2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Adaptive Weight Optimization for Ship Detection

Peng Sheng and Ruifu Wang(✉)

School of Information Communication, Dalian University of Technology, Dalian, China
wrf678265@mail.dlut.edu.cn

Abstract. Ship detection is crucial for maintaining maritime sovereignty and monitoring ocean pollution. However, deploying this technology in complex marine environments presents significant challenges, especially when detecting small vessels. Their subtle features, multi-scale variations, and the interference of complex backgrounds often result in poor target localization and classification accuracy in existing models. To address these issues, this paper presents a ship detection model based on multi-modal fusion. The model leverages pre-trained parameters from public datasets to extract features, enhances target identification through a cross-modal synergy mechanism, and introduces an uncertainty loss function to dynamically adjust loss weights, significantly improving detection accuracy across different ship sizes and complex backgrounds. Experimental results on the Levir-Ship dataset, which includes optical remote sensing images, demonstrate the model's effectiveness with AP, AP_{50}, AP_{75}, and AR, scores of 33.7%, 84.8%, 16.1%, and 45.4%, respectively. These results validate the model's superiority in ship detection, offering strong technical support for maritime surveillance and pollution monitoring, and paving the way for future advancements in marine monitoring technologies.

Keywords: Ship detection, deep learning, multi-modal fusion.

1 Introduction

Ships, as the primary carriers of maritime transportation, hold significant importance in territorial security, economic development, and environmental protection. With the advancement of remote sensing technology, the use of optical remote sensing images for continuous and accurate monitoring of small vessels has been applied in fisheries management and military security. However, due to the limitations of remote sensing image resolution, accurately detecting small fishing boats and cargo ships in optical remote sensing images remains a challenging task.

Traditional ship detection primarily relies on algorithms based on image features and physical characteristics. For instance, Miao Kang et al. [1] improved the Faster R-CNN [2] using the traditional Constant False Alarm Rate (CFAR) [3], where the generated target proposals serve as the protection window for the CFAR algorithm to extract small-sized targets. Chonglei Wang et al. [4] proposed an Intensity-Spatial (IS) domain CFAR ship detector, which fuses the intensity of each pixel and the correlation between

pixels into a feature, namely the Intensity-Spatial Index (IS Index), to fully utilize image information. However, these methods exhibit poor detection accuracy and adaptability when faced with complex backgrounds and multi-scale targets. With technological advancements, the emergence of Convolutional Neural Networks (CNNs) [5] has provided new solutions to these problems. Linfeng Jiang [6] proposed the DSFPAP-Net, which increases the number of feature layers and aggregates depth to address the challenge of detecting small targets in low-resolution remote sensing images. Due to the sensitivity of CNNs to noise and their limited performance in complex scenes, Transformers [7] have been introduced into the field of ship detection. Yue Zhou et al. [8] proposed the PVT-SAR ship detection framework, constructing a Pyramid Vision Transformer paradigm to address issues of target density and data insufficiency, thereby enhancing small target detection capabilities. Xiao Ke et al. [9] proposed a method based on Swin-Transformer [10] and FEFPN to overcome the limitations of CNN-based SAR ship detection in complex backgrounds and small target detection. The Swin-Transformer serves as the backbone to model long-range dependencies and generate hierarchical feature maps, while the FEFPN progressively enhances the semantic information of feature maps at various levels, improving the quality of shallow feature maps. In contrast, most of these methods rely on single-modal data, and their inadequacies in information utilization become evident in complex marine environments such as adverse weather and complex lighting conditions. Specifically, single-modal data (e.g., visible light, infrared, or SAR) have inherent limitations under different environmental conditions: visible light images significantly degrade in quality at night or in hazy weather; infrared images are sensitive to temperature and susceptible to environmental heat source interference; SAR images, while unaffected by lighting and weather, are prone to false alarms and missed detections in target-dense areas. Moreover, single-modal methods struggle to fully utilize the complementary information from multi-source data, leading to limited detection performance for small targets, dense targets, and occluded targets in complex scenes. Additionally, traditional methods often rely on handcrafted features or fixed network structures when dealing with multi-scale targets, lacking adaptability to different scenarios, which further restricts their generalization and robustness in practical applications. These issues indicate that there is still considerable room for improvement in the detection capabilities of single-modal methods in complex marine environments.

To address these challenges, this paper proposes a multi-modal model based on transfer learning. The model utilizes the Swin-Transformer as the image backbone network and BERT [11] as the text backbone network to extract and enhance multi-scale image and text features. It employs a feature enhancer that includes deformable self-attention (to enhance image features), ordinary self-attention (to enhance text features), and an image-text cross-attention module for cross-modal feature fusion. In language-guided query selection, a specific module selects features strongly related to the input text from the image features as decoder queries to guide target detection, with the queries containing both content and position parts. The position part is initialized by dynamic anchor boxes, while the content part is learnable during training. In the design of the cross-modal decoder, each cross-modal query sequentially passes through a self-attention layer, an image cross-attention layer, a text cross-attention layer, and a FFN

layer. Compared to the DINO [12] decoder layer, an additional text cross-attention layer is included to inject text information into the query, better achieving modal alignment. The joint application of these three fusion techniques significantly enhances the model's overall performance on existing benchmarks. Furthermore, by optimizing the adaptive loss function, the model's performance is further improved. The main contributions of this study are as follows:

- A multi-modal framework employing transfer learning is proposed. By pre-training the model on public datasets and fine-tuning it on ship datasets, the model's performance is significantly enhanced.
- Cross-modal attention learning is applied to text and images to enhance feature complementarity, and a language-guided method is used to select the most relevant features for the decoder's queries. Additionally, an adaptive loss function is employed to further optimize the model's performance.
- The model is tested on the LEVIR-Ship [13] dataset. By comparing our model with nine baseline models, the effectiveness of our approach is validated, demonstrating its potential in ship detection and its applicability in other fields.

