fusion methods of camera and radar can be broadly categorized as data-level fusion, feature-level fusion, and decision-level fusion. Data-level fusion [19–21] integrates raw or preprocessed data from radar and camera sensors at the early stages of deep learning models. Such methods necessitate addressing the correspondence between the camera and mmWave radar data, often requiring object matching or association operations. Long et al. [19] introduced Radar-Camera Pixel Depth Association (RC-PDA), which enhances and densifies radar images by associating radar point clouds with nearby image pixels. This approach resolves the challenge of associating radar point clouds with image pixels. Nobis et al. [20] input cascaded camera and radar point clouds into a network and extract features from the combined data using VGG [22]. However, data-level fusion methods typically impose high computational complexity and real-time requirements due to the potential disparate update rates between the camera and mmWave radar data.

Feature-level fusion [23–27] combines features extracted from radar data and camera images at the intermediate stages of deep learning-based fusion networks. Leveraging the distinct characteristics and advantages of these two sensor types, fusing their features provides a more comprehensive description of target objects. Chadwick et al. [24] proposed generating image and radar features separately using ResNet [28] and subsequently fusing them through concatenation and addition operations. Li et al. [25] introduced a feature pyramid layer attention module that integrates radar information, extending the feature pyramid module through the input interface of radar-projected images and attention modules. Nevertheless, feature-level fusion methods face challenges in striking a balance between fusion and aligning different sensor features.

Decision-level fusion [29] entails conducting separate object detection using camera and radar, followed by combining their results through weighted averaging or voting to obtain a comprehensive outcome. By amalgamating detection results from multiple sensors, the reliability of object detection experiences significant improvement. Jha et al. [29] employed YOLOv3 [30] as the image detector, projecting radar-detection results onto the image plane using transformation matrices, and subsequently aligning independently detected objects from the two sensors. Compared to the first two fusion methods, decision-level fusion exhibits greater robustness and adaptability, facilitating adaptive adjustments based on real-world scenarios and requirements. However, decision-level fusion encounters challenges associated with data inconsistency.

The existing methods primarily focus on road scenes, where visual information plays a dominant role and radar information serves as a supplementary source. However, the water surface environment is considerably more complex, characterized by water reflections and a prevalence of small objects. Relying predominantly on visual information in such scenarios can lead to a higher rate of false detections. Currently, there are limited camera–radar fusion methods specifically designed for water surface detection. Only RISFNet [23] has been proposed, which maps radar point clouds onto the image plane. It incorporates global attention and self-attention mechanisms to achieve deep multi-scale feature fusion between the two sensors, demonstrating robustness in detecting small objects on the water surface. Nevertheless, feature-level fusion alone fails to address the issue of unreliable camera sensors, and RISFNet heavily relies on accurate extrinsic parameters between radar and camera.

3. Our Method

For the task of small object detection on water surfaces, vision-based detection methods always generate false detection due to the sunlight reflection and surrounding scene reflection. The mmWave radar is robust to different lighting conditions but contains limited semantic information compared to the RGB image, which makes it difficult to distinguish objects of similar sizes using the radar-based detection method. Besides, the radar-based detection method may generate false detection on water surfaces due to the water clutter. Therefore, to improve the accuracy and robustness of small object detection on water surfaces, we propose a radar-aided visual small object detection method on water surfaces.

3.1. Network Overview

Due to the inherent shortcomings of camera and radar sensors, in the water surface small object detection task, both vision-based and radar-based detection methods have false detection. However, the reasons that the two sensors generate false detection are different, and the statistical probabilities of error occurrence in detection methods based on the two sensors are also independent. Hence, we adopt a detection method based on the decision-level fusion of vision and radar data. The visual object detection results are gained first, and then the detection results are associated with radar data to reduce false detection.

However, for the decision-level fusion method, the spatial position correlation of different sensors is of vital importance and acquires accurate extrinsic parameters. Due to the sparse and glittery characteristics of mmWave radar point clouds, corner reflectors or LiDAR are usually needed as the auxiliary in the extrinsic calibration between radar and camera, which involves complex calibration procedures [31]. For the applications of USVs, there can be certain variations in the extrinsic parameters between the radar and camera due to vibrations, shocks, or structural deformations of USVs during operations. In this case, we propose a new image–radar association model based on the metric learning model. By training the model using data based on the provided initial extrinsic parameters, the model is adaptable to variations in extrinsic parameters in practical application.

As shown in Figure 1, there are two main stages in the proposed radar-aided visual small object detection method: the detection stage and the association stage. Next, we will introduce more details about the two stages.

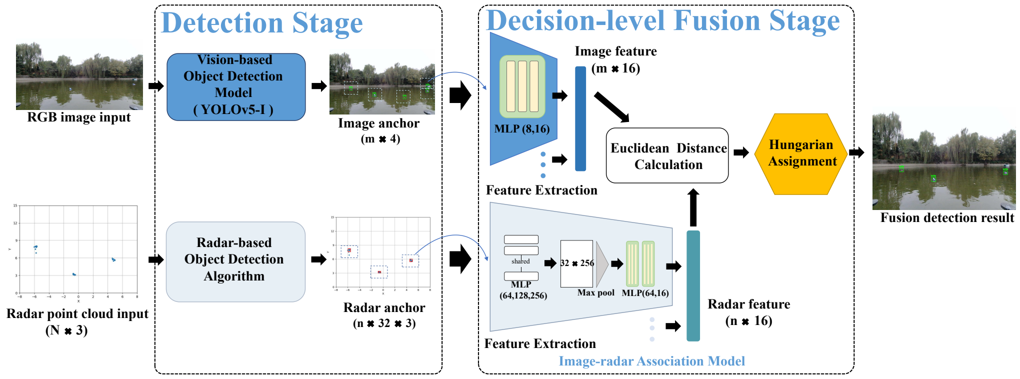


Figure 1. The architecture of the proposed method.

