MBCDE: Towards Plug-and-Play Spatial-Frequency Fusion for Aerial Object Detection

Haodong Li, Haicheng Qu*
School of Software, Liaoning Technical University
472321734@stu.lntu.edu.cn and quhaicheng@lntu.edu.cn

***Abstract*—The combination of computer vision and remote sensing imaging mechanisms drives the development of aerial object detection technology toward intelligence and precision. Existing methods are limited by the locality of spatial representation and insufficient use of frequency domain information, leading to the annihilation of small object features and sensitivity to noise interference. To address these limitations, we propose a multi-branch cross-domain enhancement (MBCDE) method, which incorporates three novel modules: an edge enhancement frequency-domain module to fuse multi-directional gradient and frequency features in shallow layers for refined texture detail, a Gaussian filtering frequency-domain module to suppress high-frequency noise in deep layers via a heat-diffusion Gaussian kernel, and a multi-branch large-Kernel frequency-domain attention module to mine and enhance shallow features during fusion for improved small object detection. Extensive experiments on the low-altitude drone dataset VisDrone2019 and the high-altitude satellite dataset DIOR demonstrate the effectiveness of MBCDE, showcasing its plug-and-play compatibility across diverse detection architectures.**

***Index Terms*—Object detection, Frequency-domain information, Large kernel convolution, Aerial image**

I. Introduction

Aerial image object detection [1], [2] is not only a core task in the field of computer vision but also a key component of modern earth observation systems. It plays an irreplaceable role in the construction of smart cities [3], emergency response management, public safety [4], and other fields. With the rapid development of drones and remote sensing technology, air-based platforms enable multi-scale continuous observation of surface objects. However, this technology still faces significant challenges in practical applications. The extreme differences in object size, the complexity and variability of the background environment, and the particularity of the imaging perspective together form the main bottleneck of current detection algorithms. These challenges not only restrict the application of existing technologies but also provide a clear research direction for algorithm innovation. Developing a detection paradigm that adapts to the characteristics of aerial images has become a key breakthrough in enhancing earth observation capabilities and intelligence levels.

The current aerial object detection methods can be categorized into three types: one-stage detectors, two-stage detectors, and DETR series detectors. Early two-stage detectors, such as D2Det [5], improve the positioning accuracy of drone scenes through dense feature sampling and bounding box refinement mechanisms. However, due to their large computational workload and slow inference speed, they struggle to meet real-time requirements, leading to a decline in related research in recent years. DETR pioneers the end-to-end use of Transformers for object detection, and subsequent optimization efforts in various scenarios continue to emerge. For instance, although the aerial object detector ARS-DETR [6] enhances global modeling capabilities, its training difficulty and large number of parameters limit its application on edge devices. In contrast, one-stage detectors have become the mainstream choice in the industry due to their high efficiency. The YOLO model, represented by TPH-YOLOv5 [7], introduces the P2 prediction head and attention mechanism to significantly improve small object detection performance, though this also greatly increases computational complexity. However, these classic aerial object detection methods remain limited to spatial feature modeling, overlooking the potential of frequency domain information to enhance texture expression and suppress noise interference.

At the same time, frequency domain learning, as a complementary approach, gradually gains attention in the field of computer vision. In the early stages, Xu et al [8]. and FcaNet [9] applied the frequency domain reweighting mechanism to object detection and instance segmentation tasks. By dynamically adjusting spectral weights and selecting key frequency components, these methods improve the model's ability to perceive fine-grained features. Although such methods demonstrate the effectiveness of frequency domain analysis in computer vision, their application in aerial object detection remains in its infancy. Moreover, existing methods generally fail to systematically model the functional differences between various levels of the network and often indiscriminately eliminate high-frequency components, leading to the loss of shallow edge and texture information. Additionally, the key frequency domain information of small objects in the shallow features of P2 is often overlooked during the multi-scale fusion process. To address these issues, this paper proposes a multi-branch cross-domain enhancement (MBCDE) method and an edge-Gaussian frequency-domain backbone network (EGFNet), with the following contributions:

- We propose MBCDE, which achieves collaborative optimization of spatial and frequency domain features in a plug-and-play manner, effectively improving the performance of object detection from an aerial perspective.
- We design three core modules: edge enhancement frequency-domain module (EFM) to enhance shallow tex-

