An Enhanced RT-DETR with Dual Convolutional Kernels for SAR Ship Detection

Chushi Yu
School of Electronic Engineering
Soongsil University
Seoul 06978, Korea
Email: csyu@soongsil.ac.kr

Yoan Shin*
School of Electronic Engineering
Soongsil University
Seoul 06978, Korea
Email: yashin@ssu.ac.kr

2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) | 979-8-3503-4434-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAIIC60209.2024.10463389

***Abstract*—In recent years, ship detection based on remote sensing images has emerged as a crucial task for coastal nations due to the advancement of remote sensing technology. Among active imaging sensors in remote sensing, synthetic aperture radar (SAR) stands out as one of the most significant ones because of its immunity to cloud cover and ability to operate day and night. However, ship targets in SAR images pose challenges such as indistinct contour information, intricate backgrounds, and intense scattering. Despite commendable results achieved by ship detection algorithms based on deep learning neural networks, they still suffer from numerous missed detections and false alarms. In this study, we propose an enhanced real-time detection transformer (RT-DETR) with dual convolutional kernels (DualConv) for accurate ship detection in SAR images. Numerical experiments conducted on the high-resolution SAR image dataset (HRSID) demonstrate the effectiveness of the proposed method, improving detection accuracy and model's robustness and capability in complex marine environments.**

Keywords—Synthetic aperture radar, remote sensing, ship detection, deep learning, real-time detection transformer

I. INTRODUCTION

Synthetic aperture radar (SAR) is a microwave imaging sensor that can actively detect in all-day and all-weather conditions. It has excellent applicability for monitoring oceans with changing climates [1]. SAR remains unaffected by the changeable ocean weather, providing real-time monitoring capabilities for ship targets in all directions [2].

Computer vision has undergone a revolutionary transformation driven by the arrival of convolutional neural Networks (CNNs) and deep learning architectures. Object detection, a crucial aspect of computer vision, aims to locate and categorize specific objects in images or videos. In recent decades, various object detection methods have been proposed, including feature-based, template-based, and deep learning-based approaches. Among these, deep learning methods have made great progress, especially methods based on CNN, such as the faster region-based convolutional neural network (Faster R-CNN) [3] and you only look once (YOLO) [4], have achieved good results. CNN architectures are generally heavy on memory and computational requirements which makes them infeasible for embedded systems with limited hardware resources.

DEtection Transformer (DETR) is a relatively new object detection algorithm that was introduced in 2020 by researchers at Facebook AI Research (FAIR) [5]. Unlike the traditional two-stage pipeline, DETR replaces it with a transformer, providing the advantages of an end-to-end architecture and global context modeling. By leveraging the self-attention mechanism, allowing the model to comprehend contextual features and their correlation, transformer-based methods have emerged as a recent breakthrough compared to CNN-based detectors. However, the unaddressed issue of the high computational cost of DETRs limits their practical application and preventing them from fully exploiting the benefits of no post-processing, such as non-maximum suppression (NMS).

Real-Time DEtection TRansformer (RT-DETR) was proposed to eliminate the inference delay caused by NMS and outperform YOLO detectors of the same scale in both speed and accuracy [6]. They design an efficient hybrid encoder to efficiently process multi-scale features by decoupling the intra-scale interaction and cross-scale fusion, and propose intersection over union (IoU)-aware query selection to improve the initialization of object queries. The proposed detector supports flexibly adjustment of the inference speed by using different decoder layers without the need for retraining, which facilitates the practical application of real-time object detectors. In addition, dual convolutional kernels (DualConv) for constructing lightweight deep neural networks was proposed by [7]. DualConv can be employed in any CNN model by some structural innovations, and significantly reduces the computational cost and number of parameters of deep neural networks while surprisingly achieving slightly higher accuracy than the original models.

In this paper, based on the RT-DETR, we conduct training optimization for the characteristics of the SAR ship detection combined with DualConv. The high resolution SAR images dataset (HRSID) [8] are used for verifying the effectiveness and the applicability of the proposed scheme.

