World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:18, No:3, 2024

RV-YOLOX: Object Detection on Inland Waterways Based on Optimized YOLOX through Fusion of Vision and 3+1D Millimeter Wave Radar

Zixian Zhang, Shanliang Yao, Zile Huang, Zhaodong Wu, Xiaohui Zhu, Yong Yue, Jieming Ma

Abstract—Unmanned Surface Vehicles (USVs) hold significant value for their capacity to undertake hazardous and labor-intensive operations over aquatic environments. Object detection tasks are significant in these applications. Nonetheless, the efficacy of USVs in object detection is impeded by several intrinsic challenges, including the intricate dispersal of obstacles, reflections emanating from coastal structures, and the presence of fog over water surfaces, among others. To address these problems, this paper provides a fusion method for USVs to effectively detect objects in the inland surface environment, utilizing vision sensors and 3+1D Millimeter-wave radar. The MMW radar is a complementary tool to vision sensors, offering reliable environmental data. This approach involves the conversion of the radar's 3D point cloud into a 2D radar pseudo-image, thereby standardizing the format for radar and vision data by leveraging a point transformer. Furthermore, this paper proposes the development of a multi-source object detection network, named RV-YOLOX, which leverages radar-vision integration specifically tailored for inland waterway environments. The performance is evaluated on our self-recording waterways dataset. Compared with the YOLOX network, our fusion network significantly improves detection accuracy, especially for objects with bad light conditions.

Keywords—Inland waterways, object detection, YOLO, sensor fusion, self-attention, deep learning.

I. Introduction

OBJECT detection in surface environments plays a crucial role in various applications, including autonomous vehicles, surveillance, and environmental monitoring. It enables systems to recognize and locate objects within their surroundings, facilitating safer navigation, enhanced security measures, and informed decision-making. This technology is particularly pivotal in the development and operation of USVs, which are revolutionizing the way time-consuming and hazardous missions are executed on water surfaces.

The deployment of USVs leverages advanced object detection capabilities to undertake various critical tasks, including storm forecasting [1], [2], water quality monitoring [3] and floating waste cleaning [4], [5], among others.

Similar to the road environment, autonomous driving technologies are crucial for ensuring safe and efficient operation on inland waterways. Among these, reliable environmental perception of the surrounding area is essential for the effective functioning of USVs. Currently, researchers employ a variety of sensors for intelligent environmental perception. The vision sensor is the most widely used for object detection, allowing USVs to gather detailed information about the surface environment. However, compared to the road environment, pure visual perception faces more challenges in the surface environment, as described below:

- High variance in visual size: Due to the expansive scenes on the surface, the distance between the camera and objects covers a wide range. When the distance from the camera is great, the space the target occupies in RGB images decreases significantly.
- Interference from sunlight reflection or adverse weather conditions like fog or drizzle: The unpredictable environment can lead to overexposure in images, resulting in a loss of detail and a washed-out appearance.

On the other hand, 3+1D Millimeter-wave (MMW) radar properties are complementary to vision. It is immune to poor weather conditions and long-distance conditions and has the ability to evaluate the target velocity and depth using the Doppler theory. Moreover, compared with traditional automotive radar, recently appearing 3+1D MMW radar provides a much more dense point cloud and one extra dimension: elevation; even though, pure radar detection has some limitations on the surface environment.

- Limited resolution of MMW radar: Even for 3+1D millimeter-wave (MMW) radar, the resolution remains limited, meaning the radar may struggle to distinguish objects that are closely spaced or have similar sizes and shapes.
- Surface obstacles disturbance: Obstacles like the surface of the water and buoy may absorb or scatter radar waves, making it difficult to detect objects close to the surface.
- Environmental clutter: Objects in the environment, such as birds and insects, can generate false returns, impacting radar accuracy.