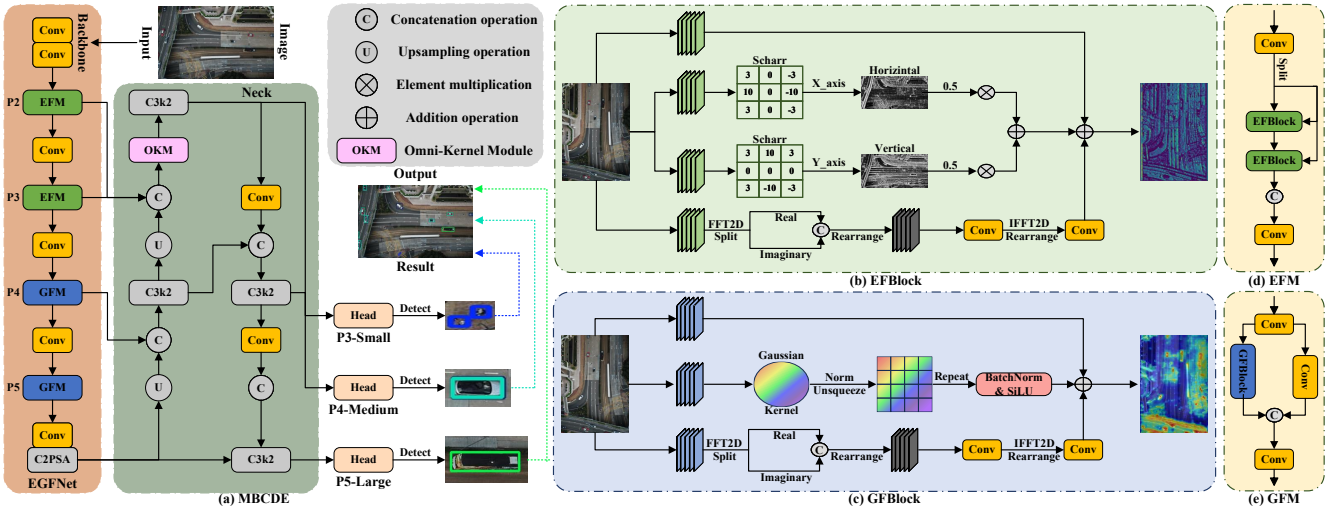

Fig. 1. The overall architecture of the MBCDE. Two layers of EFM are embedded in the shallow layer of the backbone to capture fine-grained contour perception, and two layers of GFM are embedded in the deep layer of the backbone to remove the interference of high-frequency background noise. The features of the P2 layer are reused in the neck network and OKM attention is introduced to enhance the representation ability of objects of different scales.

ture details, Gaussian filtering frequency-domain module (GFM) to suppress deep background noise, and multi-branch large kernel convolution frequency-domain attention module (OKM) to fully exploit and optimize the high-frequency features of the P2 layer, which together constitute a complete multi-scale cross-domain enhancement framework.

- Extensive experiments validate the effectiveness of the method. On the VisDrone2019 and DIOR datasets, MBCDE achieved mAP of 27.8% and 67.2%, respectively, and its generalizability was also verified in various one-stage and two-stage detectors.

II. Related Work

A. Object Detection in Aerial Images

Aerial images can be divided into high-altitude satellite and low-altitude drone perspectives. Their unique imaging characteristics lead to severe challenges in object detection, such as difficulty in identifying small objects, excessive background interference, and dense overlap. Existing methods primarily improve detection performance by optimizing multi-scale fusion strategies or modeling specific problems for different datasets and scenarios. For example, MFEFNet [10] designs an adaptive feature fusion module that allows the model to dynamically evaluate the importance of feature maps at different scales, thereby suppressing unnecessary redundant information. To solve the problem of detail loss and weak semantic information in small objects, MAGC-YOLO [11] proposes a multi-scale fusion method based on a graph neural network with self-attention, enabling the model to detect small objects accurately and efficiently in complex environments. ABFL [12] models the angle of the object bounding box based on the characteristics of the aerial perspective and von Mises distribution to alleviate the issue of discontinuous positioning angles, performing well in dense scenes. EARL [13] addresses the problem of large length and width changes, as well as unbalanced sampling of remote sensing objects, by introducing a dynamic elliptical distribution auxiliary sampling strategy that flexibly selects high-quality samples to fit the object shape while filtering out low-quality object samples. Although these methods have made significant progress in various aerial object detection scenarios, their optimization is still limited to spatial domain representation, overlooking the potential of frequency domain information. Moreover, they lack solutions for optimizing the feature extraction stage and suffer from issues such as insufficient decoupling of object edge-texture information and background noise interference.

B. Frequency-domain Learning for Object Detection

By mapping spatial information to the frequency domain through fast Fourier transform (FFT), different frequency components of the image are effectively separated, helping the model better understand the overall layout and local details of the image. However, these features are often difficult to obtain directly or are easily ignored in single spatial domain processing. In recent years, frequency domain learning has gained increasing attention in object detection tasks. For example, CFIL [24] and SFI [25] propose frequency domain feature extraction mechanisms and frequency domain feature interaction mechanisms to compensate for the lack of spatial information, deepening the network's understanding and perception of object details. NLE-YOLO [26] introduces low-frequency filters to remove high-frequency noise in low-light scene interference. Another study [27] uses discrete wavelet transform to construct a high-frequency coefficient fusion model that efficiently and accurately fuses infrared and visible light images, alleviating the problem of inadequate understanding of complex road conditions by visual sensors and making a practical contribution to the implementation of autonomous

TABLE I
COMPARISON WITH OTHER METHODS OF MEDIUM SIZE ON THE VISDRONE2019 DATASET.