3.2. Detection Stage

The detection stage includes a vision-based detection model and a radar-based detection algorithm. We adopt YOLOv5-l [32] as the vision-based model. YOLOv5-l shows good performance in visual object detection tasks and it is a lightweight model which can carry out real-time inference in an embedded system.

3.2.1. Vision-Based Detection

To make the object detection model specialize on our fusion algorithm, we modify the original YOLOv5-l [32] as the vision-based model. As our fusion detection algorithm can remove the false positive detection results efficiently through the radar-detection results and vision detection results, we need to generate more detection results to improve the recall rate of the vision-based model. The framework of enhanced YOLOv5-l is illustrated in Figure 2. We adjust the prediction head of YOLOv5-l using a double prediction head and transformer decoder module, then we will introduce the architecture of the prediction head in detail.

(1) Double prediction heads. YOLOv5 object detector uses a single prediction head to predict the location and classification of the detected bounding box at the same time. In our vision-based model, we design a double prediction head including a classification head and location regression head to predict, respectively, the location and classification of objects. Independent double prediction heads will benefit from searching both the location and classification of objects. While we utilize the full connection (FC) layer to obtain more semantic information about objects in the classification head, we obtain the position of detection objects in the location regression head.

(2) Transformer decoder module. Inspired by the vision transformer [33], we use a transformer decoder module to replace the convolution blocks in the prediction head. Compared with convolution operation, the transformer decoder module can capture global information and abundant contextual information. Each transformer decoders contain a

multi-head attention layer and a fully-connected layer. Furthermore, there are residual connections between each sublayer. As the prediction head is at the end of the network and the feature map has low resolution, applying a transformer decoder module in a low-resolution feature map explores the feature representation potential with a self-attention mechanism and enlarges the receptive field of the prediction head with low computation and memory cost.

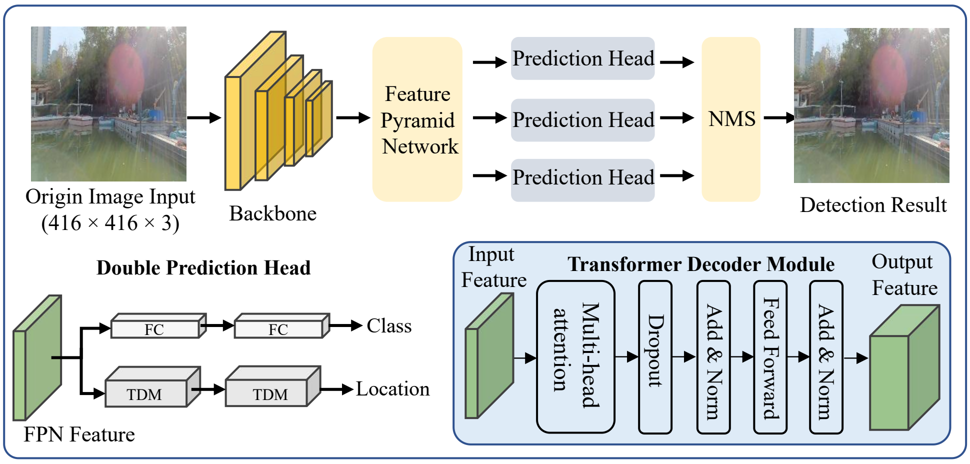


Figure 2. The framework of enhanced YOLOv5-l.

After applying the vision-based detection model to an RGB image, the image anchors, expressed as $\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_m$, where m denotes the number of image anchors, are extracted. Each image anchor $\boldsymbol{B}$ contains four parameters, including the u-axis position, v-axis position, box width, and box height in the u-v image coordinate system. Therefore, for each image, the output size in the detection stage is $m \times 4$.

3.2.2. Radar-Based Detection

A mmWave radar system senses its surroundings by transmitting and receiving FMCW signals. The transmitted and reflected signals are mixed using a frequency mixer to obtain beat signals. Then, 1D (range) Fast Fourier Transformation (FFT) and 2D (velocity) FFT are applied to the sampled beat signals along the fast time and slow time, respectively, resulting in the well-known range–Doppler matrix (RDM). The cells with strong energy in the RDM are detected as targets. The most commonly employed detector for FMCW signal processing is the constant false alarm rate (CFAR) detector, which adaptively estimates the noise level based on nearby cells relative to the cell under test. After detection, the direction of arrival (DOA) is estimated for each detected target using signals from multiple antennas. Consequently, we obtain what is referred to as 4D radar point clouds, representing various detected targets with distinct 3D positions and Doppler velocities. The illustration of the radar signal processing chain is shown in Figure 3.

For radar-based detection, we use the spatial information of mmWave point clouds and the size of the input radar point cloud is $N \times 3$, where N denotes the number of radar points in the current frame and each point contains three coordinates information x,y,z. The radar point clouds are clustered into groups and the discrete radar clutter points are also removed using DBSCAN [34]. The point clouds are divided into n clusters $\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_n$, where n denotes the number of radar point clusters. Then, we use the farthest point sampling (FPS) [35] to sample the point clouds of each group $\boldsymbol{C}_i$ into a fixed number 32. Therefore, the final size of outputs of radar-based detection is $n \times 32 \times 3$.

Through the detection stage, the vision-based and radar-based detection results are gained. Then, the detection results are sent to the fusion association stage to generate fusion detection results.

