IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10641003

CDE-DETR: A REAL-TIME END-TO-END HIGH-RESOLUTION REMOTE SENSING OBJECT DETECTION METHOD BASED ON RT-DETR

Anrui Wang, Yang Xu, He Wang, Zebin Wu, Zhihui Wei*

School of Computer Science and Engineering,
Nanjing University of Science and Technology, Nanjing, China

ABSTRACT

High-resolution remote sensing object detection is a research field with significant application value and challenges. However, existing methods cannot directly predict or produce a large number of redundant bounding boxes. They perform poorly in the face of issues such as multi-scale, dense small objects, and complex backgrounds in high-resolution remote sensing images. Therefore, a real-time end-to-end high-resolution remote sensing object detection method based on RT-DETR (CDE-DETR) is proposed. Through introducing cascaded group attention, we propose CGA-IFI for intra-scale feature interaction. The DRB-CFFM is designed with a dilated reparam block to facilitate cross-scale feature interaction. Furthermore, we enhance the bounding box regression loss function with EIoU. Experimental results demonstrate that the accuracy mAP value of our method is 2.9% higher than the baseline, FPS is increased by 33.8%. The number of parameters is reduced by 9.9%, and FLOPs is reduced by 16.0%. Compared with other methods, the proposed method has obvious accuracy and lightweight advantages.

Index Terms— High-resolution remote sensing, feature interaction, cascaded croup attention, dilated reparam block, EIoU

1. INTRODUCTION

High-resolution remote sensing object detection refers to the utilization of high-resolution satellite or aerial images for automatically identification and localization of objects of interest in images, such as vehicles, buildings, ships, aircraft, etc. This technology provides technical support for military reconnaissance, maritime rescue, disaster control and other major tasks, providing high practical value. However, it also encounters some challenges, such as background interference, multi-scale and dense small objects.

Several studies have been conducted in the field. Dong et al. [1] proposed a high-resolution remote sensing object detection method based on convolutional neural networks (CNN) with suitable object scale features, addressing the issues about region of interest (ROI) and object feature representation. Zhang et al. [2] proposed a multi-scale hard-example-mining network (MSHEMN) to deal with the problems of multi-scale objects and complex image background.

The traditional object detection methods do not involve direct predictions. Instead, the object detection task is treated as either a regression or classification challenge, relying on existing initial hypotheses for making these predictions. Their performance is very related to the initial predictions. Therefore, how to do NMS is crucial to performance, but it will also produce a large number of redundant bounding boxes.

Different from the above methods, DETR [3] uses a completely different approach. Transformer and CNN feature extraction models are combined to directly predict the category and bounding box of the object in an end-to-end manner, without relying on some of the conventional steps in the traditional object detection process, such as NMS.

However, the high detection cost of DETR limits its detection speed. Therefore, Baidu proposed the first real-time end-to-end object detection method, named Real-Time DETR (RT-DETR) [4]. An efficient hybrid encoder was designed for intra-scale interaction and cross-scale fusion. The method demonstrates superior performance on accelerated platforms like CUDA, surpassing other real-time methods.

In this paper, CDE-DETR, an improved high-resolution remote sensing object detection method based on RT-DETR is proposed to improve the object detection accuracy in the high-resolution remote sensing scene and maintain the real-time lightweight deployment capability. Regarding the issue of high computational cost in multi-head self-attention for intra-scale feature interaction, CGA-IFI is proposed to reduce time consumption and improve the diversity of attention. In response to the limited receptive field in cross-scale feature interaction, DRB-CFFM is proposed to enhance the understanding of image details. In response to the issue of small overlap areas between bounding boxes and real bounding boxes, the EIoU bounding box loss function is introduced to improve the detection ability of overlapping objects.

The structure of this paper is as follows: Section 2 provides a concise overview of our method. Section 3 is dedicated to presenting a thorough experimental evaluation and detailed discussion of the method in question. Finally, Section 4 concludes the paper.