Multiple research efforts have demonstrated that the fusion of camera and radar technologies can achieve more accurate results in object detection compared to using a single type of sensor. However, based on our review, both the fusion of 3+1D MMW radar and camera for object detection and the combination of data from images and radar to detect objects in aquatic environments are topics that have not been extensively explored. To enhance the accuracy and robustness of object detection in surface environments using 3+1D MMW radar and camera, this paper proposes an object detection

Zixian Zhang, Shanliang Yao, Zile Huang, Zhaodong Wu, Xiaohui Zhu*, Yong Yue and Jieming Ma are with School of Advanced Technology , Xi'an Jiaotong-Liverpool University, Suzhou, China (*corresponding author, e-mail: {zixian.zhang21, shanliang.yao19, zile.huang, zhaodong.wu}@student.xjtlu.edu.cn, {xiaohui.zhu, yong.yue, jieming.ma}@xjtlu.edu.cn

Open Science Index, Electronics and Communication Engineering Vol:18, No:3, 2024 publications.waset.org/10013553.pdf

network based on the fusion of these technologies. The fusion mechanism and radar preprocessing methodology are the primary focus of this research. In summary, this paper contributes to the following aspects:

- A radar-camera fusion object detection network (RV-YOLOX) for USVs in inland waterways environment is proposed. This detection network is based on the structure of the YOLOX network, optimizing by adding a multi-stage attention fusion block, which is responsible for integrating radar and vision features. This network is evaluated by our self-recorded data.
- A radar feature image generation model based on a neural network is proposed. This model can automatically extract radar features and transform them into radar feature images by BP (Back Propagation) radar semantic information. The generated radar feature images are used as the input of our proposed detection network.

II. Related Work

A. Object Detection on Inland Waterways Environment

Recently, the interest in object detection algorithms for the inland waterways environment has gradually increased due to their wide range of applications in navigation, collision avoidance, and other tasks. For pure vision detection, Hammedi et al. [6] tested different vision-based algorithms using a dataset of inland objects and assessed their performance in detecting these objects. Yang et al. [7] utilized an improved SSD (single-shot detector) model to optimize the detection performance of small and occluded targets. For radar-camera fusion detection, Yuwei et al. [8] proposed a global attention fusion mechanism to improve detection performance for small objects. However, the weather conditions and scene contents in their evaluation datasets are limited. To the best of our knowledge, object detection algorithms for complex scenes remain relatively scarce.

B. Object Detection Based on Vision Sensor

The most effective and commonly utilized deep learning models for object detection in images are convolutional neural networks (CNNs). CNN-based detectors can be categorized into two general types: two-stage detectors and one-stage detectors. Two-stage detectors, such as RCNN [9] and Fast-RCNN [10], generate a set of region proposals in the first stage, which are then refined and classified in the second stage. This approach usually results in higher accuracy but is slower compared to one-stage detectors.

On the other hand, one-stage detectors, such as YOLO (You Only Look Once) [11] and SSD [12], directly map feature maps to bounding box regression and approach object detection as a regression problem. This method is faster but traditionally less accurate compared to two-stage detectors. However, with advancements in the YOLO family, some novel iterations have achieved accuracy comparable to two-stage object detection methods while maintaining faster detection speeds. Redmon et al. first introduced the YOLO network in 2016. Subsequently, many researchers have applied state-of-the-art (SOTA) mechanisms to YOLO. In 2017, YOLOv2 [13] incorporated anchors [14]. YOLOv3 [15] utilized Residual Net [16]. More recently, YOLOv4 [17] and YOLOv5 [18] were proposed, achieving impressive performance and speed advantages. A year later, Ge's team released the YOLOX network [19]. YOLOX has five different configurations, balancing higher accuracy or faster speed. The mAP of YOLOX-L is comparable to YOLOv5-L on the COCO dataset, while the inference speed of YOLOX-L is faster than that of YOLOv5-L. However, these models still face limitations in challenging environments, such as adverse weather conditions. Therefore, this paper selects the YOLOX network as the foundation for improvements based on radar-camera fusion.

C. Object Detection Based on Radar and Image Fusion

The use of multiple sensors, such as cameras and MMW radar, for object detection in vehicles holds significant practical and theoretical value. MMW radar and cameras are highly complementary, providing abundant semantic and relative speed information to enhance the robustness and accuracy of object detection. Early researchers adopted traditional methods to fuse MMW radar and camera data. In 2002, Bruno et al. [20] proposed a low-level fusion system for vehicle detection. Bombini et al. [21] utilized radar information as a region of interest in images to improve vehicle detection accuracy.