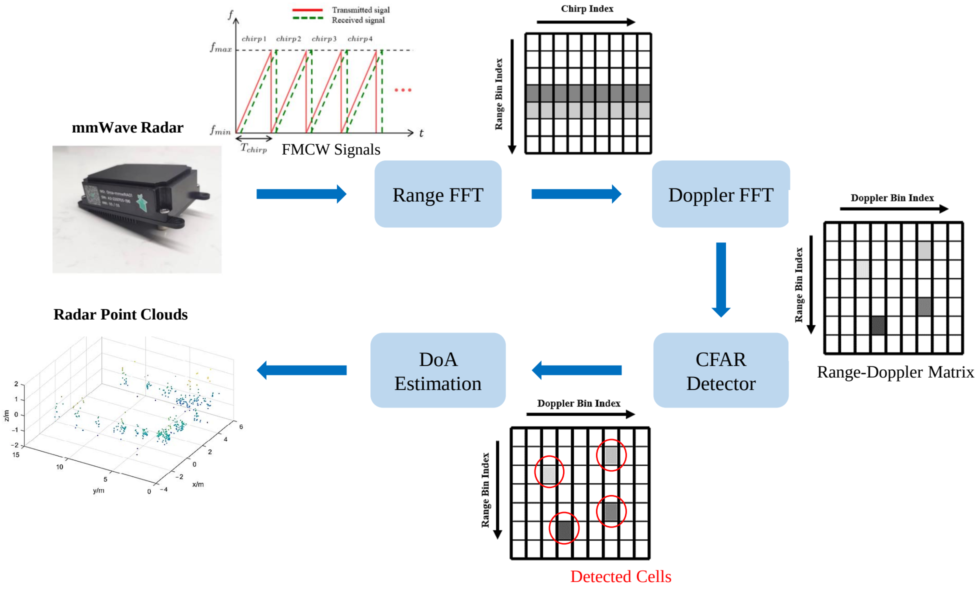


Figure 3. Illustration of the radar signal processing chain.

3.3. Fusion Association Stage

The fusion association stage extracts image feature vectors from object detection bounding boxes in the image plane and extracts radar feature vectors from radar point clouds. The image features can represent the position and size of detection-bounding boxes of corresponding objects in the image plane. The radar feature contains the spatial information of objects in the radar coordinate system as well as the shape information of objects. Therefore, by measuring the similarity between image and radar features, the association of vision and radar-detection results can be achieved. The Hungarian algorithm [36] is used for matching image and radar data according to the L2 distance between two feature vectors. Thus, an end-to-end spatial correlation between image and radar data can be achieved without the extrinsic parameter calibration procedure.

Next, we will introduce the radar and vision feature extraction model in detail. For a frame of RGB image, m bounding boxes are generated from the detection stage and the size of each bounding box is 1×4. We use the multi-layer perception (MLP) to extract the image feature $\boldsymbol{F}_{img}$ whose size is $m \times 16$ from each vision detection result $\boldsymbol{B}_i$:

$$\boldsymbol{F}_{img} = \{\boldsymbol{F}_{img_i} | \boldsymbol{F}_{img_i} = \mathrm{MLP}(\boldsymbol{B}_i), i = 1, 2, ..., m\}, \tag{1}$$

where $\boldsymbol{F}_{img_i}$ is the feature tensor of ith vision detection result $\boldsymbol{B}_i$.

A frame of radar data contains n point cloud groups and the output size of the radar-detection result is $n \times 32 \times 3$, where 32×3 denotes each group consisting of 32 points with each point containing x, y, z coordinates. For radar feature extraction, we adopt the mini-PointNet [37] architecture, which is a famous method to extract point cloud features. Through the shared weighted MLPs, the max-pooling, and another MLP, each point cluster generates a feature of size 1×16. The $n \times 16$ radar feature vector of a whole frame is generated by combining the n cluster features. The radar feature extraction can be represented as follows:

$$\boldsymbol{F}_r = \{\boldsymbol{F}_{r_j} | \boldsymbol{F}_{r_j} = \mathrm{MLP}\big(\mathrm{maxpool}\big(\mathrm{MLP}(\boldsymbol{C}_j)\big)\big), j = 1, 2, ..., n\}, \tag{2}$$

where $\mathrm{F_{r_j}}$ denotes the radar point cloud feature of the ith cluster.

After obtaining a frame of image feature $\boldsymbol{F}_{img}$ and corresponding radar feature $\boldsymbol{F}_r$, we compute the L2 distance between each object's image feature $\boldsymbol{F}_{img_i}$ and each object's radar feature $\boldsymbol{F}_{r_j}$ and obtain a cost matrix of size $m \times n$. Based on the cost matrix, within the minimum distance threshold, the matching results are gained using the Hungarian assignment algorithm.

Through the fusion association stage, the final fusion detection results which contain the vision-based detection box, object classification result, and the range and azimuth of the objects can be gained.