Methods	Year	Aw	Bi	Bu	Ca	Mo	Pd	Po	Ti	Tu	Va	AP_{50}↑	AP↑	FPS↑
DoubleHead-RCNN-R50 [14]	2020	16.8	**20.5**	56.0	75.1	43.8	40.2	33.6	31.4	40.7	48.5	40.7	25.6	14.2
Dynamic-RCNN-R50 [15]	2020	11.0	9.8	47.0	69.3	27.7	31.0	22.6	15.4	28.6	43.4	30.6	19.2	28.0
Sparse-RCNN-R50 [16]	2021	9.3	13.2	36.7	59.8	36.9	31.2	25.8	20.0	27.8	30.1	29.1	16.8	18.3
DAB-DETR-R50 [17]	2022	8.2	11.3	33.5	62.8	36.5	30.4	25.6	20.2	30.5	30.2	28.9	15.0	25.3
DINO-R50 [18]	2023	17.3	17.7	59.5	82.7	48.2	46.4	36.3	32.7	41.3	49.2	43.1	25.5	19.8
DDQ-R50 [19]	2023	18.0	18.3	61.2	83.9	**51.2**	**58.1**	**44.3**	30.7	38.5	49.6	45.4	26.7	21.6
RTMDET-M [20]	2022	14.6	12.1	56.1	75.2	40.4	34.2	28.3	25.1	36.3	41.4	36.4	21.5	36.7
Gold-YOLO-M [21]	2023	18.8	14.4	59.8	80.4	46.6	43.5	34.2	30.3	40.9	45.5	41.4	25.0	58.0
YOLOv12-M [22]	2025	17.5	14.8	60.8	81.0	46.7	45.7	35.1	31.2	40.4	47.5	42.1	25.5	**64.6**
YOLOv11-M [23] (Baseline)	2024	18.5	14.9	59.4	80.5	47.2	45.3	34.1	30.7	41.2	48.0	42.0	25.4	64.3
MBCDE (Ours)	-	**20.1**	18.6	**63.8**	**84.5**	50.9	49.5	38.5	**35.0**	**44.8**	**50.3**	**45.7**	**27.8**	58.7

driving technology. Although these methods highlight the value of the frequency domain in feature decoupling and noise suppression, the strategies they propose fail to fully consider the semantic changes of features in the depth direction of the network. This results in the weakening of edge and texture clues contained in the shallow layer during high-frequency suppression. Additionally, the frequency domain information of the shallow feature layer, which contains a large amount of small object information, is ignored during feature fusion, limiting the model's ability to recognize small objects.

III. METHOD

A. The Overall Architecture of MBCDE

Figure 1 (a) shows the overall architecture of the MBCDE proposed in this paper. In the feature extraction stage, we use the EGFNet backbone network, which includes two shallow EFM and two deep GFM. A OKM is introduced into the neck network, enhancing the model's contextual perception and semantic expression capabilities for objects of different scales by efficiently integrating the multi-scale features of the P2, P3, and P4 layers in the backbone network.

B. Edge Enhancement Frequency-domain Block

In the shallow layers of the backbone network (layers with 128 and 256 channels), we design four-branch EFBlock (shown in Figure 1 (d)) and EFM (shown in Figure 1 (b)) to enhance the backbone network's ability to capture fine-grained features, such as object contours and internal textures, from the original spatial domain, x gradient, y gradient, and frequency domain. For the input image $x(x, y)$, the Scharr edge detection operator effectively captures the horizontal and vertical edges of objects in the image, as shown in the following formula:

$$G_x = \begin{bmatrix} 3 & 0 & -3 \\ 10 & 0 & -10 \\ 3 & 0 & -3 \end{bmatrix}, \quad G_y = \begin{bmatrix} 3 & 10 & 3 \\ 0 & 0 & 0 \\ -3 & -10 & -3 \end{bmatrix} \tag{1}$$

Here, G_x and G_y represent the Scharr operators in the horizontal and vertical directions, respectively, and are used to calculate the horizontal and vertical gradients of the image. Additionally, we use FFT to convert the spatial domain image $x(x, y)$ into the frequency domain image $X(u, v)$, as shown in the following formula:

$$X(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} x(x,y) e^{-j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{2}$$

Here, (x, y) represents the spatial image pixel position, (u, v) denotes the frequency domain coordinates of the frequency domain image, and M and N are the width and height of the image. The frequency domain features after FFT are represented by the real part and the imaginary part:

$$X(u,v) = X_{\text{real}}(u,v) + jX_{\text{imag}}(u,v) \tag{3}$$

Then, the low-frequency and high-frequency information of the image is processed through frequency domain convolution. Finally, the spatial domain image $y(x, y)$ is restored using the inverse fast Fourier transform (IFFT), as shown in the following formula:

$$y(x,y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} X(u,v) e^{j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{4}$$

Finally, the four-branch features are weighted and embedded into the EFM, enabling the model to enhance its ability to extract local details of objects from different perspectives.