As deep learning technology has advanced, feature-level fusion for radar-camera systems has garnered increasing attention recently. For feature-level fusion, where the fusion process occurs at a higher level of abstraction, radar feature extraction and the fusion detection framework are crucial. Chadwick et al. [22] were among the first to transform radar point cloud data into images, fusing radar and vision images using ResNet to enhance performance under challenging scenes. Nobis et al. [23], Chang et al. [24], and Li et al. [25] adopted a similar approach to processing radar point cloud data. For the detection framework, John et al. [26] proposed a framework called RVNet for combining image and radar data using deep learning techniques. This framework effectively merges features from both sensor types and can detect obstacles in real time. In 2020, Kowol et al. [27] introduced YOdar, which uses two separate branches for extracting radar and vision features and combines them in an uncertainty-aware manner. In 2022, Song et al. [28] proposed a fusion framework based on YOLOv5, achieving impressive performance on their custom datasets. However, most existing processes for radar features are in the image dimension, overlooking the potential of radar features in the 3D dimension. Moreover, the detection frameworks tend to be relatively simple.

III. Radar Data Preprocessing

Given the differences between MMW radar data and vision data, as well as the discrepancies in sensor coordinate systems, it is crucial to standardize the data format and spatial relationship between radar and vision data. Two prevalent methods exist for processing this data. The first involves directly transforming radar point clouds into radar

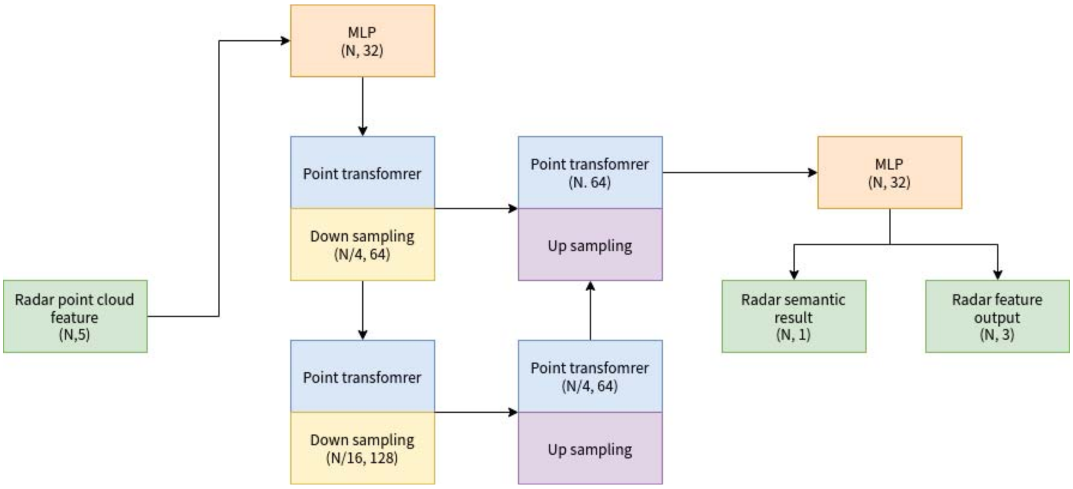


Fig. 1 Schematic diagram of the radar feature generation model

pseudo-images based on radar-camera calibration information, as demonstrated in works like [22]–[25]. The second method maps radar information onto mask maps using neural networks, as seen in [28]. Both approaches convert radar features into a 2D image format, overlooking the potential of radar's 3D positional features. In contrast, this paper aims to amplify the impact of radar features during the fusion process by proposing a radar feature image generation model that fully leverages radar's 3D positional attributes.

A. Radar Image Generation

Self-attention networks have made significant advancements in various fields, including natural language processing and image classification. Recently, researchers have begun applying self-attention networks to the processing of 3D point clouds. Inspired by these developments, we extract radar feature images using a self-attention network. This network is tasked with processing MMW radar point cloud data and transferring the extracted radar features to RGB images.

The structure of our feature-extraction self-attention network is illustrated in Fig. 1. Five original radar features serve as the inputs to this network, including 3D localized features (X, Y, Z), Radar Cross Section (RCS) value, and Doppler velocity. The network aims to extract scene segmentation details and a radar point 3-channel feature, which is then used for transforming into an RGB image for sensor fusion. Consequently, this network features two output branches: one for the backpropagation (BP) network based on radar segmentation information, and another for outputting the radar point cloud's 3-channel feature. The network's backbone is constructed based on the original Point Transformer [29], with the following modifications to better accommodate radar point cloud data:

- We reduced the depth of the U-net structure because our experiments showed that a structure that is too deep is counterproductive for extracting radar features.
- We replaced the transition down and transition up blocks in the original Point Transformer network with direct down-sampling and up-sampling operations. This adjustment is based on the observation that the low density of radar point cloud data render the K-nearest neighbors (KNN) and interpolation operations in the transition blocks unnecessary for our feature-extraction network.

