Sensors and Materials, Vol. 36, No. 8 (2024) 3351–3365
MYU Tokyo

S & M 3736

LinkedFormer: Radar Communication and Multiscale Imaging for Object Detection under Complex Sea Background

Xing Pu,[1] Xisheng Xu,[2] and Yi Yu[2*]

[1]State Radio Monitoring Center, No. 80 Beilishi Rd, Xicheng District, Beijing 100037, P.R. China
[2]Huaxin Consulting Co., Ltd., China, 999 Chunbo Road, Binjiang District, Hangzhou, Zhejiang, China

(Received April 2, 2024; accepted May 21, 2024)

Keywords: underwater autonomous communication, object detection, multisensor fusion techniques, 4D radar

The advent of deep learning has propelled significant advancements in object detection, thereby enhancing the intelligence of underwater autonomous driving systems. In this paper, we explore the cutting-edge applications of autonomous driving technology in the field of underwater exploration, addressing the pivotal role of target detection in navigating and executing tasks within challenging marine environments. In this study, the object detection capability of such systems is enhanced by integrating deep learning and multisensor fusion technology, especially by combining high-precision sensor data with multitask learning models to achieve efficient and robust detection. Our study has three principal contributions. First, we introduce a novel light perception detection system that combines monocular camera technology with 4D radar. It enriches environmental perception by weaving in radar signals and significantly enhances the accuracy and stability of target detection. Second, we have developed a dual-modal detection framework, named Radar-Picture Detection, which utilizes a parallel sequence prediction method. This approach prioritizes radar signal processing, aiding in the improvement of target detection accuracy in intricate underwater environments. Third, we conducted a comprehensive evaluation of our model's performance using the FloW Dataset, which is specifically curated for identifying floating waste in inland waters through unmanned vessel footage. We not only propel forward the field of target detection for underwater autonomous systems but also establish new avenues and a solid foundation for deploying deep learning and multisensor fusion technology in marine environmental perception. Insights and methodologies from this study are poised to spearhead further developments in autonomous marine exploration, enhancing safety, efficiency, and our understanding of underwater environments.

1. Introduction

With the rapid advancement of autonomous driving technology, underwater autonomous driving systems(1) are playing an increasingly vital role in enhancing the efficiency and safety of ocean exploration. Specifically, object detection technology is crucial for achieving high-

[*]Corresponding author: e-mail: yuyi.hx@chinaccs.cn
https://doi.org/10.18494/SAM5062

precision underwater navigation and task execution. In recent years, significant progress has been made in the field of object detection through deep learning, substantially advancing the intelligence of autonomous driving systems. However, the unique challenges of the underwater environment, such as light attenuation, underwater clutter, limited visibility, and complex and variable marine weather conditions, including strong winds, huge waves, and storms, pose unprecedented challenges to the perception capabilities of underwater autonomous driving systems. Addressing these challenges, in this study, we aim to enhance the object detection capabilities of underwater autonomous driving systems through deep learning and multisensor fusion techniques.(2) Given the unstable performance of visual models under adverse conditions and the benefits of multitask learning in improving system performance, we believe that integrating high-precision sensor information with multitask learning models is key to achieving efficient and robust underwater object detection.

We used the FloW Dataset in the experiment, which is uniquely curated for identifying floating waste in inland waters and serves as a pioneering resource for developing methodologies to detect and mitigate floating waste effectively. It offers a comprehensive compilation of 2000 images, synchronized radar point cloud data, and inertial measurement unit (IMU) attitude information, making it an invaluable asset for enhancing the detection of small targets under varied environmental conditions.

The main contributions of this study are as follows: (1) We have proposed an innovative light perception detection system based on the fusion of monocular cameras and a 4D radar. This system not only overcomes the limitations of single visual models in complex underwater environments, but also, by integrating radar signals, provides environmental information, significantly improving the accuracy and robustness of object detection. (2) We have developed a dual-modal detection framework based on transformers, named Radar-Picture Detection, which employs a parallel sequence prediction method to prioritize the processing of targets in radar signals, thereby achieving more accurate object detection in complex underwater environments. (3) We have conducted extensive experiments on the FloW-Dataset, comparing our method with several advanced object detection methods, and the results validate the effectiveness and superiority of our approach.

Through the application of these innovative technologies, we not only advance research in the field of object detection for underwater autonomous driving systems but also provide new directions and practical foundations for the application of deep learning and multisensor fusion technology in underwater environmental perception.

2. Related Work

2.1 Object detection in deep learning

The application of deep learning in maritime complex environment search and rescue operations, particularly in the field of target detection for autonomous ships(3) and unmanned surface vehicles (USVs),(4) has seen notable advances in recent years. The development of technology has primarily focused on enhancing the accuracy, speed, and robustness of target detection to address the variability and uncertainty of maritime environments. Against this backdrop, YOLO (You Only Look Once) v5, convolutional neural networks (CNNs), and Transformer technologies have played pivotal roles in autonomous driving systems(5) for maritime search and rescue tasks.