3.4. Loss Function

In our method, the detection model and the image–radar association model are trained separately. The training loss function of vision-based detection model $\mathcal{L}_{\text{vis}}$ is the same as the YOLOv5 object detection model, which is computed as:

$$\mathcal{L}_{\text{vis}} = \alpha_b * \mathcal{L}_{\text{box}} + \alpha_o * \mathcal{L}_{\text{obj}} + \alpha_c * \mathcal{L}_{\text{cla}}, \tag{3}$$

where $\mathcal{L}_{\text{box}}$ denotes the location loss, $\mathcal{L}_{\text{obj}}$ denotes the confidence loss, and $\mathcal{L}_{\text{cla}}$ denotes the classification loss. The three loss weights α_b, α_o, and α_c are constants. For the training of the image–radar association model, we choose the triplet loss [38], which is commonly used as the training loss function in the metric learning. Each training data pair for triplet loss contains three samples: a vision-based detection bounding box $\boldsymbol{B}_i$ as a base anchor, a positive radar sample $\boldsymbol{C}_j$, which is the radar-detection cluster corresponding with $\boldsymbol{B}_i$, and a negative radar sample, which is randomly selected from rest of the radar-detection clusters. The image and radar features are extracted from the training data pair, and the triplet loss is used to minimize the L2 distance d_{pos} between the image feature and positive radar feature while maximizing the L2 distance d_{neg} between the image feature and negative radar feature using:

$$\mathcal{L}_{\text{triplet}} = \max(d_{\text{pos}} + d_{\text{neg}} + \beta, 0), \tag{4}$$

where β is a constant to express the minimum distance loss.

4. Experiment and Evaluation

4.1. Dataset, Evaluation Metric, and Baseline

To evaluate the performance of our method, we use the public FloW-RI dataset [39], which contains synchronized images and mmWave radar data of floating bottles on water surfaces. Besides, to test the model's generalization performance on a broader range of data, we supplement a new dataset for water surface small object detection using a USV platform equipped with an RGB camera and a Texas Instruments 77 GHz single-chip mmWave radar AWR1843. The dataset includes 1600 frames of synchronized RGB images and mmWave radar point cloud data. The newly added data are shown in Figure 4. Finally, we use 4400 frames of data as the training set and 1200 frames of data as the test set.

To quantitatively evaluate the performance of our method, we use the mean of average precision (mAP), which is widely used in object detection as the evaluation metric, and compare the performance of our method with some famous baseline methods in object detection. For vision-based methods, the YOLO [40] series object detection methods are widely used in mobile robots due to the high inference speed. Therefore, we choose the newest YOLOv5-l as one baseline method. Compared to the single-stage object detection methods, the two-stage methods are usually slower but can achieve a higher detection accuracy. Therefore, Fast R-CNN [41] and Cascade R-CNN [42] are also selected as baselines in the experiment. In addition to methods based on the convolutional neural network, in recent years, methods based on transformers also achieve SOTA performance in some tasks. Therefore, we also choose the Swin Transformer [43] as one baseline method. For the mmWave radar-based object detection method, we choose the VoteNet [44] and the method of Danzer et al. [45] as baselines. In addition, we also compare our method with other visual–radar fusion-based methods including feature-level fusion method RISFNet [23], CRF-Net [20], the method in Li et al. [25], and a decision-level fusion method [29].