2 Related Work

In the important research field of ship detection, various types of methods have demonstrated their unique characteristics and advantages, while also facing corresponding challenges, providing rich references and directions for future research.

2.1 Ship Detection

In the field of ship detection, traditional methods achieve effective ship detection by combining multiple technologies. Yongli Xu et al. [14] proposed a high-resolution SAR ship detection algorithm that employs a two-stage architecture using a pre-trained Support Vector Machine (SVM) [15] classifier and the Kapur-Sahoo-Wong (KSW) [16] optimal entropy threshold algorithm, which enhances detection speed while maintaining accuracy and effectively suppressing false alarms. Armando Marino et al. [17] utilized polarization data acquisition to detect targets with polarization characteristics different from the sea surface through spatial perturbation analysis.

CNN-based methods have made significant progress in ship detection. Haopeng Zhang et al. [18] proposed a high-resolution feature generator (HRFG) method, adopting a specific network architecture and introducing a background degradation strategy to solve the problem of small ship detection in optical remote sensing images, significantly improving detection accuracy and robustness. Jianwei Li et al. [19] constructed a complex scene dataset and optimized the Faster R-CNN framework in multiple aspects to enhance ship detection performance.

Transformer-related methods have brought new ideas and improvements to ship detection. Runfan Xia et al. [20] combined the advantages of Transformer and CNN, proposing the CRTransSar framework based on contextual joint representation learning to

address the complex features of SAR image targets. Kuoyang Li et al. [21] proposed ESTDNet, using Cascade R-CNN Swin as a benchmark, constructing the FESwin module and AFF module to optimize feature extraction and fusion.

2.2 Transfer Learning

In ship detection-related research, various methods based on transfer learning have emerged. V. Ganesh et al. [22] adopted deep learning methods, using the TensorFlow object detection API and the MASATI-v2 dataset for object detection, considering the video characteristics of real-time satellite monitoring, and using models trained on this dataset for video processing to detect ships. The model training leveraged the SSD MobileNetV2 algorithm and employed transfer learning techniques. Xu Cheng et al. [23] proposed a semi-supervised transfer learning method SAFENESS, utilizing abundant source ship data and limited target ship data, through a data alignment algorithm, two attention mechanisms, and a multi-category adversarial discriminator to train the model, solving problems in traditional sea state estimation methods.

In this research, we innovatively propose a transfer learning-based multi-modal framework for small target ship detection. This model, with its unique multi-modal fusion capability, can effectively integrate optical remote sensing images with other possible modalities (such as infrared images), providing more comprehensive and accurate information for ship detection. In complex marine environments, the model's powerful target localization and recognition accuracy can effectively overcome the issues of insufficient small target detection accuracy and susceptibility to interference in complex backgrounds present in other methods. Moreover, the model employs transfer learning, enabling efficient and stable ship detection on the LEVIR-Ship dataset without the need for extensive retraining. Moreover, its flexible computational performance and task adaptability allow for the adjustment of computational resources according to actual detection needs, ensuring fast detection speeds while meeting detection accuracy requirements. This model is expected to bring new solutions and breakthroughs to the field of ship detection, playing an important role in national security, marine monitoring, and environmental protection.

3 Approach

Currently, the majority of deep learning-based ship detection methods primarily focus on ship localization in high-resolution (HR) remote sensing images, while research targeting medium-resolution (MR) remote sensing images remains relatively scarce. To address this issue, we propose a ship detection model based on MM-Grounding-DINO [24]. The framework of the model is illustrated in Fig. 1. Its core components include a backbone network, a feature enhancer, a language-guided query selection module, and a cross-modal decoder. Initially, the backbone network extracts features from both the input image and the corresponding text. These features are then fed into the feature enhancer for cross-modal feature fusion and enhancement. The language-guided query selection module processes the enhanced features to select appropriate cross-modal

queries from the image features. These queries are subsequently passed into the cross-modal decoder to perform object detection and simultaneously update information from both image and text modalities. Finally, the model outputs the predicted object bounding boxes along with the extracted textual information.