The YOLO series, a landmark in the field of target detection, with its latest iteration, YOLOv5, stands out for its exceptional speed and accuracy in complex maritime rescue operations. YOLOv5 enhances target detection efficiency and performance through improvements in network architecture, training methods, and anchor mechanisms. In maritime environments, YOLOv5 can swiftly identify and locate potential rescue targets such as ships, personnel, and floating objects, maintaining high recognition accuracy even under dim lighting or poor visibility conditions. This feature makes it a preferred technology for autonomous driving systems in maritime search and rescue tasks. CNNs represent one of the most fundamental and widely applied technologies in deep learning, especially in image recognition and processing. In maritime search and rescue autonomous driving systems, CNNs are utilized to extract complex features from marine environments, such as sea waves and surface reflections.(6) Through their deep network structures, CNNs automatically learn hierarchical features from images, facilitating efficient target detection and classification. This is crucial for distinguishing rescue targets from natural marine features, thereby enhancing the accuracy and efficiency of search and rescue operations. Initially proposed and applied in the field of natural language processing (NLP), Transformer models and their variants have been successfully adapted for image recognition and target detection tasks, such as vision transformer (ViT).(7) Compared with traditional CNNs, Transformers are better suited to handle sequential data and provide global contextual information, which is particularly important for target detection tasks in complex maritime scenes. Within maritime search and rescue autonomous driving systems, Transformers can assist models in capturing long-distance dependences(8) in sea surface images, allowing for the more accurate identification and location of small targets at a distance or targets within complex backgrounds.

In practice, combining YOLOv5, CNNs, and Transformers can further improve target detection performance in complex maritime environments. For example, CNNs can be used for feature extraction, YOLOv5 for rapid target detection, and Transformers for recognizing and classifying targets in complex scenes. This multimodal fusion approach fully leverages the strengths of each model, enhancing the overall performance of maritime search and rescue autonomous driving systems. With the continuous development and optimization of deep learning technology, future autonomous driving systems for complex maritime environment search and rescue operations will become more intelligent, efficient, and accurate, effectively increasing the success rate and safety of search and rescue missions.

2.2 Light Detection and Ranging (LiDAR)-camera 3D detection

In the context of deep learning and autonomous driving technologies, especially in addressing complex maritime search and rescue operations, the fusion technology of LiDAR(9) and camera for 3D perception(10) has achieved significant advancements in recent years. This technology combines the high-precision distance measurement capabilities of LiDAR with the rich color and texture information provided by cameras, offering a more comprehensive and precise environmental perception capability for autonomous driving systems. Notably, the evolution of deep learning has markedly enhanced the accuracy of target detection, classification, and tracking through multimodal fusion techniques,(11) particularly the integration of LiDAR and camera data. Researchers have developed various methods to integrate these two types of data, including early fusion, late fusion, and feature-based fusion. Early fusion typically combines different data sources at the initial stage of sensor data processing, whereas late fusion combines the independently processed results at the decision-making level. Feature-based fusion focuses on how to effectively integrate features from different sensors within deep learning models.

Furthermore, in the case of enhancing real-time processing capabilities to meet the immediate needs of maritime search and rescue operations, researchers are dedicated to accelerating the processing of LiDAR and camera data fusion. By optimizing deep learning models and algorithms, such as employing more efficient CNN architectures and computation acceleration technologies, complex multimodal data can now be processed more rapidly, achieving real-time or near-real-time target detection and tracking.

Concurrently, the development of end-to-end learning models has attracted significant attention in the LiDAR-camera fusion domain in recent years. Such models learn the mapping required for executing specific tasks (e.g., target detection or semantic segmentation) directly from the raw LiDAR point clouds and image data, eliminating the need for manually designed feature extraction steps. This approach enhances the generalization capability of the model and simplifies the training process.

Lastly, in terms of enhancing environmental understanding capabilities, deep learning models, through the fusion technology of LiDAR and cameras, are not only capable of detecting and recognizing specific objects but also understanding the 3D structure of the entire marine environment. This is crucial for autonomous navigation, obstacle avoidance, and executing complex search and rescue operations. Models are now capable of distinguishing between the sea surface, ships, buoys, and other potential rescue targets, even maintaining high accuracy under complex sea conditions and adverse lighting.

Despite significant progress, the fusion of LiDAR and cameras in complex maritime search and rescue environments still faces challenges, including how to effectively process a large volume of multimodal data, enhance robustness in extreme weather conditions, and optimize resource consumption. Future research may focus on developing more advanced fusion algorithms, improving system reliability and adaptability.

3. Methods

3.1 System overview

In this section, we outline the system architecture comprehensively, covering data collection operations and model implementation details. Figure 1 illustrates a conceptual narrative diagram of the system. The system is equipped with a single-chip 77 GHz 4D millimeter-wave radar,