C. Gaussian Filtering Frequency-domain Block

In the deep layers of the backbone network (layers with 512 and 1024 channels), we design the three-branch GFBlock (shown in Figure 1 (e)) and GFM (shown in Figure 1 (c)). While retaining the spatial and frequency domain branches to learn the global structure of the image, we introduce a Gaussian filtering to smooth the image and suppress irrelevant background noise. The Gaussian kernel is defined as follows:

$$\text{Gaussian}(x,y) = \frac{1}{2\pi\sigma^2} \exp^{\left(-\frac{x^2+y^2}{2\sigma^2}\right)} \tag{5}$$

Here, the value range of x and y consists of five discrete values, $\{-2, -1, 0, 1, 2\}$, with a Gaussian kernel size of 5 × 5 and a σ value of 1. The design of the frequency domain branch is the same as the EFBlock in Section III-B. Finally, by integrating the fused three-branch GFBlock into GFM, the global semantic consistency of the deep information is enhanced, and the features are aligned for the subsequent multi-scale fusion in the neck network.

TABLE II
COMPARISON WITH OTHER METHODS OF MEDIUM SIZE ON THE DIOR DATASET.

Methods	AL/GF	AO/HA	BE/OP	BK/SH	BI/SA	CI/SO	DM/TC	ES/TS	ET/VE	GF/WM	AP_{50}↑	AP↑
DoubleHead-RCNN	92.8/89.8	87.9/63.0	94.3/71.1	91.6/74.6	56.1/96.1	89.6/65.5	76.0/94.8	93.6/72.1	80.3/51.6	87.2/94.1	81.0	57.3
Dynamic-RCNN	92.8/88.9	88.4/61.5	93.1/68.3	91.0/73.2	55.2/94.8	89.0/63.3	75.0/94.8	67.4/62.5	63.6/48.7	86.7/90.1	79.6	57.2
Sparse-RCNN	91.3/86.4	85.9/50.5	92.9/68.8	88.6/61.7	52.3/93.6	89.6/59.1	75.0/92.9	91.2/65.9	83.9/56.0	86.9/94.8	78.4	54.2
DAB-DETR	89.6/89.8	89.9/39.8	91.0/73.9	86.7/59.8	55.7/96.3	**94.2**/60.6	**85.9**/92.3	92.9/69.5	78.8/57.6	88.2/90.4	79.2	52.9
DINO	95.0/89.6	91.0/69.1	94.1/71.0	90.2/76.1	55.3/94.6	91.7/77.7	76.7/95.4	93.7/66.9	82.4/62.2	87.8/93.3	82.7	59.3
DDQ	96.7/90.6	92.5/67.7	95.4/72.4	91.1/93.8	56.9/95.3	92.5/86.3	79.7/96.4	93.9/72.5	**91.3/70.8**	89.9/94.0	86.0	62.7
RTMDET	96.2/90.8	94.2/77.8	94.4/71.9	**94.3**/77.7	56.7/96.5	93.6/79.2	78.4/96.9	96.3/**78.1**	86.0/65.0	90.8/92.0	85.3	63.1
Gold-YOLO	96.4/90.5	93.3/75.0	95.9/71.9	92.6/94.5	59.3/95.9	92.4/88.8	81.4/97.0	96.2/72.7	89.6/67.9	88.5/94.2	85.7	64.6
YOLOv12	96.0/91.1	92.9/78.0	95.6/73.2	92.8/94.3	60.4/96.0	92.5/88.4	81.7/97.1	96.1/74.1	87.5/68.0	89.8/94.5	86.0	65.3
YOLOv11 (Baseline)	96.4/91.4	92.9/73.1	95.9/73.1	93.3/94.7	60.0/96.8	91.1/88.0	80.3/97.4	96.0/72.2	87.9/67.2	88.6/**95.8**	86.1	65.0
MBCDE (Ours)	**97.3/92.1**	**94.5/78.1**	**96.9/74.0**	94.0/**95.6**	**62.2/97.5**	93.4/**89.7**	83.8/**98.1**	**97.9**/76.7	90.6/70.1	**91.6**/95.7	**88.5**	**67.2**

D. Frequency-domain Attention Module

In order to effectively enhance the model's detection ability for objects of different sizes (especially small objects) and fully tap the complementary information between the spatial domain and the frequency domain, we reuse the feature information of the P2 layer in the neck network and introduce a multi-branch large-kernel convolution frequency-domain attention OKM [28] based on joint modeling of the spatial domain and the frequency domain. OKM consists of three parts: the large-kernel branch, the global branch, and the local branch. The large-kernel branch expands the receptive field through 1 × 31, 31 × 1, and 31 × 31 large-kernel convolutions to obtain contextual information of objects of different shapes. The global branch first introduces the frequency-domain channel attention FCA to weight the features in the channel dimension and adjust the frequency components of the global features to improve the expression of global information. The formula is as follows:

$$X_{\text{FCA}} = \mathcal{F}^{-1}\left(\mathcal{F}(X_{\text{global}}) \odot W_{FCA}^{1\times1}\left(\text{GAP}(X_{\text{global}})\right)\right) \quad (6)$$

Here, $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent FFT and IFFT, respectively; $\odot$ denotes element-wise multiplication; and GAP stands for global average pooling. Next, the frequency-domain spatial attention (FSAM) fine-tunes spatial features in the frequency domain. The formula is as follows:

$$X_{\text{FSAM}} = \mathcal{F}^{-1}\left(\mathcal{F}(W_{FCA}^{1\times1}(X_{\text{FCA}})) \odot W_{FCA}^{2\times1}(X_{\text{FCA}})\right) \quad (7)$$

$W_{FCA}^{1\times1}$ and $W_{FCA}^{2\times1}$ are 1 × 1 and 2 × 1 convolution kernels. The final local branch uses 1 × 1 depth convolution to effectively capture local detail information and prevent information degradation of small objects during feature transfer.