After extracting radar features using our radar feature-extraction self-attention network, we generate 2D radar images from the extracted 3-channel radar features. These 3-channel radar features are projected onto image pixel locations, which are calculated as follows:

$$X_A = I_A P_{AR} T_R \tag{1}$$

where I_A is the 3×4 camera intrinsic matrix, T_R is the extrinsic matrix between the camera and radar, and P_{AR} is the 3D localized features of the radar point clouds. After projection, the pixel values of the radar image are converted based on the values of the 3-channel radar features. Since the 3-channel radar features have been normalized to the range (0, 1) in the previous network, they can directly represent the RGB value for the radar image. For areas not covered by radar features, all channels are set to 0. However, we observed that the density of areas containing radar points is too low. Therefore, inspired by Chang et al. [24], we expand radar points into radar circles with a radius of r pixels. The area within these radar circles shares the same pixel value, with the center of the radar circles representing the location of radar points in pixel coordinates.

IV. Detection Network

The deep learning object detection network with multi-data source fusion presented in this paper is based on the YOLOX framework [19]; hence, it is named RV-YOLOX. Given the impressive performance of YOLOX, as evidenced by evaluation results, we employ YOLOX as the foundation for our fusion detection network. To effectively utilize radar features for object detection across various scales, we fuse image and radar features of different sizes using the self-attention mechanism.