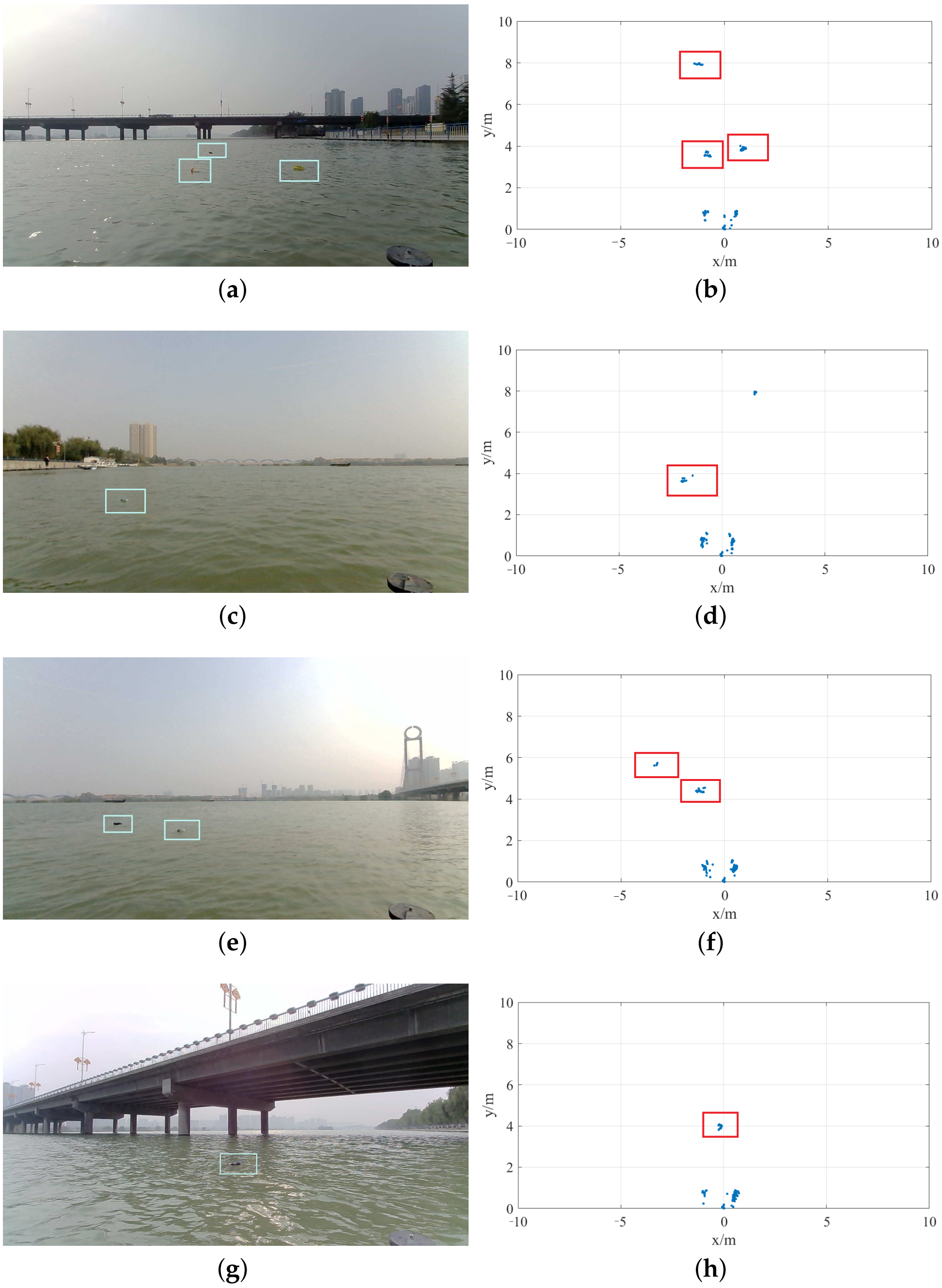


Figure 4. Examples of the supplementary data. The blue and red boxes indicate the targets in the image and radar point cloud respectively. (**a**,**c**,**e**,**g**) Images; (**b**,**d**,**f**,**h**) Corresponding radar point clouds.

4.2. Training Details

In our method, the detection model and the image–radar association model are trained separately. In the training of the vision-based detection model, our model based on PyTorch implements is pre-trained from the COCO dataset [46] and is trained on an Nvidia GTX 3090 with an initial learning rate set to 10^{-3} and the momentum of 0.937. The network is trained for 100 epochs using the SGD optimizer with a batch size of 8 and the mini-batch StepLR descent algorithm with step-size = 1, gamma = 0.94. Besides, in order to enhance the performance of the detection model, multiscale data augmentation methods such as image resizing image placing, color adjustment, and image left-right flipping are used for the training images. During the training of the image–radar association model, using the extrinsic parameters provided in the dataset, we generate 4600 pairs of objects' radar point clouds and RGB image bounding boxes to train the metric learning model. The training implementation of our association network is based on Pytorch and is trained with a batch size set to 16 and an initial learning rate set to 10^{-4}. We train the image–radar association model for 200 epochs using an ADAM optimizer with weight decay of 5×10^{-4}.

4.3. Quantitative Evaluation

To evaluate the performance of our method, we compared our method with other baseline methods. All the baselines and our model are trained on the same training set. As different model training parameters influence the final results, in the experiment, the training parameters for the baseline models are primarily set by following its recommended default values with only minor adjustments.

The result in Table 1 shows that, compared with other vision-based baseline methods, the proposed method achieves better detection accuracy while keeping a relatively low computation cost. The result in Table 2 shows that our mmWave radar-aided vision detection method outperforms other radar-based and most visual–radar fusion-based methods in detection accuracy. Although the RISFNet achieves higher detection accuracy, it has a higher computation cost and requires accurate extrinsic parameters between the radar and camera. When the extrinsic parameter is inaccurate, the performance of RISFNet decreases significantly.

In addition, we also combine our image–radar association model with other vision-based object detection methods. As shown in Table 3, by directly applying our image–radar association model, other vision-based methods all achieve obvious improvement in detection accuracy. As our image–radar association model has low computational complexity and the inference speed of the image–radar association model is extremely fast with about 280 FPS, the image–radar association model can also be seen as an independent plug-and-play model to improve the detection accuracy of the vision-based detection method.

Table 1. Comparison of the detection accuracy using vision-based baseline methods and our method on the dataset.

Method	mAP (IoU = 0.35, %)	FPS
YOLOv5-l [32]	74.66	29
Cascade-RCNN [42]	78.36	17
Faster-RCNN [41]	74.34	19
Swin-Transformer [43]	77.33	15
Ours	**81.41**	**29**

Table 2. Comparison of the detection accuracy using radar-based and fusion-based methods and our method on the dataset.

Modality	Method	mAP (IoU = 0.35, %)
Radar	VoteNet [44]	45.24
	Danzer et al. [45]	32.65
Vision + Radar	CRF-Net [20]	74.35
	Li et al. [25]	77.23
	Jha et al. [29]	77.98
	RISFNet [23]	83.25
	Ours	**81.41**

Table 3. The results of combining our image–radar association model with other vision-based object detection methods.

Method	mAP (IoU = 0.35, %) (with Image–Radar Association Model)	FPS
YOLOv5-l [32]	81.41 (**+6.75**)	29
Cascade-RCNN [42]	83.62 (**+5.26**)	15
Faster-RCNN [41]	79.53 (**+5.19**)	17
Swin-Transformer [43]	82.42 (**+5.09**)	19

4.4. Robustness Analysis

In order to validate the effect of camera–radar extrinsic parameter changes on the performance of the proposed model, based on the known extrinsic parameter, we artificially add a rotation and translation bias to the overall radar point clouds, to simulate variations in the camera–radar extrinsic parameters. As shown in Table 4, when the extrinsic parameters change slightly (with ±5° rotation bias, ±1 m translation bias), the model's performance is nearly not affected, indicating the model's adaptability to small changes in the extrinsic parameters. However, when the extrinsic parameters change a lot (with ±20° rotation bias, ±4 m translation bias), there is a significant decrease in the model's performance. Nevertheless, in the practical application of USVs, extrinsic parameters are unlikely to have large variations, and the model can adapt well to most scenes.

Table 4. Comparison of results of the fusion model with different extrinsic parameter variations.