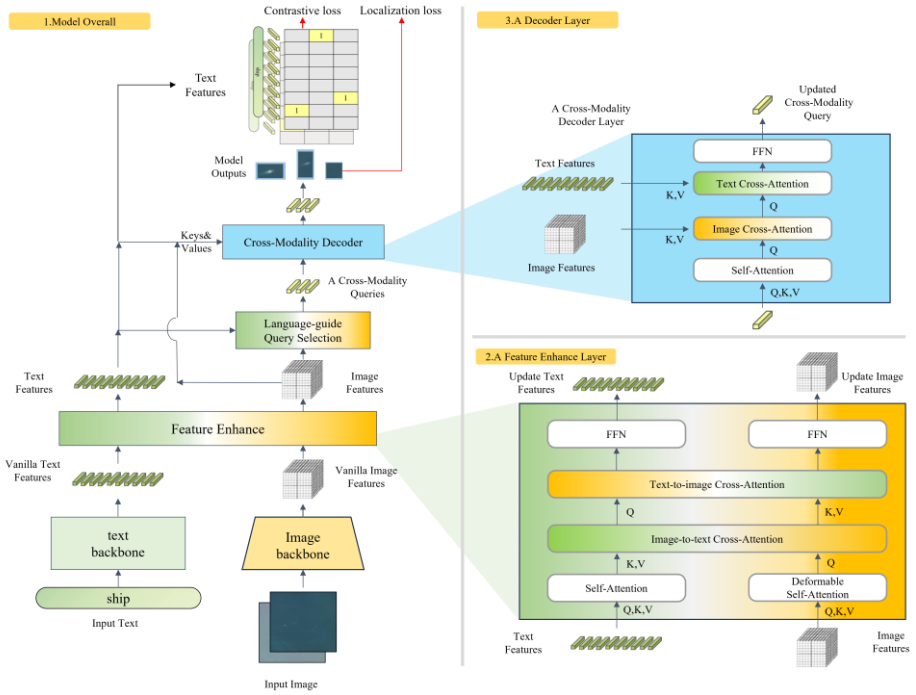

Fig. 1. The framework of our model.

3.1 Model

Feature Extraction and Enhancement. For an input pair comprising an image and its associated text, a visual backbone network (e.g., Swin-Transformer) is utilized to capture multi-scale image representations, while a textual backbone network (e.g., BERT) is applied to extract contextual text embeddings.

$$image_{features} = \text{backbone}_{img}(images) \tag{1}$$

$$text_{features} = \text{backbone}_{text}(text) \tag{2}$$

Next, both sets of features are directed into the feature enhancer module to facilitate cross-modal fusion. Inside the module, the initial integration of text and image features is performed using a Bi-Attention Block, which includes a cross-attention layer for text-to-image interactions and another for image-to-text interactions.

$$image_{fused} = \text{BiAttentionBlock}(image_{features}) \tag{3}$$

$$text_{fused} = \text{BiAttentionBlock}\left(text_{features}\right) \tag{4}$$

Subsequently, the fused text and image features undergo further refinement via a standard self-attention layer and a deformable self-attention layer, followed by a feed-forward network (FFN) layer, ultimately producing the enhanced image and text features.

$$image_{enhanced} = \text{FFN}\left(\text{Self-Attention}\left(image_{fused}\right)\right) \tag{5}$$

$$text_{enhanced} = \text{FFN}\left(\text{Self-Attention}\left(text_{fused}\right)\right) \tag{6}$$

Language-Guided Query Selection. To enhance the role of text in guiding target detection, we introduce a language-guided query selection module. This module dynamically generates or selects queries aligned with the text description by leveraging text features, enabling the model to more effectively focus on image regions relevant to the textual context.

Initially, a matching score matrix is calculated to evaluate the alignment between the image features and the text features.

$$S = image_{enhanced} \cdot text_{enhanced}^{T} \tag{7}$$

$image_{enhanced}$ and $text_{enhanced}$ represent the image and text feature fusion matrices, respectively, after undergoing the feature enhancement and fusion module. S denotes the overall matching score matrix, where S_{ij} corresponds to the matching score between the image feature at position i and the text feature at position j.

For each image, the maximum similarity with all texts is calculated.

$$max\_scores_i = \max\left(S_{i,1}, S_{i,2} \cdots, S_{i,N^T}\right) \tag{8}$$

A vector max_score of length N^T is obtained, where each element represents the maximum similarity between an image token and all text tokens.

Finally, the indices of the num_query image features with the highest matching scores are selected.

$$\text{Selected}_{\text{indices}} = \left\{i_{n-num\_query}, i_{n-num\_query+1}, \cdots i_n\right\} \tag{9}$$

num_query denotes the number of selected queries, which is configured as 900 in this research. The argsort function returns indices sorted in ascending order, and the indices with the highest scores are selected through slicing operations.

Cross-Modality Decoder. The cross-modality decoder layer in our model is designed to further integrate text and image features, enabling effective cross-modal learning.

First, each cross-modal query is fed into a self-attention layer:

$$Q_{\text{self}} = \text{self-attention}(Q) \tag{10}$$

Q represents the input query vector. The self-attention layer is responsible for processing the input queries to enhance the internal feature representation of the queries.

Next, the queries are processed through the image cross-attention layer and the text cross-attention layer separately, facilitating the fusion of image and text features.

$$Q_{image} = \text{cross-attention}(Q_{self}, K_{image}, V_{image}) \tag{11}$$

$$Q_{text} = \text{cross-attention}(Q_{image}, K_{text}, V_{text}) \tag{12}$$

Q_{image} is the output of the image cross-attention layer, while K_{text} and V_{text} are the key and value of the text features, respectively. The primary goal is to integrate text information into the queries to achieve more accurate modality alignment.

Finally, the queries are further processed and optimized through a FFN:

$$Q_{out} = \text{FFN}(Q_{text}) \tag{13}$$

Q_{text} is the output of the text cross-attention layer, and FFN is a feedforward neural network used to further process and refine the final representation of the queries.

3.2 Training Setting

In the training configuration, for the OVD task, all categories from the detection dataset are merged into a single continuous string, such as "Car. Tree. Dog. Bicycle."; for the PG and REC tasks, adhering to M-DETR, each object referenced in the text is annotated during the pre-training stage. For instance, with the description "A man in a red shirt standing beside a bicycle," our model is trained to predict bounding boxes for the man, the red shirt, and the bicycle. Regarding model variations, the case-insensitive pre-trained BERT model is selected as the language encoder, while Swin Transformer serves as the image backbone. For data augmentation, beyond standard techniques like random scaling, cropping, and flipping, random negative samples are introduced by combining categories or descriptions from unrelated images to effectively mitigate the model's hallucination tendency, ensuring it does not predict non-existent objects in the image. We train the model for 100 epochs on 8 NVIDIA 4090 GPUs with a batch size of 64.