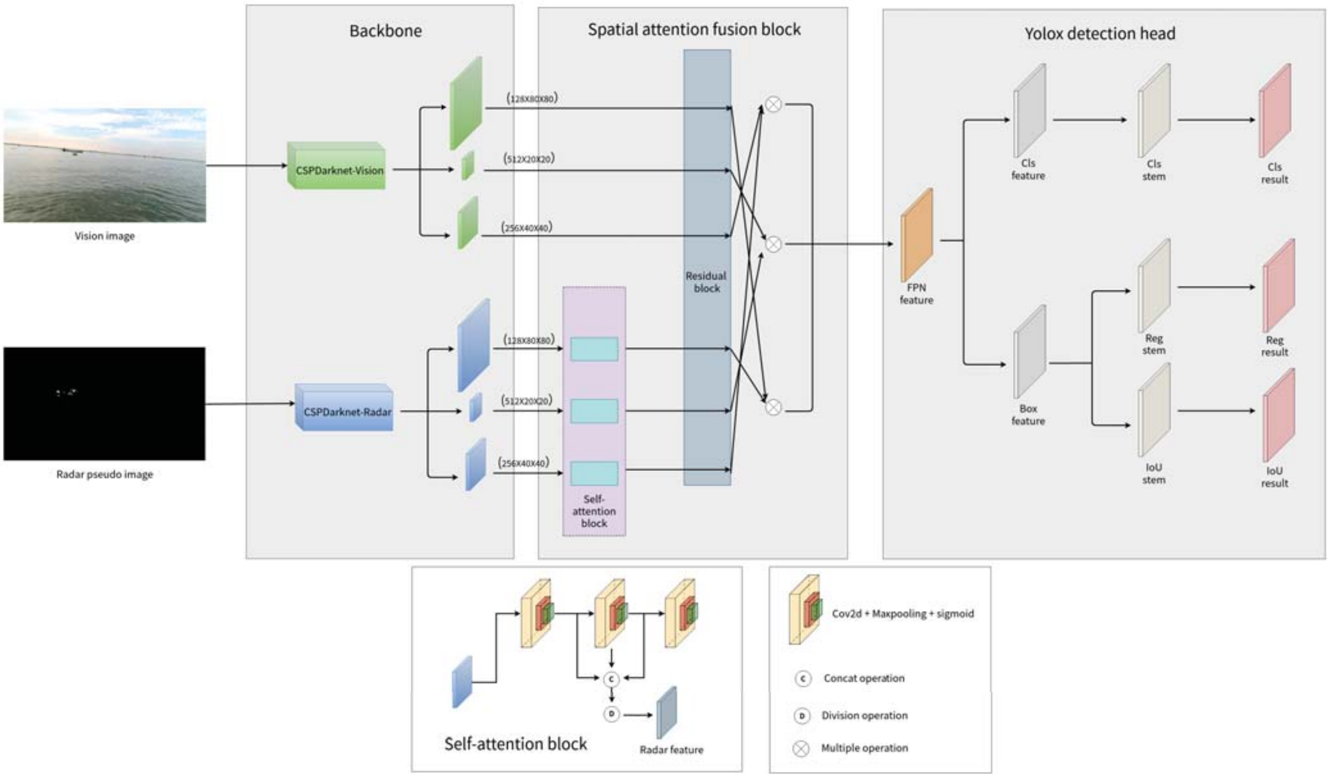


Fig. 2 The RV-YOLOX network architecture

The network architecture is illustrated in Fig. 2. This network primarily consists of three parts: the backbone, the self-attention fusion block, and the YOLOX detection head. The backbone is made up of two parallel structures designed to extract features from vision images and radar pseudo images, respectively. The self-attention fusion block is tasked with fusing multi-scale image and radar features, extracted by the backbone, across all channels using the self-attention mechanism. Finally, the fusion results are passed to the YOLOX decoupled head for the prediction of detection bounding boxes. Further details of our detection network are provided below:

- *Backbone:* Fig. 2 presents the specifics of the backbone block. The inputs to this block are the source vision image and the radar pseudo image produced by the preceding radar image generation model. This block comprises two backbone networks: CSPDarknet-vision and CSPDarknet-radar, each tasked with extracting features from vision and radar data, respectively. The network outputs feature maps of multiple sizes for subsequent multi-scale fusion. The dimensions of the final image and radar features are $128 \times 80 \times 80$, $256 \times 40 \times 40$, and $512 \times 20 \times 20$.
- *Self-attention fusion block:* As is widely recognized, general environmental information surrounding objects can be extracted from radar points. Hence, leveraging radar points as a guide to direct vision sensors in feature extraction can be beneficial. Our aim is to enhance detection performance by increasing the emphasis on radar features for small and blurred objects, which are typically lacking in clear vision features. Moreover, for objects that possess abundant vision information, radar points are expected to provide a positive contribution as well. Additionally, in contrast to data-level fusion, which may overlook objects in regions without radar points, our proposed fusion strategy thoroughly accounts for the condition of areas devoid of any radar points.
The data fusion scheme employed in our detection framework features a self-attention fusion block tasked with re-weighting the feature maps from the vision branch. This re-weighting is based on a 2D matrix that is biased and iteratively adjusted according to the features from the radar branch. Specifically, the features of radar images are encoded into a 2D spatial attention matrix. Subsequently, all channels of the feature maps from the vision branch are re-weighted based on this biased 2D spatial attention matrix. The mathematical representation is detailed below:

$$\mathbf{F}_{\text{fusion}} = \mathbf{F}_{\text{image}} \cdot \mathbf{W_1} \tau \left(\mathbf{W_0} \left(\sigma \left(\mathbf{F}_{\text{radar}} \right) \right) \right) \qquad (2)$$

where τ represents ReLU function and σ denotes sigmoid function. $\mathbf{F}_{\text{fusion}}$, $\mathbf{F}_{\text{image}}$ and $\mathbf{F}_{\text{radar}}$ represent the fused features, image features, and radar features respectively. $\mathbf{W_1}$, $\mathbf{W_0}$ denotes MLP weights for assigning radar weights to image features.

A. Loss Function

The convergence speed of the network is enhanced by employing an appropriate loss function, similar to that used

TABLE I
DETECTION RESULTS COMPARISON FOR RV-YOLOX AND YOLOX BASED ON OUR SURFACE RADAR-CAMERA DATASET

Model	*AP(100)*	$AP^{.50}$*(100)*	$AP^{.75}$*(100)*	AP^{s}*(100)*	AP^{m}*(100)*	AP^{l}*(100)*
YOLOX-S	60.3	82.9	66.4	42.7	60.3	71.3
YOLOX-L	63.6	88.3	71.9	46.5	62.5	74.3
YOLOX-M	62.4	85.2	68.9	45.2	61.7	73.7
YOLOX-X	64.9	87.2	70.9	47.5	62.9	77.7
RV-YOLOX-S	*65.4*	84.5	71.0	48.7	64.3	73.5
RV-YOLOX-L	*72.6*	92.3	79.6	58.7	68.0	80.6
RV-YOLOX-M	*70.8*	89.7	75.4	54.3	67.3	79.7
RV-YOLOX-X	*73.4*	91.2	81.2	57.3	70.8	84.6
Model	*AR(1)*	*AR(10)*	*AR(100)*	AR^{s}*(100)*	AR^{m}*(100)*	AR^{l}*(100)*
YOLOX-S	12.2	57.3	65.5	54.1	69.2	76.7
YOLOX-L	14.4	62.1	72.3	60.0	72.6	81.5
YOLOX-M	13.2	61.6	71.2	58.7	70.9	79.3
YOLOX-X	14.8	63.2	74.4	62.1	73.1	83.3
RV-YOLOX-S	*13.2*	58.8	67.9	56.4	71.0	77.8
RV-YOLOX-L	*16.0*	68.7	78.4	67.1	73.6	86.8
RV-YOLOX-M	*15.3*	67.2	66.5	65.8	71.6	84.2
RV-YOLOX-X	*16.4*	69.4	79.3	70.4	75.4	87.5