*Corresponding author

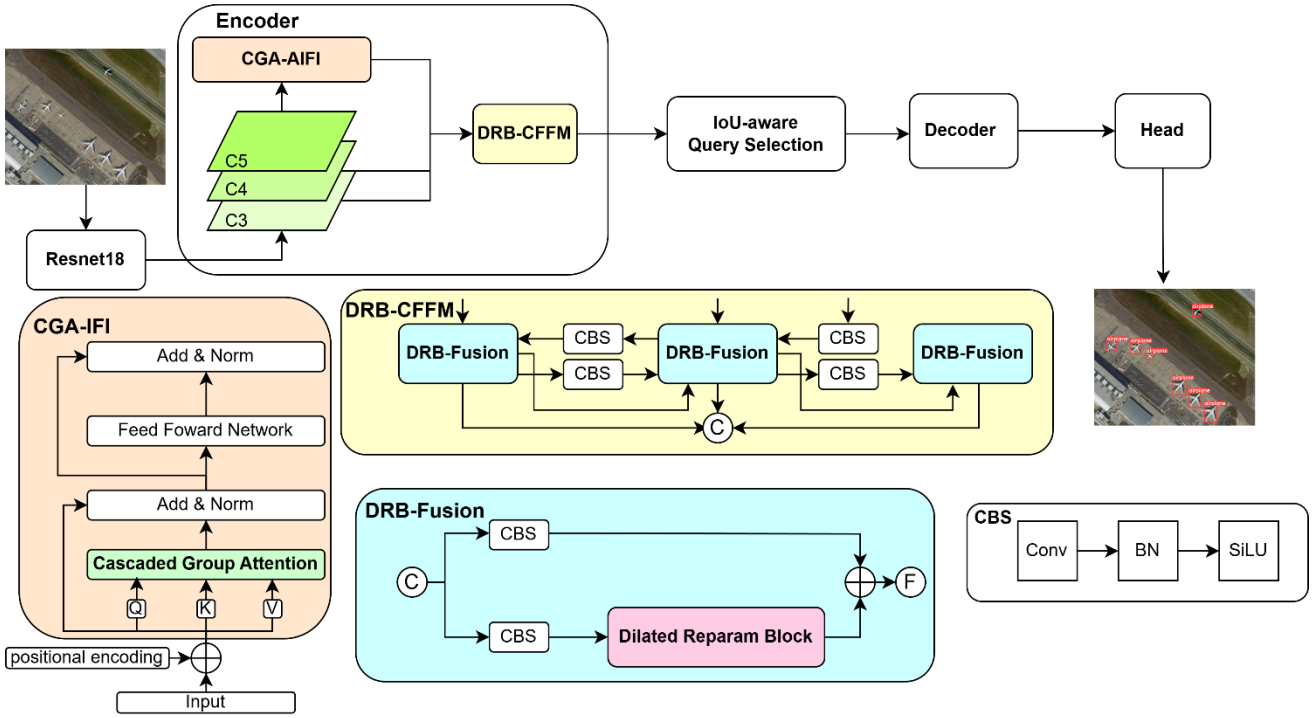

Fig. 1. The overall architecture of CDE-DETR

2. METHOD

Fig. 1. shows the overall architecture of our CDE-DETR. We use Resnet18 as backbone to extract features from the input image, then choose the last three stages of the backbone {C3, C4, C5} as encoder inputs. The encoder converts multi-scale features into image feature sequences through the cascaded group attention-based intra-scale feature interaction (CGA-IFI) and dilated reparam block-based cross-scale feature-fusion module (DRB-CFFM). IoU-aware query selection selects the relevant image features in the output of the encoder as the initial object query of the decoder. Finally, the decoder decodes the feature map to realize the detection of the object through the head, and generates the bounding box and confidence score.

2.1. CGA-IFI

RT-DETR uses multi-head self-attention (MHSA) for intra-scale feature interaction. However, the computational complexity of MHSA is high, resulting in an increase in memory consumption and computation time. The high computational cost reduces the detection speed in high-resolution remote sensing object detection. At the same time, the limited resources cannot be concentrated on the detection of key objects, which reduces the efficiency of detection.

As shown in Fig. 1, to address the above problems, we propose cascaded group attention-based intra-scale feature interaction (CGA-IFI). Since C5 contains rich semantic concepts, in order to reduce the interference of redundant information, we only use C5 and the corresponding positional encoding as input, and each encoder layer has a standard architecture consisting of a cascaded group attention (CGA) module and a feedforward network (FFN).

As shown in Fig. 2, The core idea of CGA [5] is to partition a complete feature into different groups, with each group corresponding to an attention head. This allows each head can focus on a distinct feature subspace. It can reduce the computational redundancy problem caused by MHSA and improve the diversity of attention. CGA can be formulated as:

$$\tilde{X}_{ij} = Attn\left(X_{ij}W_{ij}^{\mathrm{Q}}, X_{ij}W_{ij}^{\mathrm{K}}, X_{ij}W_{ij}^{\mathrm{V}}\right), \quad (1)$$

$$\tilde{X}_{i+1} = Concat\left[\tilde{X}_{ij}\right]_{j=1:h} W_i^{\mathrm{P}}, \quad (2)$$

where the j-th head computes the self-attention over X_{ij}, which is the j -th split of the input feature X_i, i.e., $X_i = [X_{i1}, X_{i2}, \ldots, X_{ih}]$, h is the total number of heads, W_{ij}^{Q}, W_{ij}^{K}, and W_{ij}^{V} serve as projection layers, each converting a segment of the input features into its respective subspace. W_i^{P} is a linear layer that projects the concatenated output features back into the same dimensional space as the input features.

