IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10641146

YO-DETR: A LIGHTWEIGHT END-TO-END SAR SHIP DETECTOR USING DECODER HEAD WITHOUT NMS

Yue Guo[1], Tianzhu Zhang[1], Shiqi Chen[2], Ronghui Zhan[1], Luzhuo Li[3], Jun Zhang[1]*
[1] National Key Laboratory of Automatic Target Recognition,
National University of Defense Technology, Changsha, China
[2] College of Information and Communication,
National University of Defense Technology, Wuhan, China
[3]Beijing University of Technology, Beijing, China
* zhanrh@nudt.edu.cn

ABSTRACT

A lightweight SAR ship detection algorithm is necessary to further meet the demands of military applications. This paper proposes an end-to-end efficient SAR ship target detection algorithm based on RT-DETR called YO-DETR. In the feature extraction network, a CNN-based backbone network is used to replace the transformer-based encoder structure, which retains the original feature extraction capability while reducing the number of parameters. Additionally, in order to retain the characteristic of long-distance feature dependency in transformers, the IRMB module is incorporated into the CNN network to enhance long-distance feature interaction. Finally, the introduction of the decoder head reduces the additional time overhead of traditional NMS during inference. Ultimately, the YO-DETR method achieves a 98.2% mAP on the SSDD dataset with only 5.37M parameters and 10.5M weight size, while the FPS (when batch size is set to 32) also achieves 220.4.

Index Terms—SAR, Ship Detection, RT-DETR, Lightweight, End-to-End

1. INTRODUCTION

In recent years, muti-scale SAR ship detection based on CNN has developed rapidly. H Zhu et al. [1] designed the Duplicate Bilateral Feature Pyramid Network (DB-FPN) which enhances the fusion of spatial information and semantic information to improve the detection performance of multi-scale ships. W Zhao et al. [2] proposed a novel ship detection base on yolov5. Which use CBAM, RFB, and ASFF modules, to enhance the extraction of target features. Z Sun et al. [3] designed bi-directional feature fusion module (Bi-DFFM) to improve the YOLO's Neck structure. Bi-DFFM employs both top-down and bottom-up path for features extraction to improve detection ability of multi-scale ships. Y Guo et al. [4] proposes the depthwise adaptively spatial feature fusion (DSASFF) module. It fuses features from muti-scale and improve the detection performance of multi-scale targets while reducing the computational effort of the model. Z Chen et al. [5] improved yolov7 by SAS-FPN module. Which uses the SA module and the ASPP module to enhance the extraction of target features. X Ren et al. [6] proposed the effective multi-scale feature fusion network (MFFNet) to obtain positional and semantic information, which improves the detection performance of multi-scale targets.

In addition, as transformer shows excellent performance in various fields, more and more researchers apply transformer in SAR ship detection. K Li et al. [7] designed a new backbone network called FESwin by combining swim transformer and CNN networks, which makes the network capable of global context as well as local feature information extraction. K Feng et al. [8] introduces a EfficientViT block in yolov5's backbone, which improves the network's ability to extract multiscale features while keeping the model lightweight. M Zha et al. [9] proposed the Local Enhancement and Transformer (LET) Module, which applies the part Conformer module to enhance the extraction of target features. N Yu et al. [10] designed a novel lightweight radar ship detector that use multiple hybrid attention mechanisms to obtain high detection precision while achieving fast inference for SAR ships detection. H Shi et al. [11] used the deformable attention mechanism to improve the original attention mechanism of swim transformer, which improves the detection accuracy of small ships. W Zhou et al. [12] introduces transformer as a spatial attention mechanism, which enables the model to better extract global information and improves the detection performance of the model.

To combine the advantages of the CNN and Transformer structures, this paper proposes a lightweight end-to-end SAR ship detection algorithm called YO-DETR. During the prediction phase, a transformer-based decoder head is utilized to achieve accurate and efficient prediction of target boxes without using Non-Maximum Suppression (NMS).

2. METHOD

2.1. Overall Structure

The overall flowchart of YO-DETR is illustrated in Figure 1. The YO-DETR can be divided into three main parts: Backbone, Neck, and Prediction. In the backbone, a feature pyramid structure is employed to perform feature extraction tasks. Additionally, the traditional convolutional modules are

enhanced with IRMB modules to improve the dynamic long-range feature interaction capability. The Neck section utilizes a PAN (Path Aggregation Network) structure to facilitate feature transfer and fusion. During the prediction phase, a transformer-based Decoder Head is employed to efficiently obtain the anchors for the target detection.