II. METHODOLOGY

A. Overview of the Proposed Method

In this paper, we propose the enhanced RT-DETR method for SAR ship targets detection, leveraging the integration of dual convolutional filters. The architecture is delineated by three main components: a backbone, an efficient hybrid encoder, and a transformer decoder with auxiliary prediction heads, as illustrated in Fig. 1. The architecture diagram of the enhanced RT-DETR model shows the utilization of the last three stages of the backbone $\{S_3, S_4, S_5\}$ as inputs to the encoder. The efficient hybrid encoder transforms multiscale features into a sequence of image features through intrascale feature interaction (AIFI) and cross-scale feature-fusion module (CCFM). To initiate the decoding process, an IoU-aware query selection mechanism is employed to select a fixed number of image features to serve as initial object queries for the decoder. Finally, the decoder with auxiliary prediction heads iteratively optimizes object queries to generate boxes and confidence scores.

*Corresponding author

This work was supported in part by the NRF grant funded by the Korea government (MSIT) (No. RS-2023-00251595), and by the MSIT, Korea, under the ITRC support program (IITP-2023-RS-2023-00258639) supervised by the IITP.

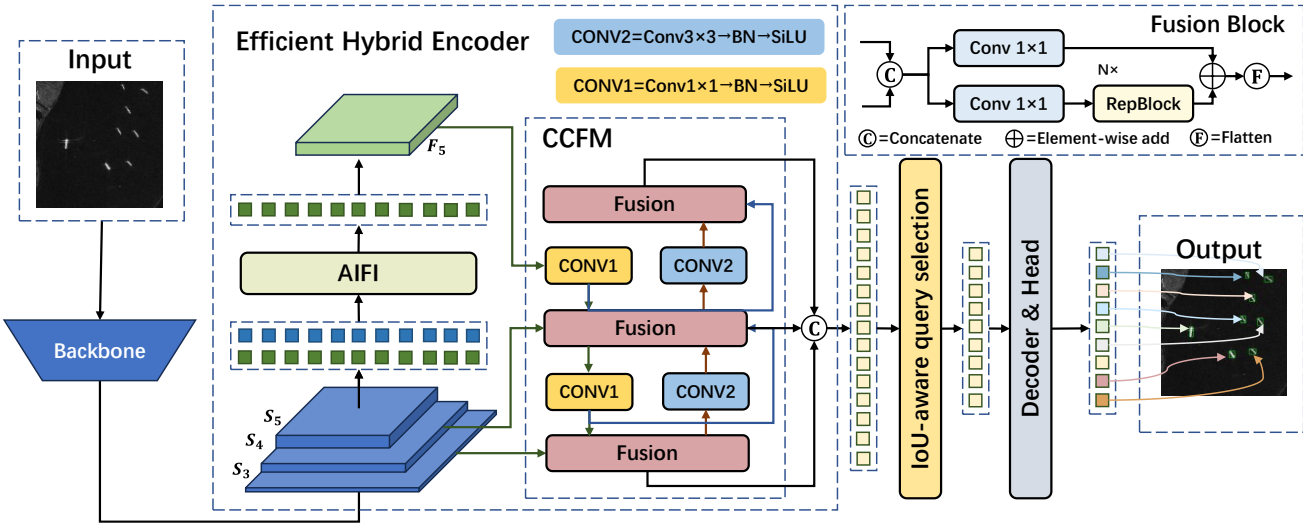

Fig. 1. Detailed illustration of the proposed model architecture based on RT-DETR.

B. Dual Convolutional Filters

We employ the dual convolutional filters instead the original convolutional operation in the backbone. DualConv integrates 3×3 and 1×1 convolutional kernels for simultaneous processing of input feature map channels, utilizing the group convolution technique for efficient filter arrangement as shown in Fig. 2. M represents the number of input channel (depth of input feature map), N is the number of convolutional filters and output channels (depth of output feature map), and G is the number of groups in dual convolution. N convolutional filters are divided into G groups. Each group processes the complete input feature map, with M/G input feature map channels concurrently handled by 3×3 and 1×1 convolutional kernels. The remaining channels $(M - M/G)$ are processed by 1×1 convolutional kernels exclusively. The results of simultaneous 3×3 and 1×1 convolutional kernels are summed up, denoted by the $\oplus$ sign in Fig. 2. The filter group structure enforces block-diagonal sparsity on the channel dimension, facilitating structured learning of highly correlated filter. Consequently, convolutional filters are not arranged in a shifted manner. The DualConv reduces parameters in original backbone network models through group convolution strategy, and promotes information sharing between convolutional layers. This is achieved by preserving the original information of input feature maps and enabling maximum cross-channel communication with M 1×1 convolutions. As a result, DualConv can be constructed without the need for channel shuffle operation.

