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

FE-DETR: A Fourier-Enhanced, Edge-Aware Framework for UAV-Based Remote Sensing Object Detection

Guocheng An[1][0009-0007-6189-9733], Wenbin Liu*[2][0009-0008-6626-0194] and Pengzhan Sheng[1]

[1] Artificial Intelligence Research Institute of Shanghai Huaxun Network System Co., LTD., Chengdu, 610074, China
{Anguocheng, shengpengzhan}@eccom.com.cn
[2] Shanghai Jiao Tong University, Shanghai, 201100, China
bilibili@sjtu.edu.cn*
*Corresponding author

Abstract. Aerial object detection in drone-based imagery presents unique challenges including sub-20px targets, motion blur, dense occlusions, and complex backgrounds. Existing methods struggle to harmonize spectral sensitivity with spatial precision while maintaining real-time efficiency. This paper proposes FE-DETR, an optimized end-to-end framework integrating Fourier-enhanced processing, adaptive attention, and edge-aware fusion. First, the Fourier-Enhanced Feature Fusion (FFF) module synergizes global frequency analysis with multi-scale dilated convolutions, amplifying faint object signatures while preserving structural integrity under motion blur. Second, the Adaptive WL-GH Attention dynamically allocates computation between local window attention and global cross-window reasoning via learnable feature statistics. Third, the Edge-Enhanced Multi-Scale Fusion neck (E^2MF) embeds physics-inspired Sobel operators to maintain structural coherence in occlusion-heavy scenes. Evaluated on VisDrone2019, FE-DETR achieves state-of-the-art 50.4% mAP50 and 31.1% mAP50-95 with 17.3M parameters and 54.9G FLOPs. Ablation studies confirm the complementary benefits of spectral-spatial fusion and edge-aware processing. The framework demonstrates robust performance across illumination variations and scale disparities, offering practical efficiency for UAV deployment. Code will be released at https://github.com/Avery5233/FE-DETR.

Keywords: Aerial Object Detection, RT-DETR, Fourier-Enhanced Feature Fusion, Edge-Aware Neck.

1 Introduction

Remote sensing object detection, a critical branch of computer vision, focuses on automatically identifying and localizing targets (e.g., vehicles, pedestrian) in aerial imagery captured by satellites or unmanned aerial vehicles (UAVs). This technology has become indispensable in urban planning, environmental monitoring, agricultural management, and disaster response due to its unique aerial perspective. While deep learning

approaches have gradually replaced traditional methods relying on handcrafted features [1,2], four persistent challenges in aerial imagery remain inadequately addressed: detection of numerous sub-20px small targets under low resolutions; severe occlusions in crowded scenarios; interference from complex background textures, and motion blur caused by UAV platform vibrations - as evidenced by the VisDrone2019 benchmark [3].

Current CNN architectures struggle with these aerial-specific challenges due to inherent limitations. While multi-scale fusion designs like FPN [4] and specialized loss functions [5] have improved small object detection, two-stage detectors suffer prohibitive computational costs for real-time UAV deployment. Single-stage models (YOLO [6], SSD [7]) introduce anchor bias and NMS-induced latency, particularly problematic in dense target scenarios. Recent Transformer-based approaches like RT-DETR [8] eliminate anchor dependencies through self-attention but still underperform on sub-20px targets due to inadequate high-frequency feature preservation and edge smearing in occlusion scenarios.

Recent UAV detection research reveals persistent challenges in harmonizing spectral sensitivity and spatial coherence. UAV-DETR [9] pioneers' frequency-domain processing but struggles with parameter redundancy from decoupled spectral-spatial branches, limiting real-time deployment. ARFP [10] advances adaptive feature pyramids yet encounters gradient decay in deep recursion layers, impairing small object discernment. HIC-YOLOv5 [11] strengthens dense target detection through hierarchical fusion but relies on static receptive fields that inadequately resolve severe occlusions. MSFE-YOLO [12] enhances multi-scale awareness through parallel convolutions at the cost of increased computational complexity, particularly in motion-blurred scenarios. While YOLOv12 [13] integrates attention mechanisms, its fixed spectral-channel interactions prove suboptimal for preserving structural edges under illumination variations. These approaches collectively face fundamental trade-offs between global spectral awareness and local spatial precision, often prioritizing architectural complexity over synergistic feature integration.

Our work addresses these gaps by proposing an optimized RT-DETR framework specifically tailored for aerial target detection called FE-DETR. Key innovations include:

1. Fourier-Enhanced Feature Fusion (FFF): Combines global frequency analysis with multi-scale spatial processing to amplify faint object signatures while suppressing background noise. By preserving phase information during FFT reconstruction, we overcome motion blur challenges that degrade conventional convolutions.