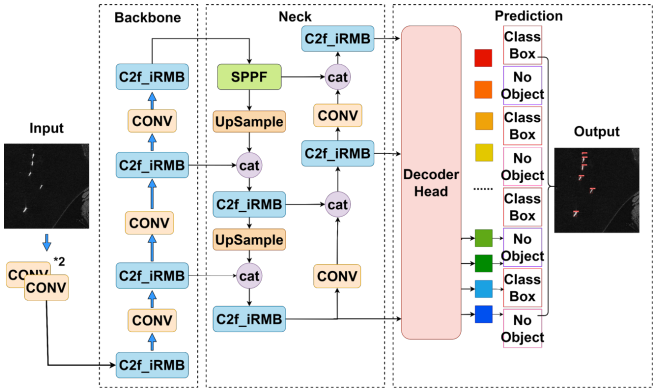

Figure 1. The network structure of YO-DETR.

2.2. IRMB Module

The Inverted Residual Mobile Block (IRMB) [13] module is a flexible and efficient inverted residual module. The structure of IRMB module is shown in Figure 2 and it rethinks the lightweight feature brought by the IRB structure of lightweight networks such as MobieNetv2. Additionally, it considers the dynamic modelling capability of transformer, and applies the IRB structure of CNN-based networks to the attention model. Meanwhile, for mobile applications, the lightweight MHSA structure of EW-MHSA is proposed. Specifically, the IRMB module uses DW-Conv and EW-MHSA cascade to replace the previous Meta-Former's efficient operator, which makes the model take into account the learning ability of local features as well as long-distance interactions, thus effectively improve the model's receptive field and enhance the model's detection performance. The introduction of self-attention modules also enhances the ability to extract features from multi-scale objects.

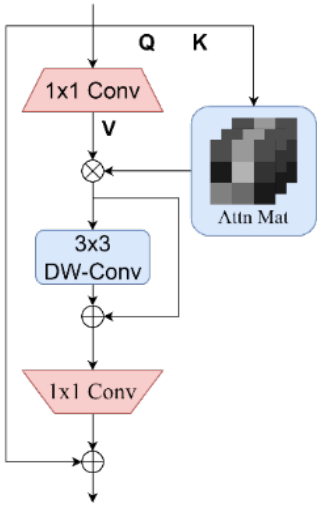

Figure 2. The structure of IRMB module.

2.3. Decoder Head

The YOLO series models use time-consuming NMS operations in post-processing, which affects the performance of the models in real-time detection. To address this problem, RT-DETR [14] uses an improved DINO decoder, which does not require NMS post-processing for prediction and only picks the final Top-k predictions from the output of the encoder module. Firstly, in order to improve the convergence speed of the model, the decoding part of RT-DETR inherits the idea of Contrastive Denoising Training (CDT) from DINO to improve the quality of matched samples, avoiding the repeated output of the model to the same target. Furthermore, DINO enhances the forward structure of Deformable DETR to enable the passing of gradient information from the next layer to the current layer, thus promoting the optimization of the current layer's parameters.

Meanwhile, in order to ensure that the category prediction is aligned with the position prediction, the decoder of RT-DETR improves the category prediction method of DINO by using IoU-aware as the category prediction label to calculate the category loss.

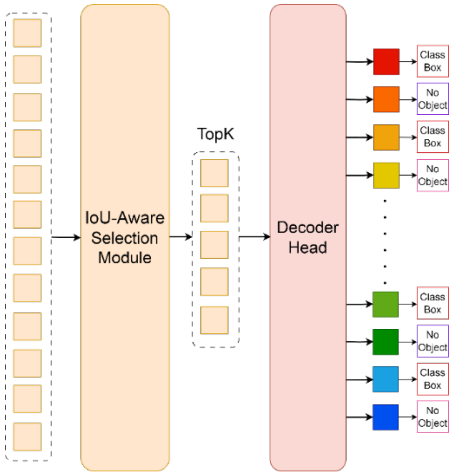

Figure 3. The network structure of Decoder Head.

3. EXPERIMENTS

3.1. Datasets and Details

In order to verify the effectiveness of the proposed algorithm, a series of experiments were conducted on the SSDD dataset. The SSDD dataset consists of 1,160 scene images collected by RaderSat-2, Terra-SAR-X, and Sentinel-1 satellites, containing a total of 2,540 ship targets. The hardware development environment used in this paper is Ubuntu 18.04, with an AMD Ryzen R7-3700X @3.80Ghz×16 CPU, NVIDIA RTX 3090 24GB GPU, and DDR4 3600MHz 32GB RAM. The software platform is based on the PyTorch 1.10 framework, Python 3.10, and uses CUDA 11.7 and CUDNN 8.5 to accelerate training. During the training process, 300 epochs were performed with a batch size of 32. The learning

rate was set to 0.0001, and the AdamW optimizer was selected for parameter optimization.