3.3 Loss Function

The loss function consists of three main components loss_bbox (bounding box regression loss), loss_cls (classification loss), and loss_iou (IoU loss between predicted boxes and ground truth boxes).

$$\mathcal{L}_{bbox} = \sum_{i=1}^{n} |\mathrm{g}_i - p_i| \tag{14}$$

As shown in Eq. 14 the L1 loss is used for bounding box regression, where g_i represents the ground truth tokenized bounding box, and p_i represents the predicted tokenized bounding box. Following the GLIP [25] model, we employ a contrastive loss

between predicted objects and language tokens for classification. Specifically, we predict the logit value for each text token by computing the dot product between each query and the text features, as shown in the equation, and then apply focal loss [26] to each logit value.

$$\text{logit} = image_{enhanced} \cdot text_{enhanced}^{T} \tag{15}$$

$$p_t = sigmoid(\text{logit}) \tag{16}$$

$$\mathcal{L}_{cls} = -\alpha_t(1-p_t)^{\chi}\log(p_t) \tag{17}$$

As shown in Eq. 17, p_t is the probability of the model predicting the correct class, α is a weighting factor used to balance positive and negative samples, and χ is a focusing parameter used to modulate the weight decay for easy-to-classify samples.

$$\boldsymbol{\mathcal{L}}_{iou} = 1 - IoU + \frac{|C-U|}{|C|} \tag{18}$$

The IoU loss between the predicted and ground truth bounding boxes is computed according to Eq. 18, where C denotes the smallest closed region enclosing both bounding boxes, and U represents the union of their spatial extents.

The traditional multi-task loss function is calculated as shown in Eq. 19, where ω_{bbox}, ω_{cls}, and ω_{iou} are the weights for the bounding box loss, classification loss, and IoU loss, respectively.

$$\mathcal{L}_{total} = \omega_{bbox}\mathcal{L}_{bbox} + \omega_{cls}\mathcal{L}_{cls} + \omega_{iou}\mathcal{L}_{iou} \tag{19}$$

Traditional methods typically rely on manually preset or fixed weight assignments, which cannot dynamically adjust the priorities of different tasks based on task difficulty, changes in data distribution, or training progress. To address this, we introduce the concept of UNCERTAINTY [27] to measure the losses of different tasks and dynamically balance them. In multi-task learning, UNCERTAINTY is used to adjust the loss weights of each task. By modeling uncertainty as learnable parameters, the model automatically adjusts task weights, there by suppressing the influence of complex or noisy tasks and ensuring effective joint training. Specifically, Eq. 20 has three tasks $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ with loss functions $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$, respectively, we adjust their weights by introducing uncertainty parameters σ_1, σ_2, and σ_3. The final optimization objective can be expressed as:

$$\mathcal{L}_{\alpha}(\sigma_1, \sigma_2, \sigma_3) = \sum_{i=1}^{3}\left(\frac{1}{2\sigma_i}\mathcal{L}_i + log\sigma_i\right) \tag{20}$$

4 Experimental results and analysis

4.1 Evaluation Metrics

We evaluate the effectiveness of the proposed method through detection accuracy. We use metrics such as AP (Average Precision), AP_{50}, AP_{75}, and AR, (Average Recall) to measure the model's detection accuracy. The calculation of AP is shown in:

$$AP = \sum_{r\in\{0,0.01,0.02,\cdots,1\}} P(r) \times \triangle r \tag{21}$$

4.2 Dataset

Most ship detection datasets focus on high spatial resolution images, such as 0.3 meters, 0.5 meters, and 2 meters, while datasets targeting small ship objects with low spatial resolution are relatively scarce. In this study, we utilized a small ship detection dataset named Levir-Ship, which has a spatial resolution of 16 meters. The images in Levir-Ship were acquired by the multispectral cameras of the GaoFen-1 and GaoFen-6 satellites, using only the red (R), green (G), and blue (B) bands. The dataset contains 85 scenes, with pixel resolutions ranging between 10,000×10,000 and 50,000×20,000. We cropped the original images to a size of 512×512, resulting in 1,973 positive samples and 1,903 negative samples.

4.3 Comparison

To provide a more comprehensive evaluation of our model's performance, we compared our model with several state-of-the-art object detection models, such as DDOD [28], Dynamic R-CNN [29], TOOD [30], and others. The specific data is presented in the Table 1.

Table 1. COMPARISION OF DIFFERENT METHODS

Model	AP(%)	AP_{50}(%)	AP_{75}(%)	AP_s(%)	AP_m(%)	AR(%)	AR_s(%)	AR_s(%)	AP_{ave}(%)	AR_{ave}(%)
AutoAssign [31]	23.0	73.8	9.9	27.1	25.8	40.7	40.9	27.5	31.9	36.4
DDOD [28]	23.7	73.9	10.3	27.2	27.8	41.4	41.6	21.5	32.6	34.8
Dynamic R-CNN [29]	24.1	69.9	7.6	24.1	25.7	36.8	36.8	32.5	30.3	35.4
TOOD [30]	26.3	72.6	10.4	26.3	28.7	40.0	40.0	36.2	32.9	38.7
FCOS [32]	25.7	71.2	8.7	25.5	33.6	39.8	39.9	38.8	32.9	39.5
ATSS [33]	27.0	73.2	12.0	26.9	33.0	41.3	41.4	36.2	34.4	39.6
FASF [34]	20.0	58.7	5.7	19.8	34.7	35.8	35.8	35.0	27.8	35.5
Grid R-CNN [35]	16.7	57.3	3.4	16.9	9.2	32.2	32.3	25.0	20.7	29.8
Libra R-CNN [36]	23.6	67.0	7.3	23.4	35.8	36.0	36.0	36.2	31.4	36.1
MM-Grouding-DINO [24]	1.0	0.7	0	0.7	0.3	8.0	10.3	32.5	0.5	16.9
Ours	33.7	84.8	16.1	33.5	37.1	45.4	46.0	51.2	41.0	47.5