2. Adaptive WL-GH Attention (WL-GH): Introduces dynamic resource allocation between local window attention and global cross-window reasoning, automatically adapting to scene complexity through learnable feature statistics.

3. Edge-Enhanced Multi-Scale Fusion neck (E^2MF): Embeds physics-inspired Sobel operators in the feature pyramid to maintain structural integrity under heavy occlusions, with adaptive fusion weights for multi-scale edge preservation.

2 Related Work

2.1 Prior-Based Object Detection Methods

Two-stage detectors, exemplified by Faster R-CNN [14], generate region proposals before classification and regression. While effective for multi-class aerial targets, their computational complexity limits real-time deployment. Cai et al. [15] introduced Cascade R-CNN to refine detection accuracy through iterative IoU thresholding, yet inference speed remains impractical for large-scale remote sensing applications.

Single-stage models like YOLOv9 [16] prioritize speed by unifying localization and classification. However, their reliance on predefined anchors and NMS introduces bias in cluttered scenes. Yang et al. [17] enhanced YOLOv3 with feature fusion for small targets, but performance degrades under low-resolution conditions common in UAV imagery.

2.2 End-to-End Detection Frameworks

The emergence of Transformer-based detectors has addressed prior dependency issues. DETR [18] pioneers set prediction via self-attention, eliminating anchors and NMS. Despite superior global modeling, its quadratic computation complexity hinders scalability. Deformable DETR [19] mitigates this through deformable attention mechanisms, focusing computation on sparse spatial locations. While efficient, it struggles with sub-pixel-scale targets due to insufficient feature granularity.

RT-DETR [8] advances real-time performance via hybrid CNN-Transformer designs and optimized encoder-decoder interactions. By decoupling multi-scale features and integrating hybrid channels, it is faster than Deformable DETR on aerial datasets. Nevertheless, its fixed-scale attention windows inadequately capture micron-level features of distant vehicles or pedestrians in satellite imagery. Recent approaches such as Drone-DETR [20] employ enhanced dual-path feature fusion and shallow feature enrichment to improve small object recognition accuracy. However, the integration of deformable convolutions and multi-scale attention mechanisms introduces computational overhead, escalating model complexity.

2.3 Small Target Detection in Remote Sensing

Recent advances in aerial detection address specific challenges through architectural refinements. Hu et al. [21] enhanced YOLOv7 with E-ELAN and model scaling, achieving 77% mAP on DOTA. However, its fixed receptive fields inadequately resolve sub-20px targets in motion-blurred scenes. Zhang et al. [22] proposed FFCA-YOLO with multiscale fusion (FFM) and spatial context modules (SCAM), yet its computational-heavy FEM struggles with real-time UAV deployment. Li et al. [23] integrated super-resolution with YOLO, boosting small-target precision via Swin Transformers but introducing latency from SR preprocessing. While these methods improve specific aspects, they neglect synergistic spectral-spatial integration: [21,22] lack

motion-blur resilience due to spatial-only convolutions; the SR of [23] induce latency conflicts with real-time needs.

3 FE-DETR

To address the unique challenges of drone-based aerial imagery—small targets, motion blur, occlusions, and illumination variations—we propose Fourier-Edge DETR (FE-DETR) which integrating global frequency analysis, adaptive attention, and edge-aware fusion. The architecture combines Fourier-enhanced spectral processing with multi-scale spatial convolutions to amplify faint object signatures, dynamically balances local precision and global context through adaptive attention, and preserves structural integrity under occlusions via physics-inspired edge enhancement. By harmonizing these components through frequency-guided feature recalibration and content-aware resource allocation, the framework achieves robust detection across diverse aerial scenarios while maintaining real-time efficiency critical for UAV deployment. The architecture of FE-DETR is shown in **Fig. 1**.

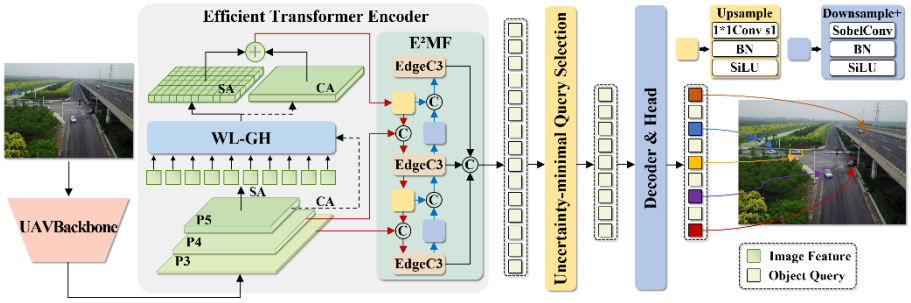

Fig. 1. The architecture of FE-DETR. The EdgeC3 is improved RepC3 by using SobelConv.