C. Efficient Hybrid Encoder

The proposed encoder comprises two integral modules, the AIFI and CCFM module. AIFI selectively engages in intra-scale interaction solely on S_5, thus effectively mitigating computational redundancy. Appling self-attention operations to high level features characterized by richer semantic concepts facilitates the establishment of connection between conceptual entities within the image, which enhances the subsequent modules' capabilities in detecting and recognizing objects. Meanwhile, intra-scale interactions of lower-level features are unnecessary due to the lack of semantic concepts and the risk of duplication and confusion in high-level feature interactions. CCFM is instantiated through the insertion of several fusion blocks, comprising convolutional layers into the fusion path. These fusion blocks play a pivotal role in

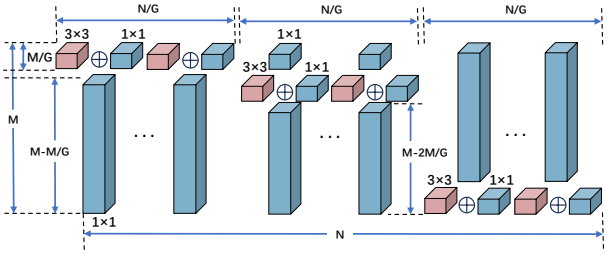

Fig. 2. The architecture of the dual convolutional filters.

amalgamating adjacent features, culminating in the creation of a novel composite feature. The structural depiction of the fusion block is illustrated in Fig. 1, which contains N RepBlocks [9], and the two-path outputs are fused by element-wise add. The process can be described as

$$Q = K = V = Flatten(S_5), \tag{1}$$

$$F_5 = Reshape\big(Att(Q, K, V)\big), \tag{2}$$

$$Output = CCFM(\{S_3, S_4, F_5\}), \tag{3}$$

where Att represents the multi-head self-attention, and $Reshape$ represents restoring the shape of the feature to the same as S_5, which is the inverse operation of $Flatten$.

D. IoU-aware Query Selection

IoU-aware query selection was proposed by constraining the model to produce high classification scores for features with high IoU [10] scores and low classification scores for features with low IoU scores during training [6]. Therefore, the prediction boxes corresponding to the top K encoder features selected by the model according to the classification score have both high classification scores and high IoU scores. The incorporation of the IoU score into objective function of the classification branch serves to enforce a consistency constraint on both the classification and localization of positive samples. We reformulate the optimization objective of the detector as follows

$$\begin{aligned} L(\hat{y}, y) &= L_{box}\big(\hat{b}, b\big) + L_{cls}\big(\hat{c}, \hat{b}, y, b\big) \\ &= L_{box}(\hat{b}, b) + L_{cls}(\hat{c}, c, IoU), \end{aligned} \tag{4}$$

where $\hat{y}$ and y denote prediction and ground truth, $\hat{y} = \{\hat{c}, \hat{b}\}$, and $y = \{c, b\}$, c and b represent categories and bounding boxes, respectively.