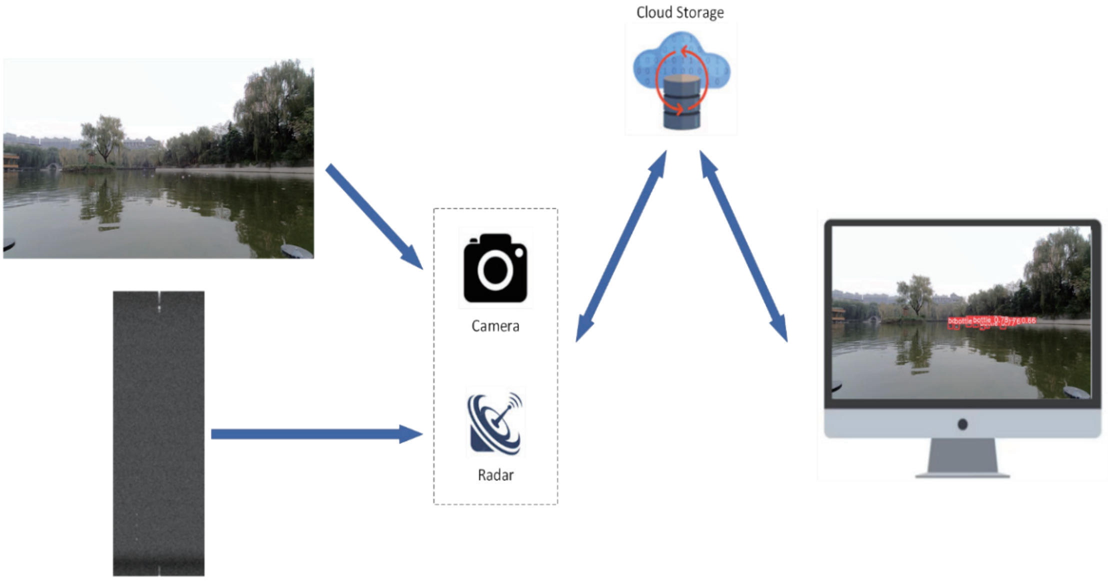

Fig. 1. (Color online) Overall structure of the system. Initially, both cameras and radar devices collect information. Subsequently, the collected data is uploaded to the cloud storage. Finally, the computer retrieves the information from the cloud storage for further processing.

featuring a three-transmitting (TX) and four-receiving (RX) antenna array, supporting a maximum bandwidth of 4 GHz. It directly generates a 4D radar point cloud and transmits data through the Ethernet port. Additionally, the system integrates a camera with a 12 Hz capture frequency and automatic exposure. The camera's native resolution is 1600 × 1200, with an option to crop to a 1600 × 900 ROI to reduce processing time and bandwidth. Subsequently, the robot uploads captured images and radar data to the cloud, enabling remote professionals to access high-quality data for further processing in real time, regardless of their geographical location.

Moving forward, we explore the neural network employed for object detection. By utilizing radar and imagery data from the cloud as inputs, we fuse information from both sensors to classify and localize objects within the 2D bounding box.

3.2 Radar-Picture Detection

We employ a dual-backbone structure for Radar-Picture Detection to capture both visual images and radar features. Adverse weather conditions such as rain or fog may invalidate collected images, necessitating the involvement of the radar in the sensing process. Since the radar cannot directly conduct visual detection, we convert the 3D radar coordinate system into a 2D camera image. An encoder tailored for radar target detection must efficiently extract visual features from each instance in a radar 2D image. Traditional convolutional methods are unsuitable for this task owing to the sparse and small size of radar point clouds, leading to numerous ineffective operations. To mitigate this issue, we incorporate LinkedFormer from the Achelous framework, aiding in capturing point cloud features while minimizing feature loss.(12)

Although both single-stage and two-stage detectors have demonstrated remarkable performance in recent years, the latter, involving box generation and classification stages, requires more computational resources and time. Therefore, we select the YOLOv5 framework for target detection, utilizing its backbone to extract image features (see Sect. 3.3 for detailed information).

The Radar-Picture Detection system processes a sequence of pictures $X = \{x_i\}_{i=1}$ ($x_i \in R^{H\times W\times C}$) and radar queries $R = \{r_i\}_{i=1}$, where r_i corresponds to the i radar picture. An image encoder generates a feature map $f_x \in R^{H\times W\times C}$ for the input detection image, whereas a radar encoder produces a radar embedding vector $f_x \in R^{E\times V_R}$ for the radar data. The shared dimension E is obtained through the linear projection of image and radar features. Subsequently, multimodal sequences S are created by flattening the features of each image and concatenating them with radar embeddings, resulting in sequences with a shape of $(H \times W + E) \times V$. Finally, the multimodal sequence and a set of M_p instance sequences are transmitted to the Transformer in parallel. The Transformer architecture chosen is based on the design used in DETR.[13]

The diagram representing our decoder process is depicted in Fig. 2. Initially, the updated multimodal sequence F_U is generated by the final Transformer encoder layer. F_U is then extracted and reshaped in both the radar and picture-related components. Subsequently, the output $F_L^{1,\ldots,n-1}$ from the first $n - 1$ layers of the image encoder is obtained. This $F_L^{1,\ldots,n-1}$ is hierarchically fused with F_U through a picture decoder D_{det} similar to Feature Pyramid Networks. Throughout the fusion process, the model produces semantically rich image feature maps, denoted as F_{det} in Eq. (1).

$$F_{det} = \left\{ f_{det}^{i} \right\}_{i=1}, \; f_{det}^{i} \in R^{V\times\frac{H}{4}\times\frac{W}{4}} \tag{1}$$

3.3 YOLOv5-based detectors

YOLOv5, introduced in 2020, represents a breakthrough in single-stage object detection. Diverging from its predecessors, YOLOv5 integrates a plethora of novel design and optimization strategies tailored to enhance detection accuracy and efficiency. Its hallmark attributes encompass efficiency, accuracy, and user-friendliness. The architecture of the YOLOv5 target detector is composed of a backbone network and a detection head. To enrich the input data, YOLOv5 employs Mosaic data enhancement, randomly selecting coordinates (x, y) for picture splicing and resizing four pictures accordingly, which are subsequently positioned within a larger image. This augmentation not only enriches the dataset but also bolsters model robustness and improves small object detection by introducing various small objects and semantic information. During training, YOLOv5 utilizes adaptive anchor box calculation functions to dynamically adjust anchors on the basis of different datasets. The Focus module in YOLOv5 slices the input image before it enters the Backbone, resulting in four complementary images with concentrated width and height information in the channel space. By quadrupling the input channels, this process significantly enhances computational efficiency without compromising information integrity. YOLOv5 adopts CSPDarknet53 as its backbone network, a modified