Method	mAP (IoU = 0.35, %)
Using origin radar data	81.41
Using radar data with slight bias	80.83
Using radar data with large bias	56.29

4.5. Ablation Analysis

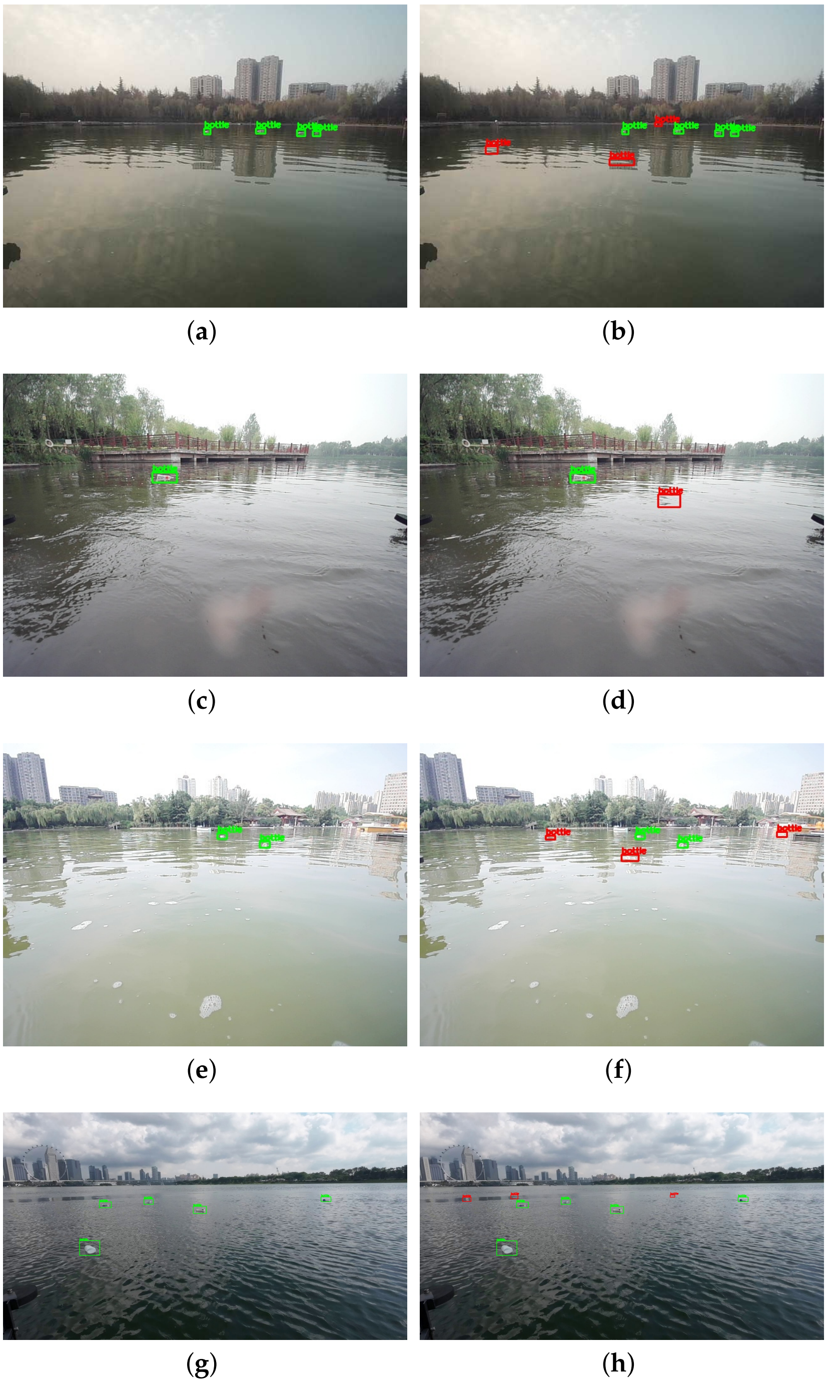
To verify the contributions of the proposed modifications to the YOLOv5 model's prediction head, we conduct an ablation analysis by replacing it with the original prediction head. Furthermore, to test the effectiveness of the newly proposed image–radar association model for data fusion, we compare it with the traditional manual configuration fusion method. The method directly projects the mmWave radar point cloud onto the RGB image plane based on the initial extrinsic parameters. Then, data association is performed based on the spatial relationships between radar point cloud clusters and 2D image boxes in the image plane with a predefined distance threshold. The results are shown in Table 5, indicating that the proposed improved double prediction heads effectively enhance the object detection accuracy. Besides, the proposed metric learning-based image–radar association model achieves better fusion results compared to the traditional manual association method.

Table 5. Results of the ablation analysis.

Method	mAP (IoU = 0.35, %)
Without double prediction heads	79.85
Without image–radar association model	78.17
Our method	81.41

4.6. Discussion

The visualization of the detection results of our fusion detection method compared with the vision-based YOLOv5-l is shown in Figure 5. As can be seen, our method achieves a lower false object detection rate in various surrounding scenes.



(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

(**g**) (**h**)

Figure 5. Visualization of results of the proposed method and vision-based YOLOv5-l. The red and green boxes in the figures represent correct and false detection results, respectively.(**a**,**c**,**e**,**g**) Ours; (**b**,**d**,**f**,**h**) YOLOv5-l.

To enhance the detection performance of the YOLOv5-l visual detector, we integrate a transformer decoder module to replace the conventional convolutional blocks within its prediction head. The transformer decoder uses multi-head attention to enhance the low-resolution feature representation capability. We visualize the input and output feature maps of the transformer decoder module in Figure 6. In Figure 6, each row represents a frame, where the first column displays the original image with the final detection results (highlighted in red bounding boxes), the second column shows the output features of the backbone network, and the third column show cases the features strengthened through the transformer decoder module. The highlighted regions denote areas of high response, indicating a higher probability of object presence in those regions. As can be seen, the results demonstrate that small objects in the enhanced features of this module are more distinguishable, so that our vision-based detector can locate all objects more accurately.