3.2. Evaluation Metrics

In this paper, Precision (P), Recall (R), and mean Average Precision (mAP) are used to evaluate the detection performance of the algorithm. Params, Weight Size, GFLOPs and FPS are used to quantify the computational complexity of the overall model. The formulas are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$mAP = \int_0^1 P(R)dR \tag{3}$$

Where Precision (P) measures the accuracy of positive predictions made by the model. Recall (R), also known as sensitivity or true positive rate, measures the ability of the model to identify all positive instances. Mean Average Precision (mAP) is a popular evaluation metric for object detection algorithms. It calculates the average precision across different object categories and provides an overall measure of the algorithm's performance.

Params refer to the learnable weights of a machine learning model and the number of parameters indicates the model's complexity and capacity to capture data patterns. Weight size represents the memory required to store the model's parameters. GFLOPs measure the computational performance of a model in terms of floating-point operations per second which helps estimate the computational demands of deep learning models. FPS means the number of frames per second and stands as a pivotal gauge for evaluating the real-time responsiveness and speed of algorithms.

3.3. Ablation Experiments

In this part, we analyze the effect of each module in detail, the results of ablation experiments for each module are shown in Table 1.

Table 1. Experimental results of the ablation study for each module.

Model	Decoder Head	IRMB Module	Parms (M)	mAP
YOLOv8s			11.15	0.971
YO-DETR	√		6.10	0.972
		√	8.66	0.974
	√	√	5.37	0.982

The experimental results in Table 1 demonstrate that the use of decoder head and IRMB module reduces the parameter count and improves the detection ability of the model. The design of the decoder head reduces the computational cost of the NMS process, where the use of the decoder head decreases 5.05M Params than YOLOv8s and the mAP has improved by 0.1%. The IRMB module learns long-distance interactions through its reverse residual structure, and it also reduces 2.49M Params and improves 0.3% mAP performance than YOLOv8s.

3.4. Comparative Experiments

In this section, the effectiveness of the proposed method will be analyzed by comparing it with four mainstream methods. All methods will be compared from the perspectives of accuracy and parameter count, and the experimental results are shown in Table 2 and Table 3.

Table 2. Comparison results of detection accuracy.

Model	P	R	mAP
RT-DETR(r18)	0.965	0.943	0.972
RT-DETR(r50)	0.953	0.941	0.975
YOLOv8s	0.959	**0.957**	0.971
YOLOv5s	0.964	0.942	0.968
YO-DETR	**0.967**	0.946	**0.982**

Table 2 shows the powerful feature encoding and decoding capabilities based on the transformer structure, compared to the traditional CNN-based YOLO series. The RT-DETR algorithm demonstrates superior performance in terms of accuracy. RT-DETR (r18) achieves 97.2% in mAP on the SSDD dataset, surpassing YOLOv5s and YOLOv8s by 0.4% and 0.1%, respectively.

Table 3. Comparison results of parameter count.

Model	Params (M)	Weight Size (MB)	GFLOPs	FPS
RT-DETR (r18)	19.87	40.5	56.9	211.9
RT-DETR (r50)	41.95	86.0	129.5	170.6
YOLOv8s	11.15	22.5	28.6	**237.5**
YOLOv5s	9.11	18.5	23.8	225.8
YO-DETR	**5.37**	**10.5**	**10.0**	220.4

From Table 3, we can conclude that the algorithm based on the transformer structure with more parameters due to the extensive use of encoding and decoding operations. The Params of the RT-DETR (r18) model is 1.78 times that of YOLOv8s, and the weight size is 1.8 times larger. The proposed YO-DETR has a parameter size of 5.37M (reduced by 14.5M compared to RT-DETR), and a weight size of only 10.5MB (reduced by 30MB compared to RT-DETR). Despite reductions in both Params and GFLOPs, our method only lags behind YOLOv8s by 17.1 frames in terms of FPS. This is due to the current lack of GPU support for parallel acceleration of transformer structures, resulting in a slower detection inference speed.

Taking into account both Table 2 and Table 3, the proposed YO-DETR achieves higher detection accuracy while having fewer parameters with a parameter of 5.37M, it achieves 98.2% in mAP. YO-DETR inherits the powerful and

lightweight feature extraction capabilities of the YOLO series based on CNN, while incorporating the unified efficiency of the transform's encoding and decoding structure, making it more suitable for deployment on mobile devices.

4. CONCLUSION

In this paper, we propose an end-to-end lightweight SAR ship detection algorithm, YO-DETR. In the backbone network, a pyramid structure improved with IRMB modules is employed as the feature extraction network. This not only retains the strong feature extraction capability of the CNN network but also establishes long-range feature interaction. A transformer-based Decoder Head is utilized as the detection head of the algorithm. The powerful encoding and decoding capabilities of transformers ensure more accurate detection. Moreover, the entire detection process is not constrained by the NMS threshold, significantly reducing the speed of anchor generation and thereby improving detection efficiency. The proposed YO-DETR achieves 98.2% in mAP on the SSDD dataset, outperforming RT-DETR in terms of both accuracy and model parameters. Future work will focus on model compression and distillation to further reduce the parameter size of the model.