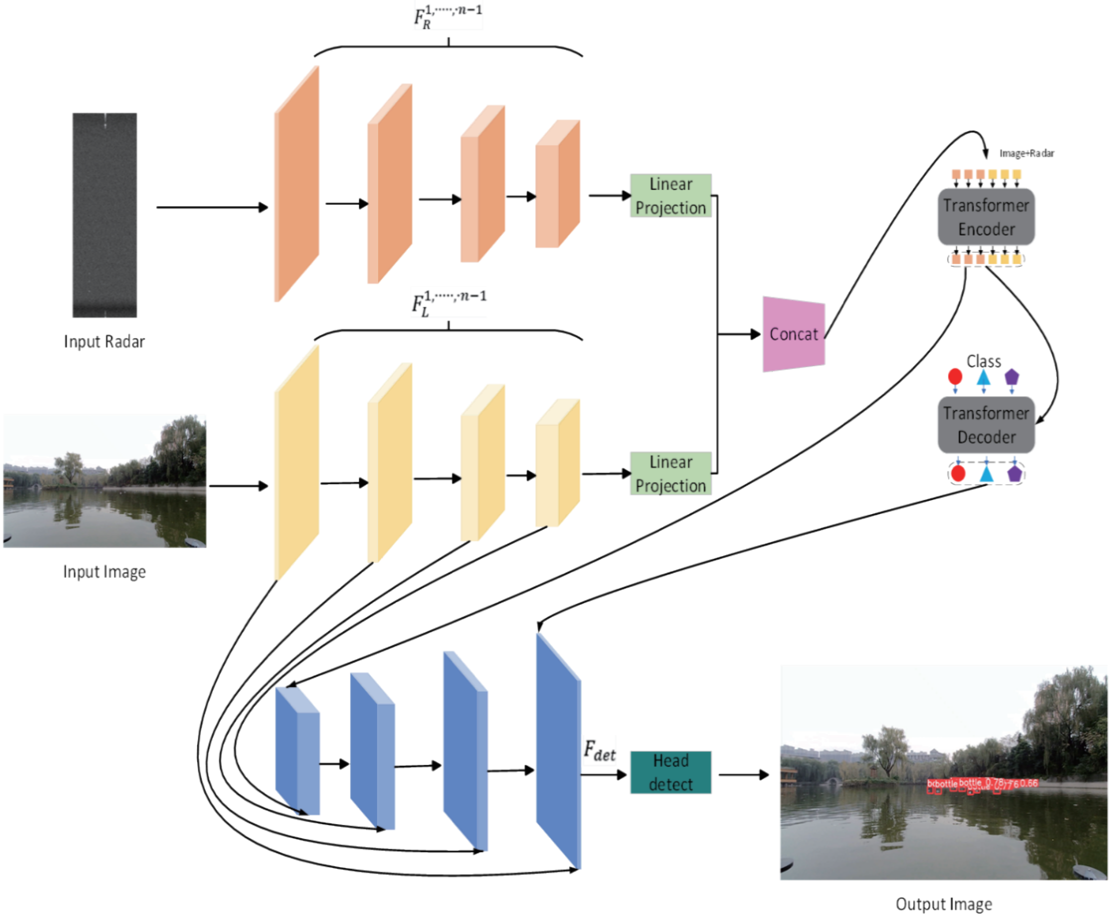

Fig. 2. (Color online) Detailed overview of Radar-Picture Detection.

version of Darknet53 featuring Cross Stage Partial connections to facilitate efficient feature fusion and parameter reduction. The detection head of YOLOv5 consists of convolutional layers and activation functions tasked with generating prediction boxes and category confidence scores. In contrast to the traditional YOLO series, YOLOv5 employs a single-scale prediction strategy, boosting both accuracy and speed by performing target detection on different scales. Additionally, YOLOv5 introduces a novel loss function that leverages information from multiscale target detection to further enhance performance. Demonstrating significant advancements in both accuracy and speed, YOLOv5 has emerged as a premier solution in target detection. With its lightweight network structure and concise code implementation, YOLOv5 is well suited for deployment in resource-constrained environments, such as mobile devices. Furthermore, it offers pretrained models and clear code implementation, streamlining rapid application and customized development for researchers and developers.

4. Experiments and Discussion

4.1 Dataset

In our study, we evaluated the performance of our Radar-Picture Detection model using the FloW Dataset, gathered in a real urban environment. Curated by Ouka Smart Hublot, the FloW Dataset is the first of its kind dedicated to detecting floating waste in inland waters. This dataset demonstrates the identification of floating waste captured through unmanned vessel footage in authentic river settings, pioneering effective methodologies for detecting and mitigating floating waste in inland waters. The dataset comprises 2000 images with a resolution of 1280 × 720, along with synchronized radar point cloud information acquired under varying lighting and wave conditions. Additionally, an IMU records attitude information at a frequency of 10 Hz, while the millimeter-wave radar operates at 77 GHz using Frequency Modulated Continuous Wave radar technology. To ensure data synchronization, readings from different sensors are temporally aligned. Notably, the dataset includes 5271 labeled instances of floating garbage objects, with more than half measuring smaller than 32 × 32 pixels. Consequently, it serves as a valuable resource for studies focusing on the detection of small targets. Exemplary images from this dataset are provided in Fig. 3.