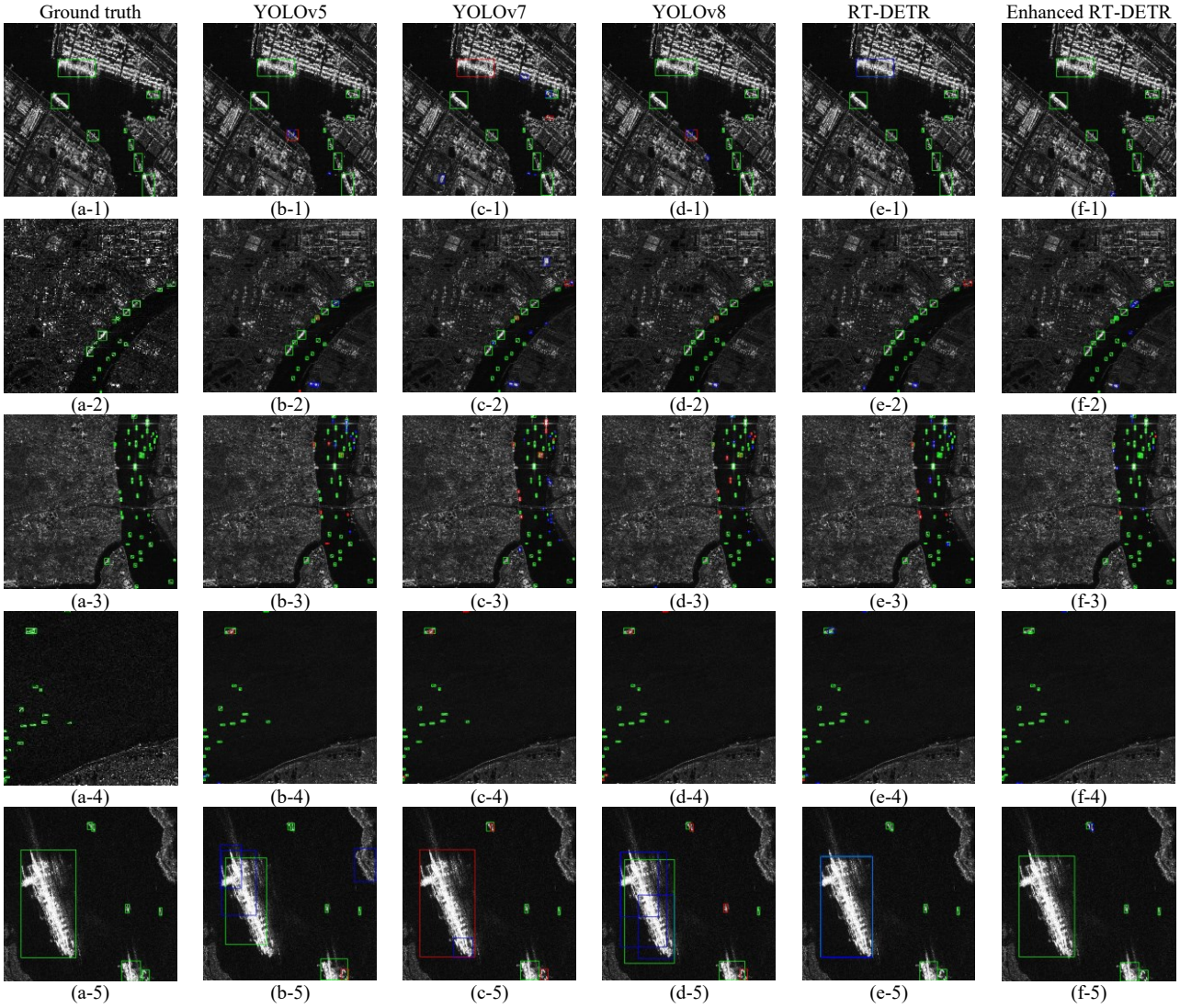

Fig. 3. Comparison of the qualitative ship detection results for the enhanced RT-DETR and other methods in the HRSID. The green boxes represent correctly-detected ships, red boxes indicate missing ships, and blue boxes denote false alarms. (a) is the ground truth of SAR images, and (b-f) illustrate the detection results of different scale ships and complex background.

III. Experimental Results

A. Dataset and Training Strategy

We use the HRSID [7] to verify the performance of the enhanced RT-DETR method. The HRSID is a well-established dataset for SAR ship detection and instance segmentation, which consists of Sentinel-1 and TerraSAR-X images with resolutions of 0.5m, 1.0m, and 3.0m as contains 5,604 high-resolution SAR images and 16,591 ship instances. It draws on the construction process of the COCO [11] datasets, including SAR images with different resolutions, polarizations, sea conditions, sea areas, and coastal ports, which is a benchmark for researchers to evaluate their methodologies. To conduct our experiments, we randomly divide images into a training set (65%) and a test set (35%). All the experiments are conducted on an NVIDIA RTX 4080 graphics processing unit (GPU) using PyTorch framework. The batch size is set to 16 and the number of training epochs is set to 300. We use the AdamW optimizer with an initial learning rate of 0.0001, and the weight decay is set to 0.0001.