in YOLOX. The YOLOX loss function comprises three components: a classification branch, a confidence branch, and a regression branch. Both the classification and confidence branches utilize the Binary Cross Entropy (BCE) loss for calculation, which is expressed as follows:

$$\mathrm{BCE} = -\log\left(\mathrm{P_t}\right) = \begin{cases} -\log(\hat{\mathrm{y}}), & \mathrm{y}=1 \\ -\log(1-\hat{y}), & \mathrm{y}=0 \end{cases} \tag{3}$$

where y = 1 represents positive sample, while y = 0 represents negative sample. On the other hand, the regression branch is calculated by IoU loss, which is expressed as:

$$\text{IoU loss } = -\sum_{i\in\{t,b,l,r\}} \ln \frac{\text{Intersection}(x_i,\widetilde{x}_i)}{\text{Union }(x_i,\widetilde{x}_i)} \tag{4}$$

where x_i denotes the ground truth area and $\widetilde{x}_i)$ denotes the prediction area.

B. Parameter Detail

In this section, we specify some key parameters used to build this detection model. The preprocessing procedure for vision and radar pseudo images is the same, including mosaic, random affine, mix-up, and random augmentation. The number of detection classes is 4, containing pier, ship, boat, and vessel. The input size of images is 640×640. The self-attention block contains three parallel convolution layers with the size of 1×1, 3×3, and 5×5 for multi-scale fusion. The training procedure is epoch-based, and the total epoch size is 80. The quadratic warm-up scheme is utilized to adjust the learning rate in the first few iterations and the base learning rate 0.01. We adopt the Adam optimization algorithm to process BP.

V. EXPERIMENTS

To convincingly demonstrate that our multi-sensor fusion network surpasses the performance of pure-vision detection networks, we conducted eight experiments. These include testing four YOLOX networks of varying sizes (YOLOX-L, YOLOX-M, YOLOX-S, YOLOX-X) and four improved fusion networks corresponding to the original YOLOX configurations. The results presented in this paper are evaluated using standard COCO evaluation metrics, which include average precision (AP) and average recall (AR).

A. Dataset Description

The training and evaluation dataset utilized in our study is a portion of our self-recorded dataset [30], encompassing recordings from creeks, canals, lakes, and rivers. These locations were chosen to represent a variety of weather and water conditions. The data were captured using a 3+1D MMW radar and a monocular camera. All recorded frames include ground truth annotations for 2D bounding boxes for vision sensors and segmentation details for the radar sensor. The dataset segment we used is categorized into four classes: pier, ship, boat, and vessel, with each class annotated with both vision and radar data. For our experiments, we used a total of 24,000 frames, dividing them into a training set of 19,200 frames and a testing set of 4,800 frames. The distribution of the number of objects for each class and the 3+1D MMW radar point density is illustrated in Fig. 4.

B. Comparison and Analysis

We conducted quantitative comparisons between RV-YOLOX models and other YOLOX networks based on the average precision (AP) and average recall (AR)