Three different scenes are selected to validate the effectiveness of the proposed method and the results are shown in Figure 4. In the first scene, due to improper selection of the NMS threshold, YOLOv8s have false alarms, while RT-DETR and YO-DETR accurately detected the ship target, with YO-DETR exhibiting higher confidence. The second scene involved a nearshore scene with multi-scale ship targets, where YOLOv8s missed small targets, while the proposed YO-DETR method successfully detected all targets. In the third scene, which featured densely arranged nearshore ship targets, the YOLOv8s-based detection algorithm generated false alarms along the coastline. However, our proposed method not only avoided false alarms but also accurately detected all dense ship targets. These results demonstrate that the proposed method, without the need for considering NMS threshold selection, effectively reduces the occurrence of false alarms to a certain extent.

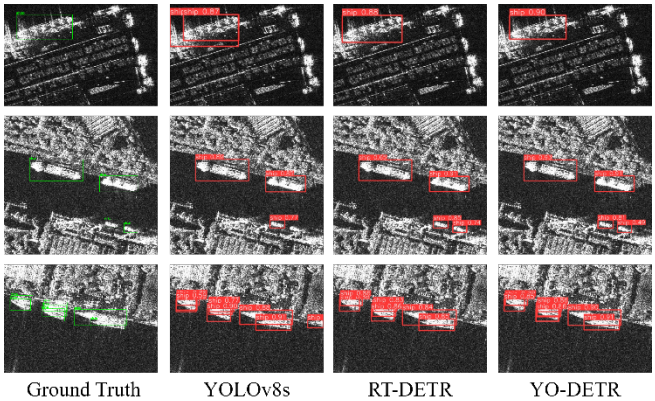

Figure 4. The visualization results of different methods.

5. REFERENCES

[1] Zhu H, Xie Y, Huang H, et al. DB-YOLO: A duplicate bilateral YOLO network for multi-scale ship detection in SAR images[J]. Sensors, 2021, 21(23): 8146.

[2] Zhao W, Syafrudin M, Fitriyani N L. CRAS-YOLO: A Novel Multi-Category Vessel Detection and Classification Model Based on YOLOv5s Algorithm[J]. IEEE Access, 2023, 11: 11463-11478.

[3] Sun Z, Leng X, Lei Y, et al. BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images[J]. Remote Sensing, 2021, 13(21): 4209.

[4] Guo Y, Chen S, Zhan R, et al. LMSD-YOLO: A Lightweight YOLO Algorithm for Multi-Scale SAR Ship Detection[J]. Remote Sensing, 2022, 14(19): 4801.

[5] Chen Z, Liu C, Filaretov V F, et al. Multi-Scale Ship Detection Algorithm Based on YOLOv7 for Complex Scene SAR Images[J]. Remote Sensing, 2023, 15(8): 2071.

[6] Ren X, Bai Y, Liu G, et al. YOLO-Lite: An efficient lightweight network for SAR ship detection[J]. Remote Sensing, 2023, 15(15): 3771.

[7] Li K, Zhang M, Xu M, et al. Ship detection in SAR images based on feature enhancement Swin transformer and adjacent feature fusion[J]. Remote Sensing, 2022, 14(13): 3186.

[8] Feng K, Lun L, Wang X, et al. LRTransDet: A Real-Time SAR Ship-Detection Network with Lightweight ViT and Multi-Scale Feature Fusion[J]. Remote Sensing, 2023, 15(22): 5309.

[9] Zha M, Qian W, Yang W, et al. Multifeature transformation and fusion-based ship detection with small targets and complex backgrounds[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[10] Yu N, Ren H, Deng T, et al. A Lightweight Radar Ship Detection Framework with Hybrid Attentions[J]. Remote Sensing, 2023, 15(11): 2743.

[11] Shi H, Chai B, Wang Y, et al. A Local-Sparse-Information-Aggregation Transformer with Explicit Contour Guidance for SAR Ship Detection[J]. Remote Sensing, 2022, 14(20): 5247.

[12] Zhou W, Shen J, Liu N, et al. An anchor-free vehicle detection algorithm in aerial image based on context information and transformer[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[13] Zhang J, Li X, Li J, et al. Rethinking mobile block for efficient attention-based models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 1389-1400.

[14] Lv W, Xu S, Zhao Y, et al. Detrs beat yolos on real-time object detection[J]. arXiv preprint arXiv:2304.08069, 2023.