IV. EXPERIMENT AND ANALYSIS

A. Dataset and Experimental Details

The first is VisDrone2019 [29] from the perspective of a low-altitude drone, which contains 10 categories: awning-tricycle (Aw), bicycle (Bi), bus (Bu), car (Ca), motorcycle (Mo), pedestrian (Pd), person (Po), tricycle (Ti), truck (Tu) and van (Va), of which 6471 images are used for training, 548 images are used for validation, and 1610 images are used for testing. The second is DIOR [30] from a high-altitude satellite perspective, including 20 categories: airplane (AL), airport (AO), baseball-field (BE), basketball-court (BK), bridge (BI), chimney (CI), dam (DM), expressway-service-area (ES), expressway-toll-station (ET), golf-field (GF), ground-track-field (GT), harbor (HA), overpass (OP), ship (SH), stadium (SA), storage-tank (SO), tennis-court (TC), trainstation (TS), vehicle (VE), windmill(WM), of which 14077 images are used for training, 4694 for validation, and 4692 for testing. We use an RTX 3090 graphics card for experiments, employ YOLOv11-M [23] as the baseline method, and train for 150 epochs on both datasets with a batch size of 8 and a learning rate of 0.01. We use the F1 score, AP_{50}, AP, and FPS as evaluation metrics for analysis.

B. Comparison with Other Advanced Methods

As shown in Table I, we compare MBCDE with recent object detection methods on the VisDrone2019 dataset. The first three methods are two-stage object detection methods, the middle three are DETR series object detection methods, and the last three are one-stage object detection methods. Experimental results show that MBCDE achieves the highest AP_{50} scores in 6 out of 10 categories, with an average AP_{50} of 45.7% and an average AP of 27.8%, both the highest among all methods. This demonstrates that MBCDE's mining and utilization of frequency domain information effectively compensates for the shortcomings of spatial domain information, and cross-domain collaboration enhances the overall performance of the model. Additionally, MBCDE, like other one-stage object detection methods, generally exhibits efficient computing speed, ranking third among the 11 methods. This indicates that MBCDE has the potential to be deployed in edge devices for real-time detection.

We also compare the same method on the DIOR dataset, with the results shown in Table II. First, MBCDE achieves the highest detection accuracy in 13 out of 20 object categories, covering objects of different sizes, such as large objects GF-92.1%, medium objects BE-96.6%, and small objects SH-95.6%. This demonstrates that MBCDE has excellent ability to identify objects of various sizes. Secondly, MBCDE achieves 88.5% and 67.2% in AP_{50} and AP metrics, respectively, which