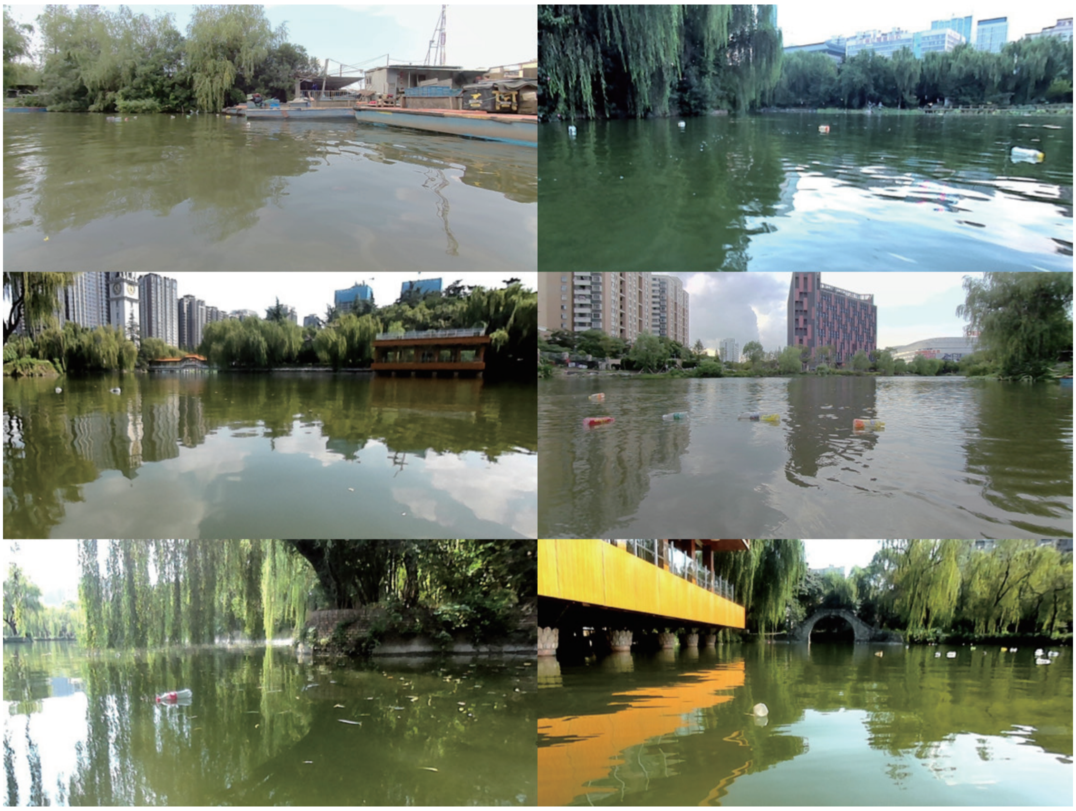

Fig. 3. (Color online) Exemplary images of the FloW Dataset.

4.2 Experimental details

We randomly divided the data samples into training, testing, and validation sets in an 8:1:1 ratio. The experiments were conducted on a PC running Windows 10 equipped with an Intel i9-11900F CPU, GeForce RTX3080 GPU, 10 GB of video memory, and 32 GB of RAM, as outlined in Table 1. To enhance dataset diversity, we applied multiscale data augmentation techniques, including image resizing, translation, color adjustment, and horizontal flipping. These augmentation strategies effectively improve the generalization capabilities of the neural network model. Throughout training, we utilized CUDA 11.3 and PyTorch 1.12 deep learning frameworks. The input image size was set to 640 × 640, the initial learning rate was 0.1, the batch size was 4, the number of epochs was 100, and the momentum of the SGD optimizer was set to 0.937. To expedite training, we employed the pretrained CSPDarknet53 model on the COCO dataset as the fusion branch of the image backbone.

4.3 Evaluation parameters

In this experiment, we conducted a comprehensive comparison of the proposed models while ensuring the consistency of the experimental environment. Our aim is to evaluate the model's impact on detection accuracy, for which we chose Mean Average Precision (*mAP*) and Recall as the primary evaluation metrics.

To assess the model's prediction performance thoroughly, we employed bounding box average precision across various Intersection over Union (*IoU*) thresholds. A higher average precision (*AP*) value for each category indicates a higher bounding box prediction accuracy and fewer missed detections. *mAP* is used to evaluate model accuracy by averaging precision values for each category. True positive (*TP*) prediction boxes, which exhibit an overlap with ground truth objects exceeding the *IoU* threshold, are considered, while false positive (*FP*) objects with less than the *IoU* threshold overlap with ground truth objects. Missed objects are classified as false negatives (*FN*s). The definitions of precision (*P*) and recall (*R*) are provided below.

Table 1
Experimental setup composition.

Composition	Specifications
Operating system	Windows 10
CPU	Intel i9-11900F
GPU	GeForce RTX3080
Video Graphics Memory	10 GB
Memory	32 GB
Computing Architecture	CUDA11.3
Deep Learning Framework	PyTorch1.12
Input Size	640 × 640
Learning Rate	0.1
Batch Size	4
Epoch	100
Optimizer Momentum	0.937

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

AP is determined by *AP* values at different recall points. *mAP* is used to evaluate the model's accuracy by averaging the *AP* values across all categories. Their definitions are

$$AP = \sum_{i=1}^{n-1} \left(R_{i+1} - R_i\right) P\left(R_{i+1}\right), \tag{4}$$

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{M}, \tag{5}$$

where *M* represents the total number of categories in the target detection task. *mAP*s at $IoU = 50$ (mAP_{50}) and $IoU = 35$ (mAP_{35}) are essential metrics for assessing the accuracy of target detection models, as they are used to evaluate model performance across different thresholds. Given the variabilities in the size and shape of targets in this study, using multiple *IoU* thresholds is imperative for accurately evaluating model performance. Selecting *AP*35 and *AP*50 as evaluation metrics helps in determining the effectiveness of target detection algorithms. Higher *AP*35 and *AP*50 values signify more precise target detection and enhanced localization accuracy.