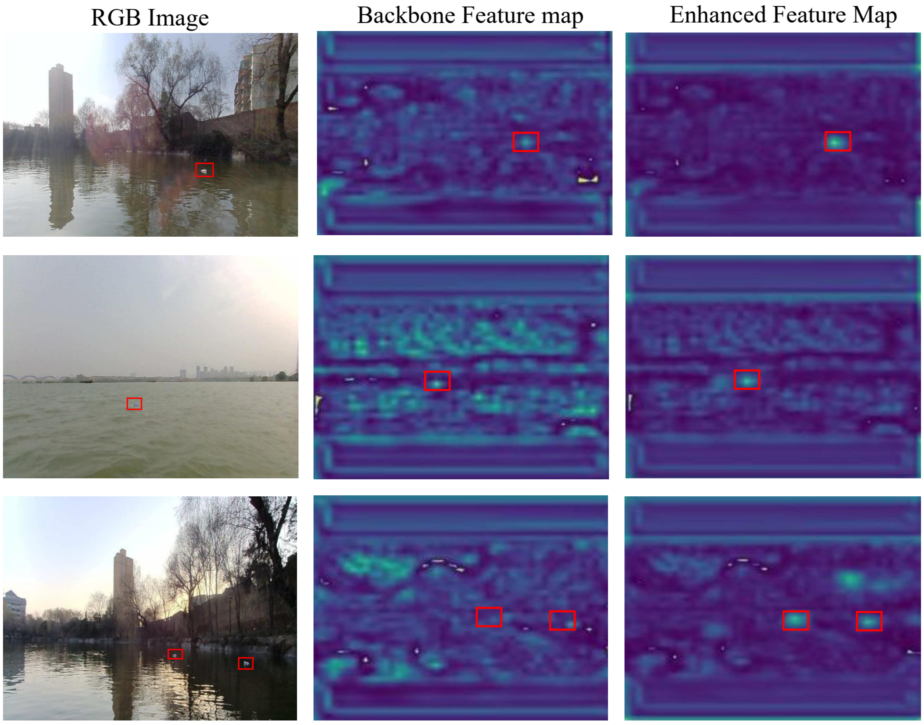


Figure 6. Feature maps of the transformer decoder module in enhanced YOLOv5-l.

5. Conclusions

In this paper, we propose a new mmWave radar-aided visual water surfaces small object detection method. The method associates mmWave radar data and images through the metric learning model, and is adapted to changes of extrinsic parameters to some degree. Through the detection stage and the fusion association stage, the proposed method outputs the final fusion detection results. Finally, we conduct experiments on the real-world dataset to test the proposed method. The results show that our method outperforms other visual detection methods on water surface small object detection.

Author Contributions: Conceptualization, J.Z., Y.Y. and Y.C.; methodology, J.Z. and Y.Y.; validation, J.Z.; data curation, J.Z.; coding and experiments, Y.C.; writing—original draft preparation, J.Z. and Y.C.; writing—review and editing, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China under Grant Number 11974286.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CFAR	Constant false alarm rate
DBSCAN	Density-based spatial clustering of applications with noise
DOA	Direction of arrival
FC	Full connect
FFT	Fast Fourier Transformation
FMCW	Frequency-modulated continuous wave
FPS	Farthest point sampling
GNSS	Global navigation satellite system
IMU	Inertial measurement unit
mAP	Mean of average precision
MLP	Multi-layer perception
mmWave	Millimeter wave radar
RDM	Range–Doppler matrix
USV	Unmanned surface vehicle