3.1 Fourier-Enhanced Feature Fusion Module

Fig. 2 illustrates the architecture of our proposed Fourier-Enhanced Feature Fusion (FFF) module, which replaces standard C2f module to build an enhanced backbone (UAVBackbone) for specifically addressing the unique challenges in aerial images mention in introduction, especially for small targets, motion blur and occlusions. Traditional convolutional modules struggle with these issues due to their limited receptive fields and local feature bias. Our FFF module overcomes these limitations through a dual-branch architecture that synergizes frequency-domain global analysis and multi-scale spatial processing, achieving superior accuracy-speed balance for drone-based detection.

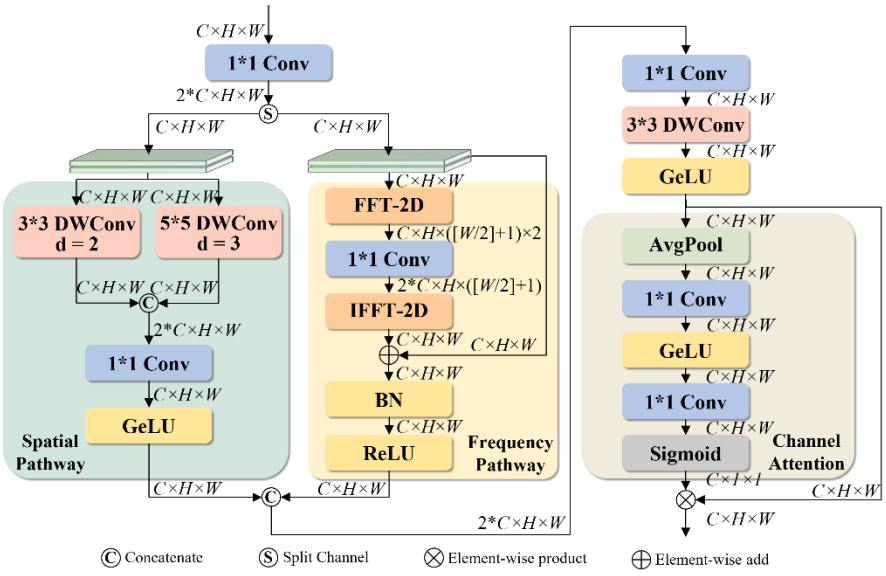

Fig. 2. The architecture of FFF Module.

Dual-Branch Feature Decomposition. The input feature map $X \in \mathbb{R}^{C\times H\times W}$ undergoes channel-wise splitting into two complementary processing streams. The frequency pathway addresses small target detection and motion blur through spectral analysis. We first apply 2D Fast Fourier Transform (FFT) with orthonormal normalization:

$$\mathcal{F}(X) = \mathrm{FFT}(X) \in \mathbb{C}^{C\times H\times \lfloor W/2\rfloor+1}. \tag{1}$$

This transformation projects spatial features into the frequency domain where high-frequency components correspond to fine-grained details (critical for sub-20px targets) while low-frequency components represent global structures. To handle motion blur prevalent in UAV-captured images, we preserve phase information during the inverse FFT (iFFT) reconstruction, as phase components contain crucial structural details about object boundaries. The real and imaginary parts of the complex tensor are processed through parallel 1×1 convolutions followed by batch normalization and ReLU activation:

$$X_{\text{freq}} = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{Re}\big(\mathcal{F}(X)\big) \oplus \mathrm{Im}\big(\mathcal{F}(X)\big)\right)\right)\right), \tag{2}$$

where $\oplus$ denotes element-wise addition. This design enables the network to learn optimal combinations of magnitude and phase information, effectively enhancing faint edges of small objects while suppressing high-frequency noise from complex backgrounds. The spatial pathway tackles occlusions and scale variations through asymmetric dilated convolutions. We employ depthwise convolutions (DWConv) with different dilation rates to maintain computational efficiency:

$$X_{\text{local}} = \text{Concat}\left(\text{DWConv}_{3\times3}^{d=2}(X), \text{DWConv}_{5\times5}^{d=3}(X)\right), \tag{3}$$

where d represents dilation rates. The 3×3 convolution with $d = 2$ expands the receptive field to 7×7, helping separate overlapping instances in crowded scenes, while the 5×5 convolution with $d = 3$ achieves a 13×13 effective receptive field to capture contextual cues around occluded targets. According to statistics, over 40% of targets in VisDrone's train set are partially occluded. This multi-scale design proves particularly effective for dense pedestrian scenarios in aerial images.