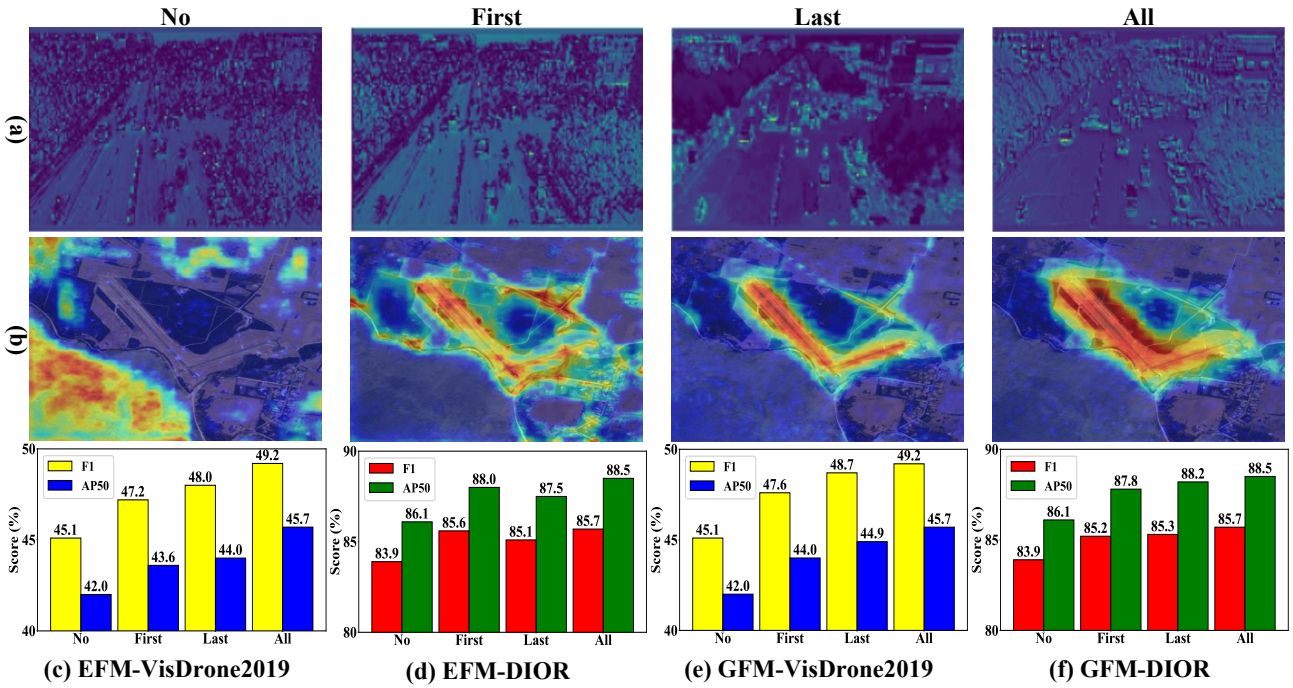

Fig. 2. Ablation study for EGFNet. (a) and (b) are visualizations of the insertion locations of EFM and GFM, respectively. (c), (d) Ablation of the insertion location of EFM. (e), (f) Ablation of the insertion location of GFM. The horizontal axis names in (c)-(f) are No, First, Last, and All from front to back, respectively, and the F1 score and AP_{50} are compared.

TABLE III
ABLATION EXPERIMENTS ON VISDRONE2019 AND DIOR DATASETS.

EFM	GFM	OKM	VisDrone2019			DIOR		
			F1↑	AP_{50}↑	AP↑	F1↑	AP_{50}↑	AP↑
✗	✗	✗	45.1	42.0	25.4	83.9	86.1	65.0
✓	✗	✗	46.7	42.6	26.1	84.7	87.1	65.4
✗	✓	✗	46.5	42.8	26.1	85.0	87.3	66.4
✗	✗	✓	47.9	43.7	26.7	85.0	87.6	66.0
✓	✓	✗	48.2	44.3	27.2	**85.8**	87.9	66.5
✓	✓	✓	**49.2**	**45.7**	**27.8**	85.7	**88.5**	**67.2**

are 2.4% and 2.2% higher than the baseline method. This further shows that, compared to the baseline method, MBCDE adapts well to high-altitude satellite perspectives for object detection and maintains stability in complex scenes.

C. Ablation Study

To verify the effectiveness of each component of MBCDE, we conduct ablation studies on two datasets. As shown in Table III, on the VisDrone2019 dataset, the addition of OKM significantly improves the detection results, with increases of 2.8%, 1.7%, and 1.3% in F1 score, AP_{50}, and AP, respectively, compared to the baseline. This reveals that OKM's strategy of reusing the P2 feature layer and the multi-branch large-kernel convolution scheme enhances the model's ability to handle extreme scenes, such as dense overlap, and effectively transmits information about small objects. On the DIOR dataset, models using EFM and GFM alone improve AP_{50} by 1.0% and 1.2%, respectively, compared to the baseline, and improve by 1.7% when used together. This shows that the shallow EFBlock and deep GFBlock enhance object texture recognition and reduce high-frequency noise interference, respectively, and can work together flexibly and effectively. Finally, MBCDE, combining all its components, achieves the best performance on both the AP_{50} and AP metrics.

1) Effectiveness Experiments within EGFNet Backbone Network: As shown in Figure 2, we conduct an ablation study on the insertion position and number of modules within EGFNet. Overall, EFM and GFM achieve the best insertion effect at all positions, and inserting a single layer performs better than not inserting any. This demonstrates that spatial-frequency domain collaboration at different levels can leverage multi-angle information and improve the model's understanding of complex scenes. Additionally, inserting modules in the deep layers generally performs better than in the shallow layers, reflecting that the deep layers contain richer details and semantic information, which better support global feature modeling.

2) Effectiveness Experiments within OKM Module: As shown in Table IV, we compare OKM with several attention algorithms based on similar concepts. It can be observed that OKM, which integrates multi-branch, large-kernel convolution, and frequency-domain information, achieves the best overall performance on both datasets. Its F1 score improves by 0.6%, 1.0%, 1.1%, 0.6%, 0.5%, and 0.7%, respectively, compared to other attention algorithms that rely on only one concept. This demonstrates that OKM has stronger adaptability

TABLE IV

ABLATION EXPERIMENTS OF SIMILAR ATTENTION AND OKM ATTENTION, WHERE MB REPRESENTS MULTI-BRANCH, LK REPRESENTS LARGE-KERNEL CNN, AND FD REPRESENTS FREQUENCY DOMAIN.

Attention	MB	LK	FD	VisDrone2019			DIOR		
				F1↑	AP_{50}↑	AP↑	F1↑	AP_{50}↑	AP↑
DBB [31]	✓	✗	✗	47.6	44.1	26.7	85.1	88.2	66.5
LSKA [32]	✗	✓	✗	48.2	44.5	26.8	85.2	88.2	67.0
MSCA [9]	✗	✗	✓	48.1	44.6	27.0	85.0	88.0	66.2
OKM [28]	✓	✓	✓	**49.2**	**45.7**	**27.8**	**85.7**	**88.5**	**67.2**

TABLE V

GENERALITY EXPERIMENTS ON DIFFERENT ONE-STAGE AND TWO-STAGE DETECTION METHODS MBCDE METHOD ON THE VISDRONE-2019 AND DIOR DATASETS.