[1] Results of MM-Grouding-DINO's test using publicly available pre-trained model weights

Our model demonstrates significant superiority over other comparative models across all metrics. Particularly in the detection tasks for small and medium-sized targets, our model exhibits stronger adaptability and higher detection accuracy. Compared to the second-best model, ATSS, in our experiments, our model achieves notable performance improvements in key metrics such as AP, AP_{50}, AP_{75}, and AR, with increases of 6.7%, 11.6%, 4.1%, and 4.1%, respectively. Other metrics also show improvements. These experimental results fully validate the effectiveness and superiority of our model in object detection tasks, especially in complex scenarios and multi-scale target detection, where it performs exceptionally well.

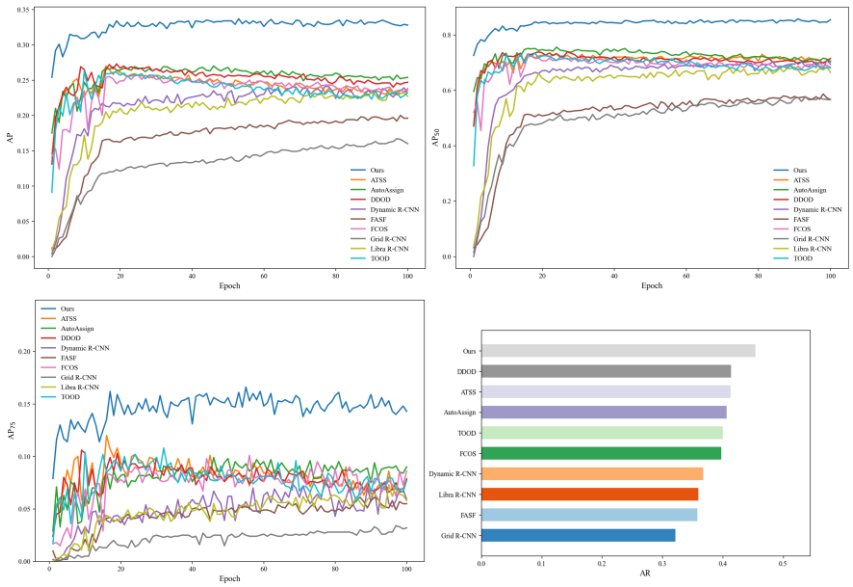

Fig. 2. Performance comparison of object detection algorithms across epochs.

As illustrated in the Fig. 2, in the line charts of AP, AP_{50}, and AP_{75} metrics, our model demonstrates efficient convergence characteristics from the early stages of training. Its curve rapidly ascends to a leading position within a very short number of epochs and maintains stable high-level fluctuations throughout the subsequent training process, significantly outperforming other comparative algorithms. This performance fully proves that our model possesses superior feature learning capabilities and optimization mechanisms, enabling it to quickly capture effective features while maintaining the stability of detection accuracy over prolonged training periods. Further analysis of the AR metric in the bar chart reveals that our model achieves a dominant position in average recall, indicating that it not only ensures high-precision localization but also achieves more comprehensive target recall in object detection tasks.

In conclusion, by conducting extensive comparisons with other leading object detection models, our model demonstrates clear advantages in key metrics such as AP,

AP_{50}, AP_{75}, and AR, fully validating its effectiveness and exceptional performance in object detection tasks. By introducing textual modality to compensate for the shortcomings of image modality in complex backgrounds and small target detection, leveraging cross-modal attention mechanisms to enhance the complementarity of image and text features, and utilizing language-guided query selection mechanisms to precisely locate targets, our model achieves significant improvements. In addition, the optimization of $\mathcal{L}_\alpha$ further enhances the model's adaptability to complex scenarios. These innovative designs enable the model to exhibit higher accuracy and robustness in ship detection tasks using medium-resolution remote sensing images while maintaining high computational efficiency.

4.4 Ablation Experiments

In this section, we systematically evaluate the effectiveness of the proposed tightly-coupled fusion localization model for open-set object detection through ablation experiments. To thoroughly evaluate the impact of each module on the model's performance, we incrementally removed different fusion modules and conducted pre-training and testing on the official dataset using the Swin-T backbone network. The experimental results are presented in the Table 2.

Table 2. ABLATIONS FOR OUR MODULE

Model	AP(%)	AP_{50}(%)	AP_{75}(%)
Ours	33.7	84.8	16.1
w/o $\mathcal{L}_\alpha$	33.0	84.7	14.9
w/o encoder fusion	33.5	84.3	15.1
w/o text cross-attention	26.9	79.0	14.3

The experimental results indicate that the text cross-attention module significantly enhances the model's performance on the Levir-Ship dataset, contributing performance gains of +25% AP, +7% AP_{50}, and +12.6% AP_{75}, respectively. This result strongly confirms the essential role of text cross-attention in cross-modal feature fusion, significantly boosting the model's ability to comprehend target semantic information and, as a result, markedly enhancing detection accuracy. Additionally, $\mathcal{L}_\alpha$ also positively impacts the model's performance, contributing gains of +7% AP, +1% AP_{50}, and +12% AP_{75}, demonstrating its importance in optimizing model prediction confidence and enhancing detection robustness.