In our experiments, we aimed to highlight the performance of our method by comparing our method with several established target detection techniques. We selected the YOLO series of vision-based approaches owing to their renowned speed in mobile robotics. Thus, we included the latest versions, YOLOv3, YOLOv4, and YOLOv5, as baseline methods. To contrast with single-stage object detection, we also included two-stage methods such as Fast R-CNN and Cascade R-CNN, sacrificing speed for higher accuracy. Additionally, we considered transformer-based approaches, such as Swin Transformer, which have shown state-of-the-art performance in various tasks. For target detection using a millimeter-wave radar, we chose VoteNet and Danzer as benchmarks. Furthermore, we compared our method with fusion-based approaches, including RISFNet, CRF-Net, and Li *et al.*'s method,(16) which combine visual and radar data.

4.4. Comparative experiment

We assessed the performance of the method proposed in this study by comparing the proposed method with other baseline methods. All baseline methods and our model undergo training on the same training set and testing on the same test set. Since different model parameters can impact performance, the parameters of the baseline model were initially set to recommended default values with minor adjustments based on practical considerations during the experiment. To validate the enhancement of detection accuracy facilitated by our proposed fusion method, we contrasted our proposed fusion method with a state-of-the-art bimodal approach using the FLOW dataset. This dataset includes a continuous sequence of simultaneous imagery and millimeter-wave radar data, meticulously annotated for accuracy. The results are detailed in Table 2, where our vision and radar fusion detection method surpasses other fusion detection methods in terms of detection accuracy. Although RISFNet exhibits high detection accuracy, its computational cost is substantial and relies on precise external parameter alignment between the radar and the camera. Inaccurate external parameters significantly diminish RISFNet's performance. Compared with the state-of-the-art LRVFNet method, we achieved mAP_{35} and mAP_{50} scores surpassing 1.29 and 1.71, respectively, demonstrating our method's competitive performance.

Furthermore, we compared our method with other vision-based object detection approaches. As demonstrated in Table 3, our method exhibits enhanced *AP* compared with the sole use of the visual detection method FCOS. Specifically, our method achieves a significant improvement in *AP*, with increases of 21.6 and 23.7% at *IoU* thresholds of 0.5 and 0.35, respectively. This improvement is attributed to the rich sensor information acquired through dual-modal fusion

Table 2
Comparison of methods based on radar and image fusion.

Method	mAP_{35}	mAP_{50}
CRF-Net(14)	79.63	57.74
Jha *et al.*(15)	77.98	–
Li and Xie(16)	85.28	64.64
RISFNet(17)	80.05	75.09
Zhu *et al.*(18)	81.4	–
YOLOv5-l + Zhu *et al.*(18)	81.41	–
Faster-RCNN + Zhu *et al.*(18)	83.63	–
Cascade-RCNN + Zhu *et al.*(18)	79.53	–
Swin-Transformer + Zhu *et al.*(18)	82.42	–
LRVFNet(19)	91.12	78.46
Our method	92.41	80.17

Table 3
Comparison of vision-based methods.

Method	mAP_{35}	mAP_{50}
SSD(20)	–	69.6
Faster RCNN(21)	77.35	57.58
YOLOv3(22)	–	45.3
Swin-Transformer(23)	77.3	–
VFNet(24)	–	55.1
TOOD(25)	–	51.1
YOLOv4(26)	78.46	57.04
EfficientDet(27)	78.62	58.52
FCOS(28)	68.71	58.57
Our method	92.41	80.17

and the implementation of efficient fusion strategies. Additionally, our method outperforms EfficientDet, showcasing performance enhancements of 21.65 and 13.79% at *IoU* thresholds of 0.5 and 0.35, respectively. This superiority arises from our network's enhanced ability to detect small and occluded objects. To offer deeper insights into the performance of our method, qualitative results are presented in Fig. 4. As illustrated in the first row of the figure, our method successfully identifies objects even when they are obscured in the image. Furthermore, as observed in the second row of the figure, our method maintains a high detection threshold even when objects in the image are distant from the observation point.