Method	VisDrone2019		DIOR	
	AP_{50}↑	AP↑	AP_{50}↑	AP↑
Dynamic-RCNN [15]	30.6	19.2	79.6	57.2
+MBCDE	33.4 (+2.8)	20.9 (+1.7)	82.0 (+2.4)	59.1 (+1.9)
DoubleHead-RCNN [14]	40.7	25.6	81.0	57.3
+MBCDE	43.9 (+3.2)	26.8 (+1.2)	83.8 (+2.8)	59.3 (+2.0)
RTMDET [20]	36.4	21.5	85.3	63.1
+MBCDE	38.6 (+2.2)	22.5 (+1.0)	87.0 (+1.7)	63.8 (+0.7)
YOLOv12 [22]	42.0	25.4	86.0	65.3
+MBCDE	45.5 (+3.5)	27.2 (+1.8)	88.2 (+2.2)	66.5 (+1.2)

and robustness for aerial object detection tasks.

D. Universality Study of MBCDE Method

Our MBCDE method can be flexibly and plug-and-play into other detectors. As shown in Table V, integrating the MBCDE consistently improves AP_{50} and AP metrics on both two-stage (Dynamic-RCNN, DoubleHead-RCNN) and two one-stage (RTMDET, YOLOv12) methods. Experiments demonstrate that the mechanisms of MBCDE in enhancing texture, suppressing noise, and optimizing feature fusion are universally effective in improving the generalization performance of detectors with different architectures.

E. Visualization

As shown in Figure 3, we compare the detection results of the baseline and MBCDE on two datasets with real labeled data. To highlight the differences more clearly, we use different colors. It can be seen that MBCDE significantly reduces missed detections and false detections across multiple scenes, such as intersections, highway toll booths, and stadiums. Additionally, in the dense scene in the fourth column, MBCDE accurately identifies small objects and ships without being disturbed by complex backgrounds. These visualization results fully verify the advantages and reliability of MBCDE in practical applications.

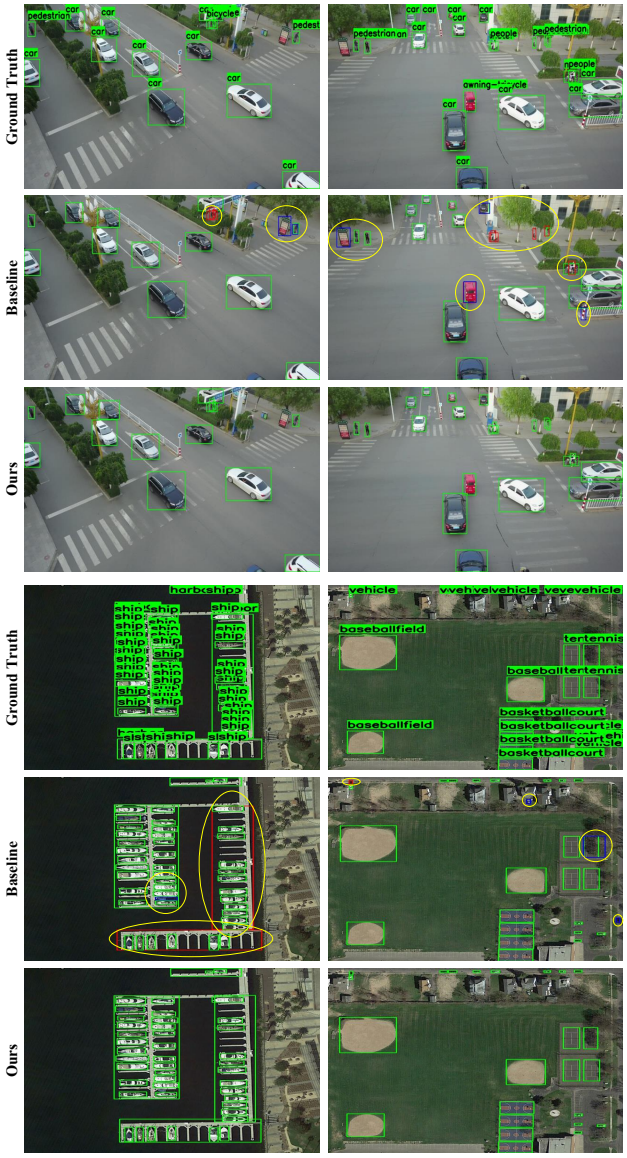

Fig. 3. The detection results of the two datasets are visualized. The top three rows are VisDrone-2019, and the bottom three rows are DIOR. Red boxes indicate missed objects, blue boxes indicate false detections, green boxes indicate correct object detections, and yellow circles highlight errors.