In contrast, the encoder fusion module provides relatively limited performance improvements on the Levir-Ship dataset, with gains of only +0.6% AP, +0.5% AP_{50}, and +6.7% AP_{75}. This phenomenon may be related to the characteristics of the dataset. Specifically, the Levir-Ship dataset has limited diversity in target types and scenarios, with relatively simple backgrounds, which means the model can adequately fit the data distribution without relying on complex feature fusion mechanisms such as encoder fusion. Therefore, the model can easily learn patterns in the data through other means (e.g., text cross-attention) without needing encoder fusion to integrate complex feature information.

4.5 Visualization

To more intuitively demonstrate the functionality of each module and its impact on model performance, we conducted a visual analysis of the feature maps and prediction results. Through visualization, we were able to diagnose the feature extraction process of each network layer during model training and gain a deeper understanding of the role of each module based on the performance of the feature maps. Specifically, we extracted feature maps generated by convolutional neural network filters layer by layer, observing how they abstracted and processed data during the training process.

The heatmaps are shown in the Fig. 3. As illustrated in the figure, it is evident that, compared to the models without $\mathcal{L}_\alpha$ (w/o $\mathcal{L}_\alpha$), without encoder fusion (w/o encoder fusion), and without text cross-attention (w/o text cross-attention), our model demonstrates more precise target localization. The red bounding boxes in the heatmaps provide clearer annotations of the targets, effectively distinguishing between target and non-target regions. Besides, the heat distribution in the target regions is more reasonable, indicating that our model can more effectively capture key information. For instance, in scenarios with complex backgrounds or target occlusion, our model still generates high-response heatmap regions, demonstrating its strong capability to capture target features.

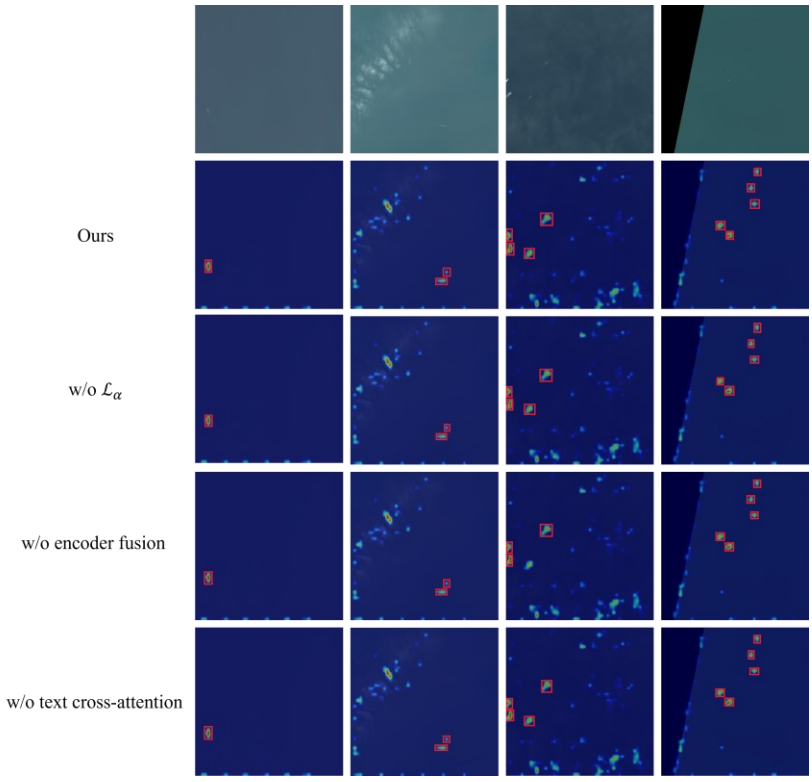

Fig. 3. Heatmaps from ablation experiments of our model.

From the visualization of the prediction results shown in the Fig. 4, it is evident that, compared to the models without $\mathcal{L}_\alpha$ (w/o $\mathcal{L}_\alpha$), without encoder fusion (w/o encoder fusion), and without text cross-attention (w/o text cross-attention), our model demonstrates significant advantages in the completeness and accuracy of target detection. Specifically, our model is able to detect more targets, particularly smaller and more challenging ones, highlighting its robustness in handling multi-scale targets. Moreover, the predicted bounding boxes align more closely with the actual target locations, indicating a notable improvement in localization precision. Furthermore, in scenarios with complex backgrounds or densely packed targets, our model more effectively identifies and distinguishes targets, reducing instances of false positives and missed detections. This further validates its strong adaptability and reliability in complex environments.