Adaptive Feature Fusion. The frequency and spatial features are fused through a novel Frequency-Guided Attention (FGA) mechanism that dynamically emphasizes task-critical components. First, we concatenate the two feature streams along the channel dimension and halve the number of channels via CAConv:

$$X_{\text{cat}} = \text{CAConv}\left(\text{Concat}\left(X_{\text{freq}}, X_{\text{local}}\right)\right) \in \mathbb{R}^{C\times H\times W}. \tag{4}$$

A squeeze-excitation block then generates channel-wise attention weights conditioned on the frequency characteristics:

$$\alpha = \sigma\left(\text{MLP}\left(\text{GAP}(X_{\text{cat}})\right)\right) \in \mathbb{R}^{C\times1\times1}, \tag{5}$$

where GAP denotes global average pooling and σ is the sigmoid function. The final fused features are computed as:

$$X_{\text{fused}} = \alpha \odot X_{\text{cat}} + X_{\text{cat}}\,. \tag{6}$$

This attention mechanism prioritizes high-frequency components essential for small object detection while suppressing irrelevant background textures. For example, in VisDrone's highway scenes, FGA can effectively increase the activation weights for high-frequency vehicle edges in complex road markings.

To further enhance robustness against illumination variations common in aerial imagery (e.g., shadows under building or sun glare), we incorporate phase-aware feature rectification before iFFT reconstruction. By maintaining the consistency between magnitude and phase components during frequency-domain processing, our module preserves structural integrity under low-light conditions where traditional convolutions suffer from gradient vanishing.

The FFF module replaces all C2f blocks in YOLOv8's backbone, forming our efficient UAV-backbone. Implementation leverages PyTorch's native FFT/iFFT operators with Hermitian symmetry preservation for numerical stability. Depthwise convolutions maintain computational efficiency following MobileNet's design principles [24], while the CSP (Cross Stage Partial) structure ensures gradient flow optimization [25]. Our method uniquely optimizes the interaction between complex spectral features and spatial context, achieving superior performance on aerial datasets without compromising inference speed.

3.2 Adaptive WL-GH Attention for Aerial Scene Understanding

To address the dual challenges of small object detection and occlusion reasoning in aerial imagery, we propose a Window Low-level and Global High-level (WL-GH) attention mechanism similar to SpectFormer [26] with dynamic level adaptation, the architecture of WL-GH Attention is shown in **Fig. 3**. This architecture disentangles low-level/high-level visual patterns through parallel processing streams, then automatically adjusts their relative contributions using learnable parameters.

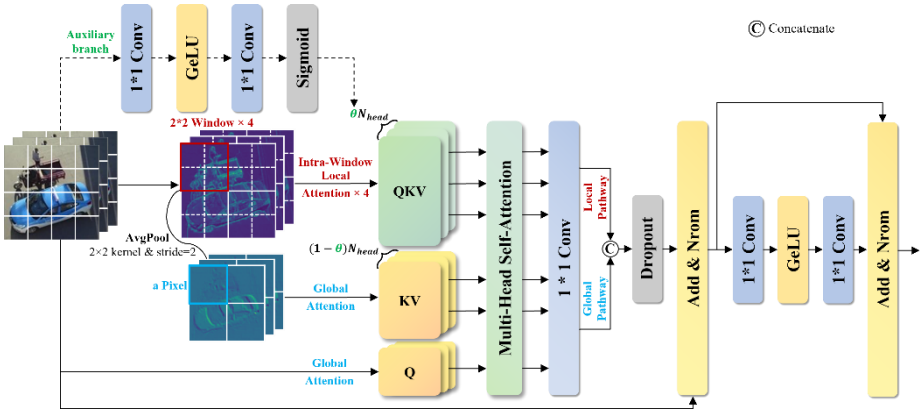

Fig. 3. The architecture of WL-GH Attention.