V. CONCLUSION

This paper proposed a multi-branch cross-domain enhancement detection method, MBCDE, which integrates spatial and frequency domain modeling, providing a new, reliable, and efficient solution for aerial object detection. By introducing the edge enhancement frequency-domain module EFM in the shallow layer of the backbone network, embedding the Gaussian filtering frequency-domain module GFM in the deep layer of the backbone network, and designing a multi-branch large-kernel convolution frequency-domain attention module in the neck network to optimize the feature fusion process, MBCDE effectively improved the model's adaptability to complex scenes and scale-varying situations. Extensive experiments on two typical aerial image datasets, VisDrone2019 and DIOR, demonstrated the advantages of this method in terms of detection accuracy and versatility, while also exhibiting good computational efficiency, making it suitable for deployment on edge devices. In the future, we plan to further optimize its architecture and apply this method to video object detection, cross-modal object detection, and other fields to explore its

potential.

ACKNOWLEDGMENT

This paper was supported by the National Natural Science Foundation of China (GrantNo. 42271409) and the Scientific Research Foundation of the Higher Education Institutions of Liaoning Province (Grant No. LJKMZ20220699). Thank you for the GPU resource support project at Liaoning Technical University.

REFERENCES

[1] ZhiLin Gao, QiXiang Meng, JinTao Wang, and FanLiang Bu. Lao-yolo: improved yolov10 model for lightweight aerial object detection. *Signal, Image and Video Processing*, 19(6):451, 2025.

[2] Zhihong Zhao and Peng He. Yolo-mamba: object detection method for infrared aerial images. *Signal, Image and Video Processing*, 18(12):8793–8803, 2024.

[3] V Tiruvikraman, D Selvakumar, and P Vijayakumar. Real-time smart surveillance and enforcement for dust throw detection and identity recognition using yolo 12 and sa-facexnet: V. tiruvikraman et al. *Signal, Image and Video Processing*, 19(12):983, 2025.

[4] M Vijayalakshmi, AJ Roshan Jose, GK Vishal, and N Vishwanath. Safety equipment detection using yolov8. *Signal, Image and Video Processing*, 19(12):994, 2025.

[5] Jiale Cao, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. D2det: Towards high quality object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11485–11494, 2020.

[6] Ying Zeng, Yushi Chen, Xue Yang, Qingyun Li, and Junchi Yan. Ars-detr: Aspect ratio-sensitive detection transformer for aerial oriented object detection. *IEEE transactions on geoscience and remote sensing*, 62:1–15, 2024.

[7] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2778–2788, 2021.

[8] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1740–1749, 2020.

[9] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021.

[10] Liming Zhou, Shuai Zhao, Ziye Wan, Yang Liu, Yadi Wang, and Xianyu Zuo. Mfefnet: A multi-scale feature information extraction and fusion network for multi-scale object detection in uav aerial images. *Drones*, 8(5):186, 2024.

[11] Jianquan Ouyang and Lingtao Zeng. Magc-yolo: Small object detection in remote sensing images based on multi-scale attention and graph convolution. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[12] Zifei Zhao and Shengyang Li. Abfl: Angular boundary discontinuity free loss for arbitrary oriented object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[13] Jian Guan, Mingjie Xie, Youtian Lin, Guangjun He, and Pengming Feng. Earl: An elliptical distribution aided adaptive rotation label assignment for oriented object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[14] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10186–10195, 2020.

[15] Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. Dynamic r-cnn: Towards high quality object detection via dynamic training. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 260–275. Springer, 2020.

[16] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021.

[17] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *International Conference on Learning Representations*.

[18] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*.

[19] Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7329–7338, June 2023.

[20] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022.

[21] Chengcheng Wang, Wei He, Ying Nie, Jianyuan Guo, Chuanjian Liu, Yunhe Wang, and Kai Han. Gold-yolo: Efficient object detector via gather-and-distribute mechanism. *Advances in Neural Information Processing Systems*, 36:51094–51112, 2023.

[22] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

[23] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023.

[24] Weijie Weng, Weiming Lin, Feng Lin, Junchi Ren, and Fei Shen. A novel cross frequency-domain interaction learning for aerial oriented object detection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 292–305. Springer, 2023.

[25] Weijie Weng, Mengwan Wei, Junchi Ren, and Fei Shen. Enhancing aerial object detection with selective frequency interaction network. *IEEE Transactions on Artificial Intelligence*, 2024.

[26] Daxin Peng, Wei Ding, and Tong Zhen. A novel low light object detection method based on the yolov5 fusion feature enhancement. *Scientific reports*, 14(1):4486, 2024.

[27] Chenwu Wang, Junsheng Wu, Aiqing Fang, Zhixiang Zhu, Pei Wang, and Hao Chen. An efficient frequency domain fusion network of infrared and visible images. *Engineering Applications of Artificial Intelligence*, 133:108013, 2024.

[28] Yuning Cui, Wenqi Ren, and Alois Knoll. Omni-kernel network for image restoration. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 1426–1434, 2024.

[29] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[30] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.

[31] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Diverse branch block: Building a convolution as an inception-like unit. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10886–10895, 2021.

[32] Kin Wai Lau, Lai-Man Po, and Yasar Abbas Ur Rehman. Large separable kernel attention: Rethinking the large kernel attention design in cnn. *Expert Systems with Applications*, 236:121352, 2024.