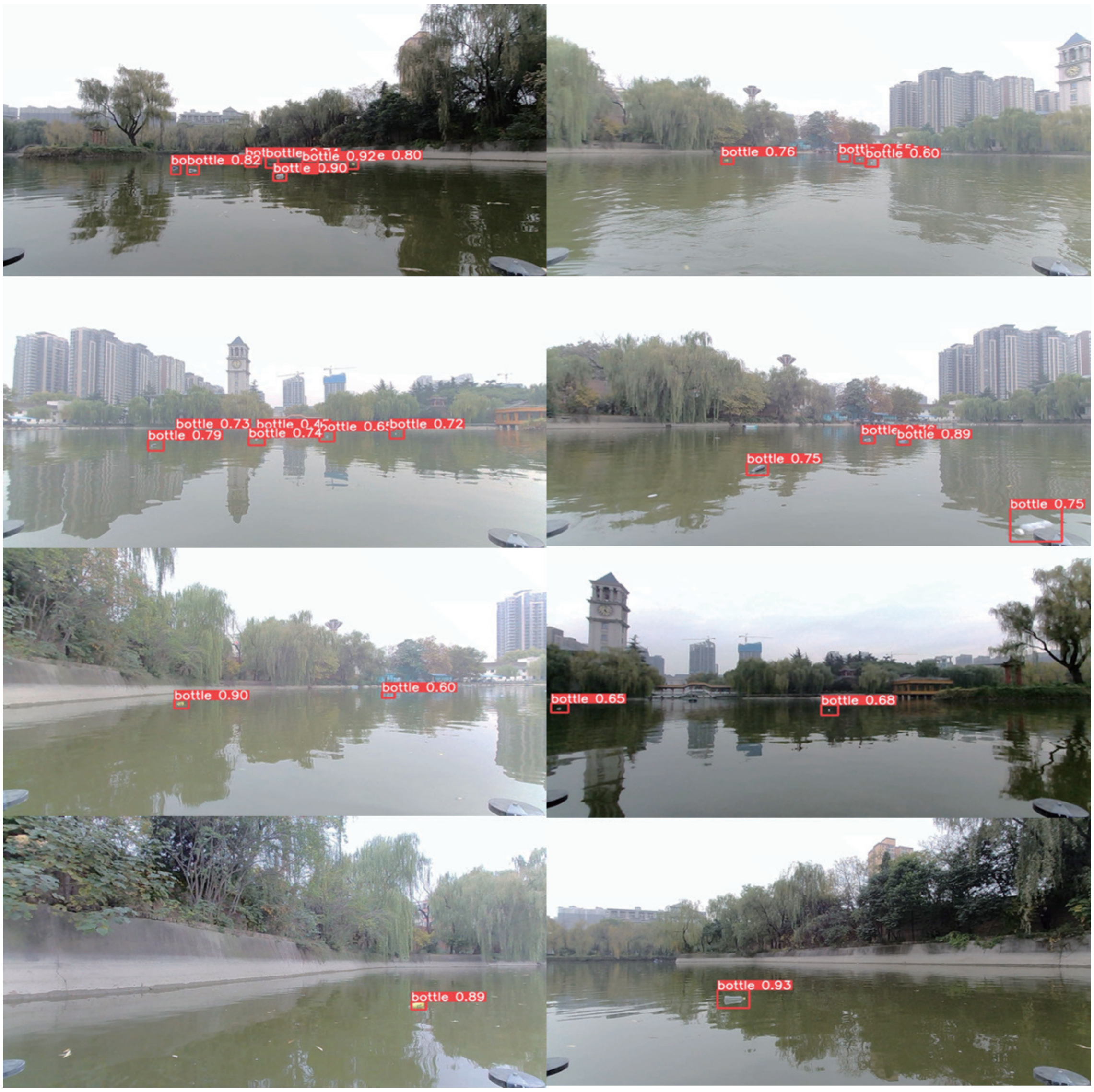

Fig. 4. (Color online) Visualization of Radar-Picture Detection.

4.5. Ablation experiment

To explore the impact of specific radar embeddings on model performance, we trained the model using two additional widely adopted transformer-based encoders: BERT-base[29] and Distill-RoBERTa-base,[30] a distilled variant of RoBERTa.[31] As illustrated in Table 4, our model demonstrates similar performance levels across various transformer-based encoders, highlighting its resilience to changes in encoder selection. However, as expected, employing simpler methods leads to a slight degradation in performance. In addition, other transformer encoders have reduced performance for distant objects.

5. Discussion and Conclusion

Our study represents a significant leap forward in the realm of underwater autonomous driving systems by harnessing the synergistic potential of deep learning and multisensor fusion technologies. Facing the formidable environmental challenges inherent in underwater navigation, such as light attenuation, clutter, limited visibility, and harsh marine weather conditions, we have successfully demonstrated the enhanced capabilities of the Radar-Picture Detection model in conducting precise and robust target detection. The innovative light perception detection system, integrating monocular camera technology with a 4D radar, transcends traditional visual model limitations in complex underwater environments. This enhancement is achieved by weaving radar signals into the fabric of environmental perception, substantially boosting the accuracy and stability of target detection. Furthermore, our dual-modal detection framework, utilizing a parallel sequence prediction method, places a premium on radar signal processing, thereby elevating target detection precision within challenging underwater contexts.

The FloW Dataset, curated specifically for the detection of floating waste in inland waters, serves as a testament to the effectiveness of our model and its superiority over existing advanced target detection methodologies. This dataset, emphasizing the detection of small targets under variable lighting and wave conditions, emerges as an indispensable resource, underscoring the practical applicability and potential of our findings. Through meticulous experimentation and the strategic deployment of these avant-garde technologies, we not only propel the field of target detection in underwater autonomous systems but also establish new avenues and a robust foundation for the application of deep learning and multisensor fusion technology in marine environmental perception.

Table 4
Effectiveness of the transformer.

Method	mAP_{35}	mAP_{50}	Recall
RoBERTa	89.74	75.37	74.36
BERT	89.10	74.89	73.94
Distill-RoBERTa	90.53	77.81	75.67
Our method	92.41	80.17	78.69

The promising results from our carefully structured training and evaluation protocol further corroborate the performance enhancements our method introduces. In particular, the Radar-Picture Detection model has shown notable improvements in *AP*, surpassing existing vision-based object detection methods and showcasing an improved capacity for detecting small and occluded objects, which underscored the critical importance of leveraging rich sensor data and effective fusion strategies to advance object detection capabilities.

The insights and methodologies derived from our research hold the potential to significantly accelerate advancements in autonomous marine exploration, enhancing safety, efficiency, and our comprehension of underwater environments. This study opens up new horizons for future research and development in the field, encouraging the exploration of innovative ideas and directions for the advancement of water surface perception algorithms. The ongoing evolution of technology and methodologies in this domain promises to yield even more sophisticated and capable underwater autonomous driving systems, ultimately contributing to our ability to protect and understand the vast and vital underwater world.