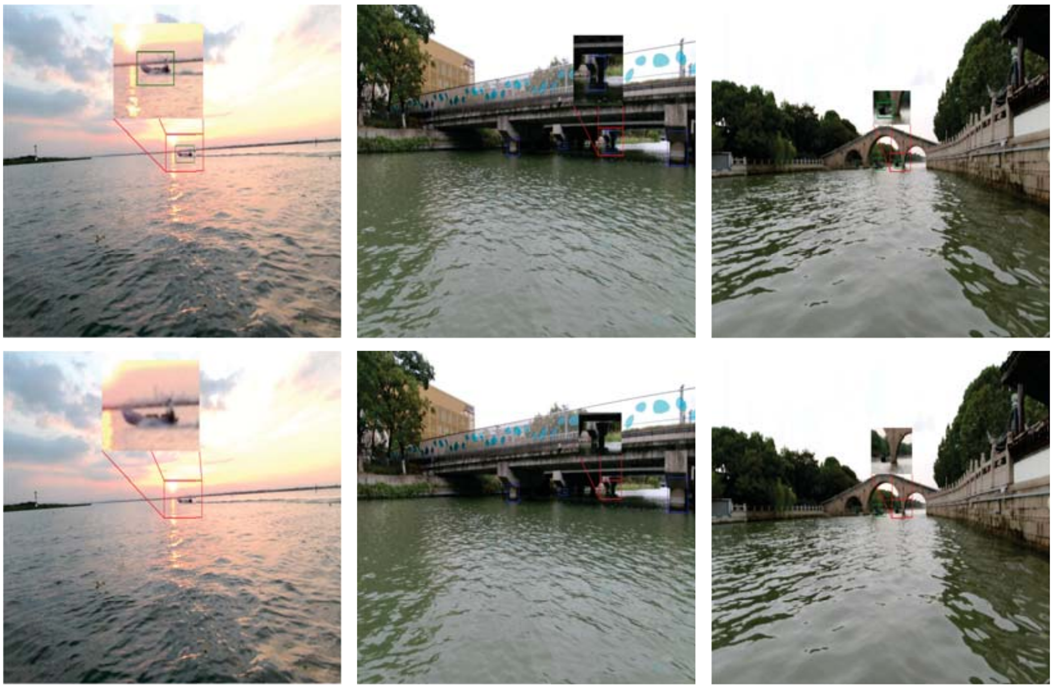
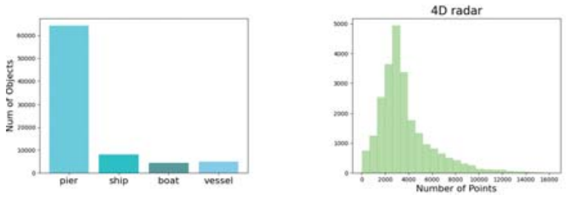


Fig. 3 Comparison of the detection capabilities of YOLOX and RV-YOLOX for tiny ship, dark pier, and invisible pier; photos in the top row are RV-YOLOX detection findings, and images in the bottom row are YOLOX detection results; the comparisons demonstrate that the recommended RV-YOLOX performs better in tiny and ambiguous targets



((a)) Number of objects per class

((b)) 4D radar point cloud density

Fig. 4 Dataset general overview including class distribution and point cloud distribution

results from the standard COCO evaluation. To ensure fair comparisons among these models, all training setups and implementation details were kept identical. The comparison results are summarized in Table I. According to Table I, the performance of object detection is consistently improved when the 3+1D MMW radar source is incorporated. This comprehensive comparison confirms that the integration of radar features results in tighter and more accurate bounding box estimations. Among the eight models compared, RV-YOLOX-X demonstrated the highest performance, achieving an average precision of 73.4%.

We also conducted a visualization comparison between the YOLOX network and the RV-YOLOX network, particularly for objects in poor lighting conditions and occluded objects. The results are displayed in Fig. 3. From Fig. 3, it is evident that the performance of YOLOX and RV-YOLOX is comparable when detecting objects with ample visual information. However, in scenarios involving poor lighting or blurred objects, the incidence of missed detections is noticeably higher for YOLOX compared to RV-YOLOX. This improvement with RV-YOLOX can be attributed to the addition of radar features, such as RCS and Doppler velocity, making the network more resilient to environmental variations. In conclusion, based on both quantitative and qualitative assessments, RV-YOLOX proves to be more effective and robust for object detection tasks in inland water surface environments.

VI. Conclusion

In this paper, we explored object detection in inland waterways utilizing 3+1D MMW radar and cameras. We proposed both a radar feature extraction model and a fusion detection network that leverages radar and vision information. The radar feature extraction model employs a point transformer architecture to extract features from 3D point clouds and convert radar-extracted features into RGB images. The fusion detection network, named RV-YOLOX, introduces an additional radar input branch and a self-attention fusion block to deeply integrate radar and vision information on the YOLOX network framework. In the experimental section, we conducted a comprehensive evaluation using different sizes of YOLOX (YOLOX-L, YOLOX-M, YOLOX-S, YOLOX-X) for both pure vision and radar-vision sources. The findings indicate that RV-YOLOX significantly enhances performance, particularly in low-light conditions, compared to pure-vision approaches. For future work, we aim to expand our research to include semantic segmentation tasks for USVs in inland waters, utilizing both vision and radar sensors.