The WL-GH module processes input features through two complementary pathways. The low-level branch preserves fine spatial details essential for sub-20px targets through constrained self-attention within non-overlapping 2×2 windows. This local attention mechanism reduces computational complexity from $O((HW)^2)$ to $O\left(\frac{HW}{s^2}(s^2)^2\right)$ while maintaining precise localization capabilities for small objects, where s denotes the size of the window. Conversely, the high-level branch employs cross-window attention on spatially pooled features (stride=s), establishing global dependencies to resolve occlusion patterns and suppress background clutter through structural coherence. The attention shared by high-level and low-level branches is as follows:

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{7}$$

A novel alpha-predictor auxiliary net enables dynamic resource allocation between branches. Implemented as a lightweight convolutional network, this module analyzes input feature statistics to predict channel-wise adaptation coefficients $\alpha \in [0,1]$. The predicted alpha automatically adjusts the head allocation ratio between low-level and high-level streams through differentiable weight redistribution:

$$\alpha = \text{Sigmoid}\left(\mathcal{F}_{1\times1}\left(\text{GeLU}\left(\mathcal{F}_{1\times1}(X)\right)\right)\right), \tag{8}$$

where $\mathcal{F}_{1\times1}$ denotes 1×1 convolutions for dimension transformation. This adaptive mechanism optimizes the trade-off between local precision and global context awareness based on input content, particularly beneficial for diverse scenarios ranging from sparse open areas to densely packed urban scenes.

The dual-stream outputs undergo channel-wise concatenation and linear projection to form final features. Integration with standard Vision Transformer architecture follows post-normalization design:

$$X'' = \text{LayerNorm}\left(X' + \text{Dropout}\big(\text{FFN}(X')\big)\right), \tag{9}$$

where FFN contains position-wise feed-forward networks. The content-aware level adaptation is particularly effective in preserving edge details of miniature objects while suppressing low-level noise in complex backgrounds.

3.3 Edge-Enhanced Multi-Scale Feature Fusion Neck

We enhance the RT-DETR neck by replacing 3×3 convolutions in down-sampling operations and RepC3 blocks within the neck with our SobelConv operators and append three consecutive Edge Fusion modules at the neck's terminal to progressively refine multi-scale edge representations. The enhanced neck is called Edge-Enhanced Multi-Scale Feature Fusion Neck. E²MF addresses the critical challenge of preserving structural edge information in occlusion-heavy UAV imagery through two synergistic components: SobelConv for edge-aware feature extraction and Edge Fusion module for adaptive feature integration. The architecture of SobelConv and Edge Fusion module is illustrated in **Fig. 4**.

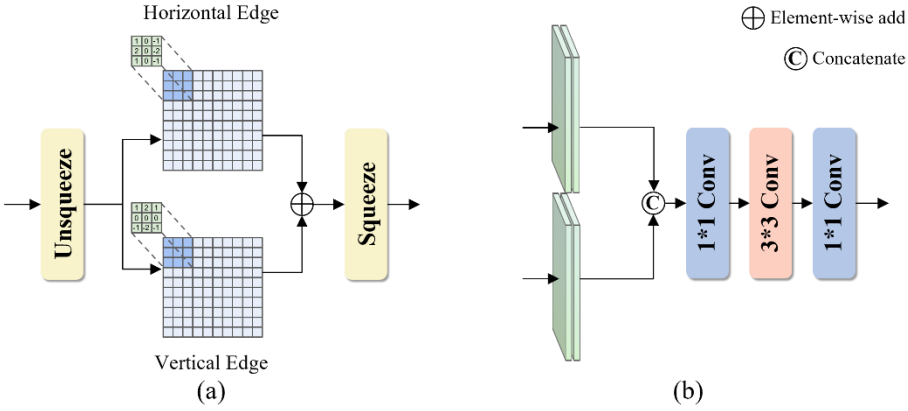

Fig. 4. The architecture of improvement components in neck. (a) SobelConv, (b) Edge Fusion.

The core innovation lies in the SobelConv operator, which implements learnable 3D convolutions with fixed Sobel kernels to explicitly model gradient patterns. Given an

input feature map $X \in \mathbb{R}^{C\times H\times W}$, the horizontal and vertical edge responses are computed through:

$$\mathbf{G}_x = \sum_{i,j=-1}^{1} \mathbf{W}_x^{(i+1,j+1)} * \mathbf{X}_{pad}^{(i,j)}, \quad \mathbf{G}_y = \sum_{i,j=-1}^{1} \mathbf{W}_y^{(i+1,j+1)} * \mathbf{X}_{pad}^{(i,j)}, \tag{10}$$

where W_x and $\boldsymbol{W}_y$ denote the predefined Sobel kernels expanded across channels through depthwise convolution. The final edge-enhanced features $E = G_x + G_y$ capture directional intensity discontinuities while maintaining channel-wise independence.

The Edge Fusion module synthesizes hierarchical edge information through a cascaded transformation pipeline. By concatenating multi-scale edge features $E_1 \oplus E_2 \oplus \cdots \oplus E_K$ along the channel dimension, the module first compresses the combined high-dimensional features into a compact representation ($\mathbb{R}^{KC} \to \mathbb{R}^{C/2}$) using a 1×1 convolution. Subsequent spatial context aggregation is performed via a 3×3 depthwise separable convolution, which enhances local geometric relationships while minimizing computational overhead. Finally, a 1×1 projection layer recalibrates the feature responses to match the target output dimension ($\mathbb{R}^{C/2} \to \mathbb{R}^{C}$), ensuring compatibility with feature hierarchy.