Author Contributions

Conceptualization, P. X. and X. X.; methodology, P. X.; software, Y. Y.; validation, Y. Y. and P. X.; formal analysis, Y. Y.; investigation, Y. Y.; resources, M. H.; data curation, P. X.; writing—original draft preparation, P. X.; writing—review and editing, X. X.; visualization, Y. Y.; supervision, X. X. project administration, X. X.; funding acquisition, X. X. All authors have read and agreed to the published version of the manuscript.

References

1 W. Hou, W. Li, and P. Li: Sensors **23** (2023) 5110.
2 J. Yu, W. Zheng, Y. Chen, Y. Zhang, and R. Huang: Front. Neurosci. **17** (2023) 1219363.
3 E. Nica, G. H. Popescu, M. Poliak, T. Kliestik, and O. M. Sabie: Mathematics **11** (2023) 1981.
4 M. J. Er, C. Ma, T. Liu, and H. Gong: Ocean Eng. **280** (2023) 114562.
5 J. Yu, Y. Xu, H. Chen, and Z. Ju: IEEE Trans. Neural Networks Learn. Syst. **35** (2022) 1.
6 R. Guan, S. Yao, X. Zhu, K. L. Man, E. G. Lim, J. Smith, Y Yue, and Y. Yue: 2023 IEEE 26th Int. Conf. Intelligent Transportation Systems (ITSC) (2023) 182–188.
7 N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko: European Conf. Computer Vision (2020) 213–229.
8 J. Yu, H. Gao, Y. Chen, D. Zhou, J. Liu, and Z. Ju: IEEE Trans. Hum.-Mac. Syst. **52** (2022) 784.
9 Y. Li, L. Zhao, Y. Chen, N. Zhang, H. Fan, and Z. Zhang: Int. J. Appl. Earth Obs. Geoinf. **116** (2023) 103156.
10 Y. Kim, J. Shin, S. Kim, I. J. Lee, J. W. Choi, and D. Kum: Proc. IEEE/CVF Int. Conf. Computer Vision (2023) 17615–17626.
11 A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain: Inf. Fusion **91** (2023) 424.
12 C. Deng, C. Jiang, C. R. Qi, X. Yan, Y. Zhou, L. Guibas, and D. Anguelov: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (2023) 20637–20647.
13 F. Li, A. Zeng, S. Liu, H. Zhang, H. Li, L. Zhang, and L. M. Ni: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (2023) 18558–18567.
14 F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp: Sensor Data Fusion: Trends, Solutions, Applications (SDF) (2019) 1.
15 H. Jha, V. Lodhi, and D. Chakravarty: 6th Int. Conf. Signal Processing and Integrated Networks (SPIN) (2019) 590–593.
16 L. Li and Y. Xie: 15th IEEE Int. Conf. Signal Processing (ICSP) 1 (2020) 366–370.
17 Y. Cheng, H. Xu, and Y. Liu: Proc. IEEE/CVF Int. Conf. Computer Vision (2021) 15263–15272.

18 J. Zhu, Y. Yang, and Y. Cheng: J. Mar. Sci. Eng. **11** (2023) 1794.

19 Y. Xiao, Y. Liu, K. Luan, Y. Cheng, X. Chen, and H. Lu: Remote Sens. **15** (2023) 4433.

20 W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg: 14th European Conf. Computer Vision–ECCV (2016) 21–37.

21 S. Ren, K. He, R. Girshick, and J. Sun: Faster r-cnn: Advances in Neural Information Processing Systems (2015) 28.

22 J. Redmon and A. Farhadi: arXiv:1804.02767 (2018).

23 Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo: Proc. IEEE/CVF Int. Conf. Computer Vision (2021) 10012–10022.

24 H. Zhang, Y. Wang, F. Dayoub, and N. Sünderhauf: Proc. IEEE/CVF Conf. Computer Vision and Pattern recognition (2021) 8514–8523.

25 C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang: IEEE/CVF Int. Conf. Computer Vision (ICCV) (2021) 3490–3499.

26 A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao: arXiv: 2004.10934 (2020).

27 M. Tan, R. Pang, and Q. V. Le: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (2020) 10781–10790.

28 Z. Tian, C. Shen, H. Chen, and T. He: arXiv 2019[J]. arXiv:1904.01355 (2019).

29 J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova: arXiv:1810.04805 (2018).

30 V. Sanh, L. Debut, J. Chaumond, and T. Wolf: arXiv:1910.01108 (2019).

31 J. Yu, H. Gao, D. Zhou, J. Liu, Q. Gao, and Z. Ju: IEEE Trans. Cybern. **52** (2021) 13738.

About the Authors

Xing Pu received his Ph.D. degree in computer application technology through a consecutive master's and doctoral program at Beijing Institute of Technology, China in 2012. He worked as a jointly sponsored doctoral student in the Joint Training Program with Georgia Institute of Technology (Georgia Tech), USA from August 2008 to September 2010. He has been working as a Senior Engineer of the State Radio Monitoring Center since May 2012, focusing on distributed computation, cloud computing, big data, industrial Internet, and radio spectrum management. (puxing@srrc.org.cn)

Xisheng Xu received his master's degree in wireless and communication from Florida Institute of Technology, USA, in 2017. Since 2019, he has been engaged in the design and research of information infrastructure as an engineer, and since 2023, he has participated in a Major Science Program supported by Zhejiang Province government. (xuxisheng.hx@chinaccs.cn)

Yi Yu received his master's degree in communication and information systems from the College of Information Science and Electronic Engineering, Zhejiang University, China in 2004. Since 2004, he has been serving as the Senior Engineer of the Huaxin Consulting Co., Ltd., responsible for the Telecommunications Business Unit. (yuyi.hx@chinaccs.cn)