2.2. DRB-CFFM

In order to tackle the problem of small receptive field in cross-scale interaction, we designed a dilated reparam block-based cross-scale feature-fusion module (DRB-CFFM). We fuse the features outputted from CGA- IFI via up-sampling with the features C3 and C4 extracted through Resnet18 via

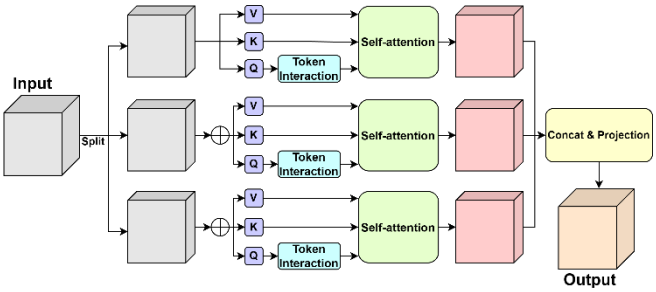

Fig. 2. Cascaded Group Attention

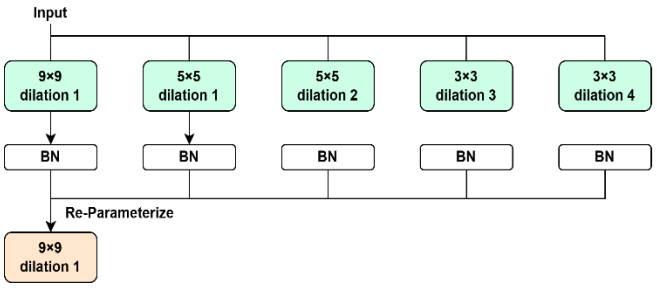

Fig. 3. The Re-Parameterize of Dilated Reparam Block

DRB-Fusion. Then, we perform cross-scale feature interaction through down-sampling and concatenate the final results. In the DRB-Fusion module, we use the dilated reparam block (DRB) [6] and two CBS for feature fusion.

As shown in Fig. 3, DRB utilizes the idea of structural re-parameterization and dilated small-kernel conv layers to enhance a non-dilated large-kernel layer. If the size of the dilated conv layer is $k \times k$, the size of the corresponding non-dilated kernel layer will be $\big((k-1)r+1\big) \times \big((k-1)r+1\big)$, where r is the corresponding number of layers. The only constraint is $(k-1)r+1 \le K$, K is the size of the large kernel.

DRB can increase the receptive field with an extended convolutional layer, so that each pixel can better understand its surrounding context information. This enhances the ability to understand the details and global structure of the image.

2.3. EIoU Bounding Box Regression Loss

Regarding the bounding box regression loss, we use the EIoU loss function for calculation. The GIoU loss, used by RT-DETR, may focus too much on the minimum enclosing rectangle in some cases, resulting in a smaller overlap area between the predicted bounding box and the true bounding box. It doesn't perform well when dealing with multi-scale problems in high resolution remote sensing object detection.

Based on GIoU, Efficient-IoU (EIoU) [7] separates the aspect ratio, and explicitly measures the differences in three geometric elements in bounding box regression: the overlap area, the center point, and the side length. This accelerates the convergence speed and improves the accuracy of regression, reduces the optimization contribution of a large number of anchor boxes with less overlap with the object box to bbox regression, making the regression process more focused on high-quality anchor boxes. The calculation formula is shown as follows:

$$\begin{aligned}\mathcal{L}_{EIoU} &= \mathcal{L}_{IoU} + \mathcal{L}_{Dis} + \mathcal{L}_{Asp} \\ &= \mathcal{L}_{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{(w^c)^2} + \frac{\rho^2(h, h^{gt})}{(h^c)^2} \end{aligned} \quad (3)$$

Among them, $\mathcal{L}_{IoU}$ represents IoU loss, $\mathcal{L}_{Dis}$ represents distance loss, $\mathcal{L}_{Asp}$ represents side length loss, w^c and h^c is the width and height of the minimum closure area covering the prediction box and the object box, w and h is the width and height of the prediction box, w^{gt} and h^{gt} is the width and height of the object box respectively.

3. EXPERIMENTS

3.1. Dataset

The NWPU VHR-10 dataset is a publicly available 10-class geospatial object detection dataset used for research purposes only. These ten classes of objects are airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. This dataset contains totally 800 very-high-resolution (VHR) remote sensing images, of which 650 have objects and 150 are negative images without objects. 80% of the images with instances are used as the training set and 20% are the test set. The HRSC2016 dataset is specifically designed for the detection of ships within high-resolution aerial imagery. It includes a wide range of ship types and sizes, from small boats to large-scale aircraft carriers, captured in different conditions and scenarios. The HRSC2016 dataset is split into three subsets: 444 images for training, 436 for testing, and 181 for validation.