References

1. Wang, W.; Gheneti, B.; Mateos, L.A.; Duarte, F.; Ratti, C.; Rus, D. Roboat: An autonomous surface vehicle for urban waterways. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Venetian Macao, Macau, 4–8 November 2019; pp. 6340–6347.
2. Chang, H.C.; Hsu, Y.L.; Hung, S.S.; Ou, G.R.; Wu, J.R.; Hsu, C. Autonomous Water Quality Monitoring and Water Surface Cleaning for Unmanned Surface Vehicle. *Sensors* **2021**, *21*, 1102. [CrossRef]
3. Zhu, J.; Yang, Y.; Cheng, Y. SMURF: A Fully Autonomous Water Surface Cleaning Robot with A Novel Coverage Path Planning Method. *J. Mar. Sci. Eng.* **2022**, *10*, 1620. [CrossRef]
4. Wu, Y.; Wang, Y.; Zhang, S.; Ogai, H. Deep 3D object detection networks using LiDAR data: A review. *IEEE Sens. J.* **2020**, *21*, 1152–1171. [CrossRef]
5. Carballo, A.; Lambert, J.; Monrroy, A.; Wong, D.; Narksri, P.; Kitsukawa, Y.; Takeuchi, E.; Kato, S.; Takeda, K. LIBRE: The multiple 3D lidar dataset. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 1094–1101.
6. Patole, S.M.; Torlak, M.; Wang, D.; Ali, M. Automotive Radars: A Review of Signal Processing Techniques. *IEEE Signal Process. Mag.* **2017**, *34*, 22–35. [CrossRef]
7. Brodeski, D.; Bilik, I.; Giryes, R. Deep radar detector. In Proceedings of the 2019 IEEE Radar Conference (RadarConf), Boston, MA, USA, 22–26 April 2019; pp. 1–6.
8. Hammedi, W.; Ramirez-Martinez, M.; Brunet, P.; Senouci, S.M.; Messous, M.A. Deep learning-based real-time object detection in inland navigation. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6.
9. Moosbauer, S.; Konig, D.; Jakel, J.; Teutsch, M. 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, Long Beach, CA, USA, 16–17 June 2019.
10. Prasad, D.K.; Rajan, D.; Rachmawati, L.; Rajabally, E.; Quek, C. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1993–2016. [CrossRef]
11. Zhou, Z.; Yu, S.; Liu, K. A Real-time Algorithm for Visual Detection of High-speed Unmanned Surface Vehicle Based on Deep Learning. In Proceedings of the 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Chongqing, China, 11–13 December 2019; pp. 1–5.
12. Zhang, W.; Gao, X.z.; Yang, C.f.; Jiang, F.; Chen, Z.y. A object detection and tracking method for security in intelligence of unmanned surface vehicles. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *13*, 1279–1291. [CrossRef]
13. Li, Y.; Guo, J.; Guo, X.; Liu, K.; Zhao, W.; Luo, Y.; Wang, Z. A novel target detection method of the unmanned surface vehicle under all-weather conditions with an improved YOLOV3. *Sensors* **2020**, *20*, 4885. [CrossRef] [PubMed]
14. Wu, Y.; Qin, H.; Liu, T.; Liu, H.; Wei, Z. A 3D object detection based on multi-modality sensors of USV. *Appl. Sci.* **2019**, *9*, 535. [CrossRef]
15. Cardillo, E.; Ferro, L. Multi-frequency analysis of microwave and millimeter-wave radars for ship collision avoidance. In Proceedings of the 2022 Microwave Mediterranean Symposium (MMS), Pizzo Calabro, Italy, 9–13 May 2022; pp. 1–4.
16. Im, S.; Kim, D.; Cheon, H.; Ryu, J. Object Detection and Tracking System with Improved DBSCAN Clustering using Radar on Unmanned Surface Vehicle. In Proceedings of the 2021 21st International Conference on Control, Automation and Systems (ICCAS), Jeju, Republic of Korea, 12–15 October 2021; pp. 868–872.

17. Ha, J.S.; Im, S.R.; Lee, W.K.; Kim, D.H.; Ryu, J.K. Radar based Obstacle Detection System for Autonomous Unmanned Surface Vehicles. In Proceedings of the 2021 21st International Conference on Control, Automation and Systems (ICCAS), Jeju, Republic of Korea, 12–15 October 2021; pp. 863–867.
18. Stanislas, L.; Dunbabin, M. Multimodal sensor fusion for robust obstacle detection and classification in the maritime RobotX challenge. *IEEE J. Ocean. Eng.* **2018**, *44*, 343–351. [CrossRef]
19. Long, Y.; Morris, D.; Liu, X.; Castro, M.; Chakravarty, P.; Narayanan, P. Radar-camera pixel depth association for depth completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12507–12516.
20. Nobis, F.; Geisslinger, M.; Weber, M.; Betz, J.; Lienkamp, M. A deep learning-based radar and camera sensor fusion architecture for object detection. In Proceedings of the 2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF), Bonn, Germany, 15–17 October 2019; pp. 1–7.
21. Nabati, R.; Qi, H. Rrpn: Radar region proposal network for object detection in autonomous vehicles. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3093–3097.
22. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
23. Cheng, Y.; Xu, H.; Liu, Y. Robust Small Object Detection on the Water Surface Through Fusion of Camera and Millimeter Wave Radar. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 15263–15272.
24. Chadwick, S.; Maddern, W.; Newman, P. Distant vehicle detection using radar and vision. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–4 June 2019; pp. 8311–8317.
25. Li, L.q.; Xie, Y.l. A feature pyramid fusion detection algorithm based on radar and camera sensor. In Proceedings of the 2020 15th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 6–9 December 2020; Volume 1, pp. 366–370.
26. Nabati, R.; Qi, H. Centerfusion: Center-based radar and camera fusion for 3D object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Online, 5–9 January 2021; pp. 1527–1536.
27. Chang, S.; Zhang, Y.; Zhang, F.; Zhao, X.; Huang, S.; Feng, Z.; Wei, Z. Spatial attention fusion for obstacle detection using mmwave radar and vision sensor. *Sensors* **2020**, *20*, 956. [CrossRef] [PubMed]
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
29. Jha, H.; Lodhi, V.; Chakravarty, D. Object detection and identification using vision and radar data fusion system for ground-based navigation. In Proceedings of the 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 7–8 March 2019; pp. 590–593.
30. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
31. Domhof, J.; Kooij, J.F.; Gavrila, D.M. An extrinsic calibration tool for radar, camera and lidar. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 8107–8113.
32. Ultralytics. YOLO-v5. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 20 August 2023).
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
34. Bäcklund, H.; Hedblom, A.; Neijman, N. A density-based spatial clustering of application with noise. *Data Min. TNM033* **2011**, *33*, 11–30.
35. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5105–5114.
36. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
37. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
38. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
39. Cheng, Y.; Zhu, J.; Jiang, M.; Fu, J.; Pang, C.; Wang, P.; Sankaran, K.; Onabola, O.; Liu, Y.; Liu, D.; et al. FloW: A Dataset and Benchmark for Floating Waste Detection in Inland Waters. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10953–10962.
40. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
41. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015, pp. 1440–1448.
42. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
43. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
44. Qi, C.R.; Litany, O.; He, K.; Guibas, L.J. Deep hough voting for 3d object detection in point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9277–9286.

45. Danzer, A.; Griebel, T.; Bach, M.; Dietmayer, K. 2d car detection in radar data with pointnets. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 17–30 October 2019; pp. 61–66.
46. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.