This architecture enables two critical capabilities: Explicit edge guidance through physics-inspired operators counteracts the feature smearing caused by heavy occlusions; and the scale-pyramid fusion mechanism adaptively weights edge responses based on object sizes - a crucial property for UAV scenes containing both large vehicles and small pedestrians.

4 Experiments

4.1 Datasets

To validate FE-DETR's effectiveness, we conducted experiments on VisDrone 2019 datasets [3]. VisDrone 2019 contains UAV-captured images with challenges including uneven illumination and object occlusion, covering 10 categories such as pedestrians, cars, and buses. The official data splits were adopted, with evaluation conducted on the validation set. This dataset, widely used in UAV vision research, offers a realistic testbed for our model. Its images, captured in diverse scenarios, provide rich variations in object appearances and interactions. The extensive annotations facilitate comprehensive evaluation of detection performance across different scales and categories.

4.2 Experiment Setup

The experimental environment utilizes a Windows 11 workstation running Python 3.8.16 and PyTorch 2.0.1, optimized for CUDA 12.1 acceleration. Hardware configuration combines cutting-edge consumer components: A 13th Gen Intel® Core™ i7-13700K processor (16 cores/24 threads, 5.4GHz turbo) handles serial computations, while an NVIDIA RTX 4090 GPU with 24GB GDDR6X VRAM accelerates parallel operations. The GPU's 3rd-generation RT cores and 4th-generation Tensor Cores

enable mixed-precision training through PyTorch's AMP (Automatic Mixed Precision) module.

As detailed in **Table 1**, the experimental configuration maintains fixed hyperparameters across all trials. Reproducibility measures include PyTorch's deterministic algorithms (enabled via torch.backends.cudnn.deterministic) and fixed random seeds across Python, NumPy, and CUDA environments. Thermal monitoring confirmed consistent GPU operation at 65°C±3°C during sustained loads. All trials completed within 5% temporal variance of projected durations, demonstrating hardware stability across extended training sessions.

Table 1. Hyperparameter configuration

Hyperparameter	Value
Input size	640×640
Batch size	8
Training epochs	300
Optimizer	AdamW
Initial learning rate	0.0001
Learning rate factor	0.01
Momentum	0.9
Warmup steps	2000

4.3 Comparison Experiments with State of the Art

We evaluate FE-DETR against state-of-the-art detectors on VisDrone2019 under 640×640 resolution (**Table 2**). Our model achieves 50.4% mAP50 and 31.1% mAP50-95 with the second lowest complexity (17.3M parameters, 54.9G FLOPs), outperforming all paradigms. FE-DETR surpasses RT-DETR-R50 by +0.3% mAP50/+0.2% mAP50-95 while using 58.6% fewer parameters, validating its efficiency-accuracy balance. Two-stage methods like EMA [27] incur prohibitive costs, while one-stage MSFE-YOLO-L [12] suffers 3× higher FLOPs from parallel convolutions. End-to-end competitors like UAV-DETR-R18 [9] exhibit spectral redundancy, whereas our frequency-spatial synergy minimizes overhead.

FE-DETR addresses aerial imaging challenges through specialized architectural innovations. For sub-20px vehicle detection, the model leverages phase-preserving fast Fourier transform reconstruction to enhance localization precision compared to conventional approaches like RT-DETR-R50. When handling occluded pedestrians, the WL-GH optimizes feature prioritization across scales, improving recognition of partially visible instances. The framework further demonstrates robustness in motion-blurred scenarios through FFF module, which maintains structural integrity during spectral processing. This capability proves particularly advantageous in dense urban landscapes where overlapping objects and transient visual artifacts complicate detection tasks.

Table 2. Performance evaluation of FE-DETR versus SOTA methods on the VisDrone-2019-DET validation set, with the best and second-best results marked in red bold and blue bold respectively for each metric.