ACKNOWLEDGMENT

This research was funded by the Suzhou Science and Technology Project (SYG202122), the Key Programme Special Fund of Xi'an Jiaotong-Liverpool University (XJTLU) (KSF-A-19, KSF-E- 350, KSF-P-02, KSF-E-54), the Research Development Fund of XJTLU (RDF-19-02-23) and Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004).

REFERENCES

[1] J. Curcio, J. Leonard, and A. Patrikalakis, "Scout - a low cost autonomous surface platform for research in cooperative autonomy," in *Proceedings of OCEANS 2005 MTS/IEEE*, 2005, pp. 725–729 Vol. 1.

[2] G. Ferri, A. Manzi, F. Fornai, B. Mazzolai, C. Laschi, F. Ciuchi, and P. Dario, "Design, fabrication and first sea trials of a small-sized autonomous catamaran for heavy metals monitoring in coastal waters," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 2406–2411.

[3] D. Madeo, A. Pozzebon, C. Mocenni, and D. Bertoni, "A low-cost unmanned surface vehicle for pervasive water quality monitoring," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1433–1444, 2020.

[4] A. Akib, F. Tasnim, D. Biswas, M. B. Hashem, K. Rahman, A. Bhattacharjee, and S. A. Fattah, "Unmanned floating waste collecting robot," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 2645–2650.

[5] N. Ruangpayoongsak, J. Sumroengrit, and M. Leanglum, "A floating waste scooper robot on water surface," in *2017 17th International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2017, pp. 1543–1548.

[6] W. Hammedi, M. Ramirez-Martinez, P. Brunet, S.-M. Senouci, and M. A. Messous, "Deep learning-based real-time object detection in inland navigation," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[7] Y. Yang, P. Chen, K. Ding, Z. Chen, and K. Hu, "Object detection of inland waterway ships based on improved ssd model," *Ships and Offshore Structures*, pp. 1–9, 2022.

[8] Y. Cheng, H. Xu, and Y. Liu, "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*, 2021, pp. 15 263–15 272.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[10] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[13] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

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

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-yolov4: Scaling cross stage partial network," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2021, pp. 13 029–13 038.

[18] G. Jocher, "YOLOv5 by Ultralytics," 5 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[19] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[20] B. Steux, C. Laurgeau, L. Salesse, and D. Wautier, "Fade: A vehicle detection and tracking system featuring monocular color vision and radar data fusion," in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 2. IEEE, 2002, pp. 632–639.

[21] L. Bombini, P. Cerri, P. Medici, and G. Alessandretti, "Radar-vision fusion for vehicle detection," in *Proceedings of International Workshop on Intelligent Transportation*, vol. 65, 2006, p. 70.

[22] S. Chadwick, W. Maddern, and P. Newman, "Distant vehicle detection using radar and vision," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8311–8317.

[23] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. IEEE, 2019, pp. 1–7.

[24] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, "Spatial attention fusion for obstacle detection using mmwave radar and vision sensor," *Sensors*, vol. 20, no. 4, p. 956, 2020.

[25] L.-q. Li and Y.-l. Xie, "A feature pyramid fusion detection algorithm based on radar and camera sensor," in *2020 15th IEEE International Conference on Signal Processing (ICSP)*, vol. 1. IEEE, 2020, pp. 366–370.

[26] V. John and S. 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*. Springer, 2019, pp. 351–364.

[27] K. Kowol, M. Rottmann, S. Bracke, and H. Gottschalk, "Yodar: uncertainty-based sensor fusion for vehicle detection with camera and radar sensors," *arXiv preprint arXiv:2010.03320*, 2020.

[28] Y. Song, Z. Xie, X. Wang, and Y. Zou, "Ms-yolo: Object detection based on yolov5 optimized fusion millimeter-wave radar and machine vision," *IEEE Sensors Journal*, vol. 22, no. 15, pp. 15 435–15 447, 2022.

[29] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 259–16 268.

[30] S. Yao, R. Guan, Z. Wu, Y. Ni, Z. Zhang, Z. Huang, X. Zhu, Y. Yue, Y. Yue, H. Seo *et al.*, "Waterscenes: A multi-task 4d radar-camera fusion dataset and benchmark for autonomous driving on water surfaces," *arXiv preprint arXiv:2307.06505*, 2023.