TABLE I. Results of the HRSID

Models	P	R	AP_{50}	$AP_{50:95}$	GFLOPs	Speed (ms)
YOLOv5	0.909	0.853	0.924	0.662	4.1	2.5
YOLOv7	0.882	0.77	0.866	0.578	13.0	3.2
YOLOv8	0.898	0.835	0.913	0.667	8.1	2.1
RT-DETR	**0.93**	0.868	0.935	0.713	56.9	3.5
Enhanced RT-DETR	0.929	**0.871**	**0.938**	**0.715**	47.3	3.4

B. Results and Discussion

In this experiment, the precision (P), the recall (R), the mean average precision (mAP), $GFLOPs$ and speed are used to evaluate the detection performance of the models.

Table 1 lists the results of the HRSID with the enhanced RT-DETR and several YOLO models. The comparative experiments with state-of-the-art methods are conducted based on the YOLOv5 [12], YOLOv7 [13, 14] and YOLOv8 [15]. The evaluation metrics demonstrate that the enhanced method outperforms the baseline YOLOv5, YOLOv7 and YOLOv8 methods in terms of accurate ship detection and maintains detection speed to meet the needs of real-time performance.

Figure 3 presents exemplary results visualizing the detection results achieved by the proposed scheme. Experimental results demonstrate the effectiveness of the model in recognizing small-scale ships amidst diverse scales and intricate backgrounds. Compared with the baseline YOLO-based models and the original RT-DETR model, our proposed method performs well identifying and locating large-size ships simultaneously.

IV. Conclusion

In this work, we introduce an enhanced real-time detection transformer (RT-DETR) incorporating dual convolutional kernels (DualConv) for precise ship detection in SAR images, following a comprehensive analysis of existing state-of-the-art object detection algorithms. The dual convolutional filters effectively reduce the computational cost of the original RT-DETR model, catering to the demands of real-time detection. Combined with the real SAR dataset, our approach aims to detect ship targets fast and accurately in the complex marine environment. Multiple YOLO-based models were trained and evaluated to facilitate a comparative assessment of test results. The experimental results reveal the promising performance of our proposed method in SAR ship detection, offering significant potential for practical applications. This study not only fulfills ship detection requirements in terms of accuracy and real-time performance but also demonstrates superior recognition accuracy for large-scale ships compared to alternative models. Our proposed method thus serves as a valuable theoretical reference for addressing similar challenges in maritime object detection.

References

[1] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou, "A tutorial on synthetic aperture radar," *IEEE Geosci. & Remote Sensing Mag.*, vol. 1, no. 1, pp. 6-43, Mar. 2013.

[2] J. Li, C. Xu, H. Su, L. Gao, and T. Wang, "Deep learning for SAR ship detection: Past, present and future," *Remote Sensing*, vol. 14, no. 11, pp. 2712, June 2022.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137-1149, June 2017.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proc. IEEE CVPR 2016*, pp. 779-788, Las Vegas, USA, Dec. 2016.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-end object detection with transformers," *Proc. ECCV 2020*, pp. 213-229, virtual conference, Aug. 2020.

[6] W. Lv, Y. Zhao, S. Xu, J. Wei, G. Wang, C. Cui, Y. Du, Q. Dang, and Y. Liu, "Detrs beat yolos on real-time object detection," *arXiv preprint arXiv:2304.08069*, April 2023.

[7] J. Zhong, J. Chen and A. Mian, "DualConv: Dual convolutional kernels for lightweight deep neural networks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 11, pp. 9528-9535, Nov. 2023.

[8] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234-120254, June 2020.

[9] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making vgg-style convnets great again," *arXiv preprint arXiv:2101.03697*, Jan. 2021.

[10] J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO: From YOLOv1 and beyond," *arXiv preprint arXiv:2304.00501*, April 2023.

[11] T-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," *arXiv preprint arXiv:1405.0312*, May 1.

[12] G. Jocher, "YOLOv5," available online at https://github.com/ultralytics/yolov5, 2020.

[13] C. Wang, A. Bochkovskiy, and H. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, July 2022.

[14] C. Yu and Y. Shin, "A deep learning-based SAR ship detection," *Proc. ICAIIC 2023*, pp. 744-747, Bali, Indonesia, Feb. 2023.

[15] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," https://github.com/ultralytics/ultralytics, Feb. 2023.