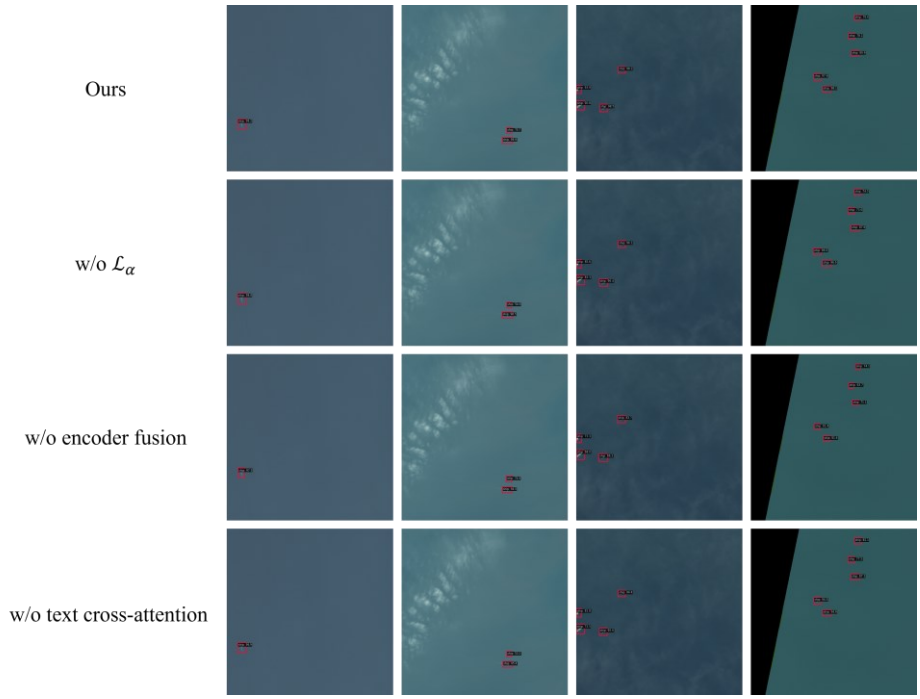

Fig. 4. Comparisons of the detection results by different methods.

To more comprehensively evaluate the model's performance, we introduced the confusion matrix as a critical tool. From the normalized confusion matrix, it is evident that the model achieves a 100% accuracy in predicting background samples, completely avoiding misclassifying the background as ships. For ship samples, the prediction accuracy reaches 83%, demonstrating high reliability in target detection. The overall classification performance is notably superior, with the model accurately distinguishing between ships and the background.

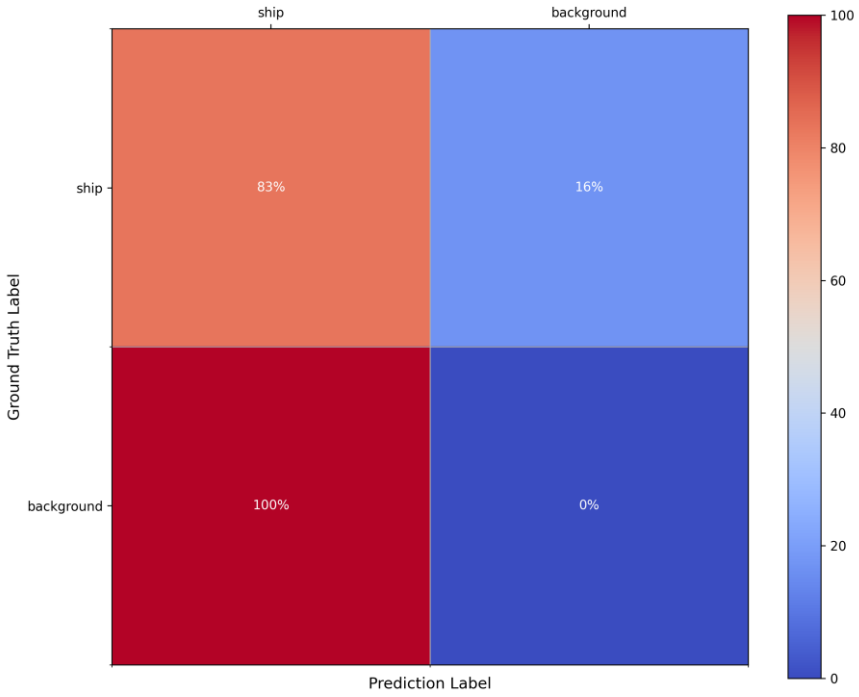

Fig. 5. Confusion matrix for our model.

5 Conclusion and Discussion

This study addresses the challenge of multi-scale ship detection in complex maritime environments by proposing an innovative framework based on multi-modal feature fusion and dynamic loss optimization. By leveraging a cross-modal complementary mechanism to enhance the representation capability of multi-scale targets and incorporating uncertainty modeling to achieve adaptive adjustment of loss weights, the framework significantly improves the detection accuracy of weak-feature targets (e.g., small ships) and robustness under complex background interference. Experimental results demonstrate the technical superiority of the framework in target localization and classification tasks, achieving metrics of AP (33.7%), AP_{50} (84.4%), AP_{75} (16.1%), and AR (45.4%) on the Levir-Ship dataset.

Despite significant advancements in existing ship detection technologies, notable limitations remain in complex maritime scenarios. Traditional ship detection methods rely on handcrafted features, which perform poorly in multi-scale target detection. Deep learning models, due to their single-modal nature, exhibit limited capability in classifying detailed ship attributes. This study integrates textual semantic information with remote sensing images, effectively addressing the shortcomings of pure visual models

in semantic reasoning and providing a new approach for target detection in complex scenarios.

The goal of this study is to enhance ship detection performance through multi-modal feature fusion and dynamic loss optimization. By fully utilizing the complementary information from multiple modalities, the accuracy and applicability of ship recognition are improved. A dynamic loss weighting strategy based on uncertainty modeling is employed to adaptively adjust the optimization weights for multi-scale targets, effectively mitigating training imbalance issues in complex scenarios. Future research will further explore the transferability of this framework to fields such as remote sensing image analysis and ship monitoring, while focusing on optimizing the real-time performance and generalization capability of multi-modal fusion.