3.2. Results and Analysis

We use RT-DETR as the baseline method, and validate the effectiveness of our proposed method CDE-DETR on the NWPU VHR-10 dataset. In addition, we also conduct experiments on the HRSC2016 dataset to verify the universality of our method.

Results on NWPU VHR-10. As shown in Table 1, we first compared the proposed method with the baseline on 10 classes of the NWPU VHR-10 dataset. Compared with the baseline, our method takes the lead in seven categories. As shown in Table 2, in order to verify the effectiveness of each improvement component, we conducted ablation experiments, and each group of ablation experiments improved the mAP_{50}. The mAP_{50} of our proposed method has increased by 2.9% compared to the baseline, the FPS has increased by 33.8%, the number of parameters has decreased by 9.9%, and the GFLOPs has decreased by 16.0%. As shown in Table 3, we also compare CDE-DETR with other methods. CDE-DETR surpasses all other comparison methods, demonstrating the superiority of our method.

Table 1. Comparison with the baseline on the NWPU VHR-10 dataset

Method	AL	SH	ST	BD	TC	BC	GTF	HAS	BR	VE
Baseline	**99.3**	**93.9**	64.6	99.0	94.4	**93.0**	99.5	76.6	82.0	88.5
CDE-DETR(Ours)	99.2	92.5	**72.5**	**99.3**	**94.6**	91.5	**99.5**	**92.2**	**87.2**	**91.0**

Table 2. Ablation Experiments on NWPU VHR-10 dataset

CGA	DRB	EIoU	mAP_{50}	FPS	Params	GFLOPs
			89.1	51.7	20.1	58.6
√			90.6	**76.9**	20.0	58.7
√	√		91.5	70.0	18.1	49.2
√	√	√	**92.0**	69.2	**18.1**	**49.2**

Table 3. Comparison with other methods on NWPU VHR-10 dataset

Methods	mAP_{50}
MS-FF	85.6
Faster-RCNN	88.4
CA-CNN	91.0
RFBNet	91.6
SCRDet	91.75
CDE-DETR(Ours)	**92.0**

Results on HRSC2016. As shown in Table 4, we compared CDE-DETR with other real-time methods on the HRSC2016 dataset to verify the universality of the proposed method. From Table 4, it can be concluded that the mAP_{50} of the CDE-DETR is higher than other real-time methods, and it also has the least number of parameters, which facilitates the real-time lightweight deployment of our method.

Table 4. Comparison with other real-time methods on HRSC2016 dataset

Methods	mAP_{50}	Params
YOLOv5m	92.1	25.1
YOLOv6m	91.3	52.0
YOLOv7	83.7	36.9
YOLOv8m	91.9	25.8
SSD	86.5	24.4
CDE-DETR(Ours)	**92.6**	**18.1**

Fig. 4. Visualization of Results

4. CONCLUSION

The accuracy and real-time requirements for high-resolution remote sensing object detection methods are exceedingly high. Therefore, this paper proposes a real-time end-to-end object detection method based on RT-DETR, proposes a CGA-IFI module for intra-scale feature interaction, proposes a DRB-CFFM for cross-scale feature interaction, and uses the EIoU to calculate the bounding box regression loss. Compared to existing methods, our approach demonstrates superior accuracy and offers the advantages of convenient real-time deployment and high-speed processing, making our method highly suitable for scenarios that require both high precision and real-time responsiveness.

5. ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant U23B2006, Grant 92370203, and Grant 62071233, in part by the Jiangsu Provincial Innovation Support Program under Grant No. BZ2023046, in part by the Jiangsu Provincial Natural Science Foundation of China under Grant BK20211570, in part by the Jiangsu Provincial Key Research and Development Program under Grant No. BE2022065-2.

6. REFERENCES

[1] Zhipeng Dong, Mi Wang, Yanli Wang, Ying Zhu, and Zhiqi Zhang, "Object detection in high resolution remote sensing imagery based on convolutional neural networks with suitable object scale features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 2104–2114, 2019

[2] Lei Zhang, Yuehuan Wang, and Yang Huo, "Object detection in high-resolution remote sensing images based on a hard-example-mining network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 10, pp. 8768–8780, 2020

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[4] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu, "Detrs beat yolos on real-time object detection," *arXiv preprint arXiv:2304.08069*, 2023

[5] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14420–14430

[6] Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan, "Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition," *arXiv preprint arXiv:2311.15599*, 2023.

[7] Yi-Fan Zhang, Weiqiang Ren, Zhang Zhang, Zhen Jia, Liang Wang, and Tieniu Tan, "Focal and efficient iou loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, 2022.