Model	Publication	Imgsize	mAP50 (%)↑	mAP50-95 (%)↑	Params (M)↓	FLOPs (G) ↓
Two-stage methods						
Faster R-CNN[14]	NeurIPS 2015	640×640	31.0	18.2	41.2	127.0
ARFP[10]	Appl Intell	1333×800	33.9	20.4	42.7	193.8
EMA[27]	ICASSP 2023	640×640	49.7	30.4	91.2	-
One-stage methods						
YOLOv12-M[13]	arXiv 2025	640×640	43.1	26.3	20.3	67.5
HIC-YOLOv5[11]	ICRA 2024	640×640	44.3	26.0	**10.5**	**31.2**
MSFE-YOLO-L[12]	GRSL	640×640	46.8	29.0	41.6	160.2
YOLO-DCTI[28]	Remote Sensing	1024×1024	49.8	27.4	37.7	-
End-to-end methods						
DETR[18]	ECCV 2020	1333×750	40.1	24.1	40.0	187.1
Deformable DETR[19]	ICLR 2020	1333×800	43.1	27.1	40.2	172.5
RT-DETR-R18[8]	CVPR 2024	640×640	47.2	28.7	20.2	57.0
UAV-DETR-R18[9]	arXiv 2025	640×640	48.8	30.4	20.5	64.3
RT-DETR-R50[8]	CVPR 2024	640×640	**49.3**	**30.6**	41.8	133.2
Ours						
FE-DETR	-	640×640	**50.4**	**31.1**	**17.3**	**54.9**

4.4 Ablation Experiments

The ablation experiments shown in **Table 3** systematically evaluate the contributions of each proposed module within FE-DETR on the VisDrone2019 validation set. Starting with the baseline RT-DETR-R18, incremental integration of components reveals progressive performance gains. Introducing the FFF module alone elevates mAP50 to 48.4%, demonstrating its efficacy in amplifying faint object signatures through spectral-spatial synergy. The standalone WL-GH improves occlusion reasoning, yielding a 0.7% mAP50 gain by dynamically balancing local-global feature prioritization. The E^2MF independently enhances edge preservation under complex backgrounds, achieving 47.8% mAP50. Combining FFF with WL-GH yields a significant leap to 49.8% mAP50, validating their complementary roles in addressing small targets and dense occlusions. Adding E^2MF to FFF further improves robustness, particularly for motion-blurred instances.

Table 3. Ablation studies using the RT-DETR baseline model were evaluated on the VisDrone2019 validation set, with the best and second-best results marked in red bold and blue bold respectively for each metric.

FFF	WL-GH	E²MF	mAP50 (%)↑	mAP50-95 (%)↑	Params (M)↓	FLOPs (G) ↓
-	-	-	47.2	28.7	20.2	57.0
√	-	-	48.4	29.5	18.7	50.9
-	√	-	47.9	29.2	18.1	56.4
-	-	√	47.8	29.0	20.9	60.8
√	√	-	**49.8**	**30.7**	**16.5**	**50.7**
√	-	√	49.5	30.6	19.2	54.9
-	√	√	49.0	30.3	18.9	60.0
√	√	√	**50.4**	**31.1**	**17.3**	**54.9**

The full FE-DETR configuration demonstrates synergistic integration, where frequency-domain reconstruction, adaptive attention allocation, and edge-aware fusion collectively optimize aerial scene understanding. Notably, parameter count reduces from 20.2M to 17.3M despite performance gains, attributable to FFF's channel splitting and WL-GH's dynamic resource allocation. FLOPs decrease by 3.6%, highlighting computational efficiency from optimized spectral processing and depth-wise convolutions. These results confirm that FE-DETR's architectural innovations harmonize accuracy and efficiency through physics-inspired feature engineering rather than mere capacity expansion.

4.5 Experimental Results

To validate the effectiveness of FE-DETR, we visualized aerial images containing typical challenges in drone-based detection (including undersized targets, occlusions in dense scenarios, and illumination variations between sunlit and shaded areas e.g.). Comparative detection results between RT-DETR and FE-DETR are presented in **Fig. 5**, with purple bounding boxes highlighting regions where FE-DETR demonstrates superior detection performance compared to RT-DETR.

Fig. 5. Comparison of detection effects between RT-DETR and FE-DETR in complex aerial photography scenes.

The detection results reveal significant performance disparities between RT-DETR and FE-DETR in handling complex urban scenarios. As shown in the figure, RT-DETR exhibits multiple labeling errors including fragmented detections, misclassifications, and incomplete object recognition. These errors predominantly occur in areas with small-scale vehicle and overlapping objects. In contrast, FE-DETR demonstrates enhanced robustness through complete word formations and accurate classification of challenging cases, particularly in regions marked by purple boxes. The comparative outcomes substantiate FE-DETR's superior capability in resolving partial occlusions and maintaining detection consistency under scale variations, addressing critical limitations observed in RT-DETR's performance on dense urban imagery.