References

1. M. Kang, X. Leng, Z. Lin, and K. Ji, "A modified faster r-cnn based on cfar algorithm for sar ship detection," in 2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP). IEEE, 2017, pp. 1–4.
2. R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015.
3. H. M. Finn, "Adaptive detection mode with threshold control as a function of spatially sampled clutter level estimates," 1968. [Online]. Available: https://api.semanticscholar.org/CorpusID:208092087
4. C. Wang, F. Bi, W. Zhang, and L. Chen, "An intensity-space domain cfar method for ship detection in hr sar images," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 4, pp. 529–533, 2017.
5. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural Computation, vol. 1, no. 4, pp. 541–551, 1989.
6. L. Jiang, Y. Li, and T. Bai, "Dsfpap-net: Deeper and stronger feature path aggregation pyramid network for object detection in remote sensing images," IEEE Geoscience and Remote Sensing Letters, 2024.
7. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762
8. Y. Zhou, X. Jiang, G. Xu, X. Yang, X. Liu, and Z. Li, "Pvt-sar: An arbitrarily oriented sar ship detector with pyramid vision transformer," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 291–305, 2022.
9. X. Ke, X. Zhang, T. Zhang, J. Shi, and S. Wei, "Sar ship detection based on swin transformer and feature enhancement feature pyramid network," in IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2022, pp. 2163–2166.
10. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021. [Online]. Available: https://arxiv.org/abs/2103.14030
11. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

12. H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, “Dino: Detr with improved denoising anchor boxes for end-to-end object detection,” 2022.
13. J. Chen, K. Chen, H. Chen, Z. Zou, and Z. Shi, “A degraded reconstruction enhancement-based method for tiny ship detection in remote sensing images with a new large-scale dataset,” IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–14, 2022.
14. Y. Xu, W. Xiong, and J. Liu, “A new ship target detection algorithm based on svm in high resolution sar images,” in Proceedings of the International Conference on Advances in Image Processing, 2017, pp. 6–13.
15. C. Cortes and V. N. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, pp. 273–297, 1995. [Online]. Available: https://api.semanticscholar.org/CorpusID:52874011
16. J. Kapur, P. Sahoo, and A. Wong, “A new method for gray-level picture thresholding using the entropy of the histogram,” Computer Vision, Graphics, and Image Processing, vol. 29, no. 3, pp. 273–285, 1985. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0734189X85901252
17. A. Marino, “A notch filter for ship detection with polarimetric sar data,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 6, no. 3, pp. 1219–1232, 2013.
18. H. Zhang, S. Wen, Z. Wei, and Z. Chen, “High-resolution feature generator for small ship detection in optical remote sensing images,” IEEE Transactions on Geoscience and Remote Sensing, 2024.
19. J. Li, C. Qu, and J. Shao, “Ship detection in sar images based on an improved faster r-cnn,” in 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA). IEEE, 2017, pp. 1–6.
20. R. Xia, J. Chen, Z. Huang, H. Wan, B. Wu, L. Sun, B. Yao, H. Xiang, and M. Xing, “Crtranssar: A visual transformer based on contextual joint representation learning for sar ship detection,” Remote Sensing, vol. 14, no. 6, p. 1488, 2022.
21. K. Li, M. Zhang, M. Xu, R. Tang, L. Wang, and H. Wang, “Ship detection in sar images based on feature enhancement swin transformer and adjacent feature fusion,” Remote Sensing, vol. 14, no. 13, p. 3186, 2022.
22. V. Ganesh, J. Kolluri, A. R. Maada, M. H. Ali, R. Thota, and S. Nyalakonda, “Real-time video processing for ship detection using transfer learning,” in International Conference on Image Processing and Capsule Networks. Springer, 2022, pp. 685–703.
23. X. Cheng, G. Li, R. Skulstad, and H. Zhang, “Safeness: A semi-supervised transfer learning approach for sea state estimation using ship motion data,” IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 5, pp. 3352–3363, 2023.
24. X. Zhao, Y. Chen, S. Xu, X. Li, X. Wang, Y. Li, and H. Huang, “An open and comprehensive pipeline for unified object grounding and detection,” 2024. [Online]. Available: https://arxiv.org/abs/2401.02361
25. L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao, “Grounded language-image pre-training,” 2022. [Online]. Available: https://arxiv.org/abs/2112.03857
26. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007.
27. Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” 2016. [Online]. Available: https://arxiv.org/abs/1506.02142
28. Z. Chen, C. Yang, Q. Li, F. Zhao, Z.-J. Zha, and F. Wu, “Disentangle your dense object detector,” in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4939–4948.

29. H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic R-CNN: Towards high quality object detection via dynamic training," arXiv preprint arXiv:2004.06002, 2020.
30. C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in ICCV, 2021.
31. B. Zhu, J. Wang, Z. Jiang, F. Zong, S. Liu, Z. Li, and J. Sun, "Autoassign: Differentiable label assignment for dense object detection," arXiv preprint arXiv:2007.03496, 2020.
32. Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," arXiv preprint arXiv:1904.01355, 2019.
33. S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," arXiv preprint arXiv:1912.02424, 2019.
34. C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 840–849.
35. X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid r-cnn," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
36. J. Pang, K. Chen, Q. Li, Z. Xu, H. Feng, J. Shi, W. Ouyang, and D. Lin, "Towards balanced learning for instance recognition," International Journal of Computer Vision, vol. 129, no. 5, pp. 1376–1393, 2021.