5 Conclusion

This paper presents FE-DETR, a novel aerial object detection framework that harmonizes spectral-spatial feature integration through Fourier-enhanced processing, adaptive attention allocation, and edge-aware multi-scale fusion. By synergizing global frequency analysis with physics-inspired edge preservation, FE-DETR achieves state-of-the-art performance on VisDrone2019 with 17.3M parameters, demonstrating superior efficiency-accuracy balance for drone-based scenarios. The proposed modules effectively address critical challenges including sub-20px targets, motion blur, and dense occlusions.

Potential extensions include integrating temporal modeling for video-based UAV detection and lightweight deployment via neural architecture search. Exploring cross-modal fusion (e.g., infrared/LiDAR) and self-supervised spectral adaptation could further enhance robustness under extreme conditions. Additionally, dynamic computation strategies tailored to scene complexity may optimize real-time performance for resource-constrained UAV platforms.

6 Acknowledge

This work was supported by the National Key R&D Program of China (Project Number:2023YFC3006700; Topic Five Number: 2023YFC3006705).

References

1. Leng, J., Ye, Y., Mo, M., Gao, C., Gan, J., Xiao, B., Gao, X.: Recent Advances for Aerial Object Detection: A Survey. ACM Computing Surveys 56(12), 1–36 (2024)
2. Liu, X., Ghazali, K.H., Han, F., Mohamed, I.I.: Review of CNN in aerial image processing. The Imaging Science Journal 71(1), 1–13 (2023)
3. Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y., 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, pp. 0–0 (2019)

4. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
5. Liu, G., Han, J., Rong, W.: Feedback-driven loss function for small object detection. Image and Vision Computing 111, 104197 (2021)
6. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
7. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21–37. Springer, Heidelberg (2016)
8. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16965–16974 (2024)
9. Zhang, H., Liu, K., Gan, Z., Zhu, G.-N.: UAV-DETR: Efficient End-to-End Object Detection for Unmanned Aerial Vehicle Imagery. arXiv preprint arXiv:2501.01855 (2025)
10. Wang, J., Yu, J., He, Z.: ARFP: A novel adaptive recursive feature pyramid for object detection in aerial images. Applied Intelligence 52(11), 12844–12859 (2022)
11. Tang, S., Zhang, S., Fang, Y.: HIC-YOLOv5: Improved YOLOv5 for small object detection. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6614–6619. IEEE, Piscataway (2024)
12. Qi, S., Song, X., Shang, T., Hu, X., Han, K.: Msfe-yolo: An improved yolov8 network for object detection on drone view. IEEE Geoscience and Remote Sensing Letters (2024)
13. Tian, Y., Ye, Q., Doermann, D.: Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524 (2025)
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6), 1137–1149 (2016)
15. Cai, Z., Vasconcelos, N.: Cascade R-CNN: High quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(5), 1483–1498 (2019)
16. Wang, C.-Y., Yeh, I.-H., Liao, H.-Y.M.: Yolov9: Learning what you want to learn using programmable gradient information. In: European Conference on Computer Vision, pp. 1–21. Springer, Heidelberg (2024)
17. Yang, Z., Xu, Z., Wang, Y.: Bidirection-fusion-YOLOv3: An improved method for insulator defect detection using UAV image. IEEE Transactions on Instrumentation and Measurement 71, 1–8 (2022)
18. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229. Springer, Heidelberg (2020)
19. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
20. Kong, Y., Shang, X., Jia, S.: Drone-DETR: Efficient small object detection for remote sensing image using enhanced RT-DETR model. Sensors 24(17), 5496 (2024)
21. Zhenyu, H.: Research on Small Target Detection in Optical Remote Sensing Based on YOLOv7. In: 2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE), pp. 804–809. IEEE, Piscataway (2023)

22. Zhang, Y., Ye, M., Zhu, G., Liu, Y., Guo, P., Yan, J.: FFCA-YOLO for small object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 62, 1–15 (2024)
23. Li, R., Shen, Y.: YOLOSR-IST: A deep learning method for small target detection in infrared remote sensing images based on super-resolution and YOLO. Signal Processing 208, 108962 (2023)
24. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
25. Wang, C.-Y., Liao, H.-Y.M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., Yeh, I.-H.: CSPNet: A new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 390–391 (2020)
26. Patro, B.N., Namboodiri, V.P., Agneeswaran, V.S.: Spectformer: Frequency and attention is what you need in a vision transformer. arXiv preprint arXiv:2304.06446 (2023)
27. Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., Huang, Z.: Efficient multi-scale attention module with cross-spatial learning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, Piscataway (2023)
28. Min, L., Fan, Z., Lv, Q., Reda, M., Shen, L., Wang, B.: YOLO-DCTI: small object detection in remote sensing base on contextual transformer enhancement. Remote Sensing 15(16), 3970 (2023)