Integrating Frequency-Spatial Features for Energy-Efficient OPGW Target Recognition in UAV-Assisted Mobile Monitoring

Lin Huang , Xubin Ren , Daiming Qu , Lanhua Li , Jing Xu *

Posted Date: 5 December 2025

doi: 10.20944/preprints202512.0582.v1

Keywords: power transmission line; small object detection; UAV images; frequency-spatial fusion

 doi:10.20944/preprints202512.0582.v1

Article

Integrating Frequency-Spatial Features for Energy-Efficient OPGW Target Recognition in UAV-Assisted Mobile Monitoring

Lin Huang [1,2], Xubin Ren [3], Daiming Qu [1], Lanhua Li [3] and Jing Xu [1,*]

[1] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China
[2] Wuhan Maritime Communication Research Institute, Wuhan 430079, China
[3] School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-Sen University, Shenzhen 518107, China
* Correspondence: xujing@hust.edu.cn

Abstract

Optical Fiber Composite Overhead Ground Wire (OPGW) cables serve dual functions in power systems, lightning protection and critical communication infrastructure for real-time grid monitoring. Accurate OPGW identification during UAV inspections is essential to prevent miscuts and maintain power-communication functionality. However, detecting small, twisted OPGW segments among visually similar ground wires is challenging, particularly given the computational and energy constraints of edge-based UAV platforms. We propose OPGW-DETR, a lightweight detector based on the D-FINE framework, optimized for low-power operation to enable reliable onboard detection. The model incorporates two key innovations: multi-scale convolutional global average pooling (MC-GAP), which fuses spatial features across multiple receptive fields and integrates frequency-domain information for enhanced fine-grained representation, and a hybrid gating mechanism that dynamically balances global and spatial features while preserving original information through residual connections. By enabling real-time onboard inference with minimal energy consumption, OPGW-DETR addresses UAV battery and bandwidth limitations while ensuring continuous detection capability. Evaluated on a custom OPGW dataset, the S-scale model achieves 3.9% improvement in average precision (AP) and 2.5% improvement in AP_{50} over the baseline. These gains enhance communication reliability by reducing misidentification risks, enabling uninterrupted grid monitoring in low-power UAV inspection scenarios where accurate detection is critical for communication integrity and grid safety.

Keywords: power transmission line; small object detection; UAV images; frequency-spatial fusion

1. Introduction

With the modernization of power systems, Optical Fiber Composite Overhead Ground Wire (OPGW) has played a crucial role in enhancing grid communication capabilities and enabling remote monitoring and control [1]. This novel cable design, installed on transmission and distribution lines, protects optical fibers from environmental conditions, such as lightning, short circuits, and load variations, thereby ensuring reliability and service life. Due to its wide applicability across various power lines, ease of installation, strong resistance to environmental interference, low line fault rate, and long lifespan, OPGW has become a preferred communication medium extensively utilized within power systems.

Regular inspection and patrol are key steps for ensuring the long-term operational stability of OPGW [1]. These measures not only facilitate early detection of potential issues but also allow preventive actions before problems escalate, thereby minimizing system failures and maintenance downtime. Among these, Unmanned Aerial Vehicle (UAV)-based automated inspection technology is a critical approach to improving patrol efficiency and accuracy. However, OPGW cables in UAV

 doi:10.20944/preprints202512.0582.v1

imagery appear slender and small-sized, often occluded by complex background elements, such as trees, towers, and conventional ground wires. This makes accurate feature extraction of the target extremely challenging, imposing significant difficulties on detection tasks.

In recent years, the Detection Transformer (DETR) and its variants based on the Transformer architecture have achieved remarkable progress in object detection by eliminating the cumbersome anchor-based design of traditional methods, enabling an end-to-end efficient detection pipeline [2]. Nonetheless, DETR series models generally demand substantial computational resources and demonstrate limited performance when detecting small objects within complex backgrounds, thus struggling to fully meet the stringent requirements of lightweight, high-accuracy, and real-time operation demanded by UAV inspection tasks. Addressing this bottleneck, D-FINE [3], a lightweight DETR variant, achieves faster convergence and reduced computational overhead by redesigning the bounding box regression mechanism and introducing a self-distillation strategy. However, its capabilities in fine-grained feature enhancement and extraction of complex texture information can be further improved.

Therefore, this paper proposes an improved model, termed OPGW-DETR, for detecting OPGW cables and conventional ground wires based on the D-FINE framework. We innovatively introduce a multi-scale convolution global average pooling (MC-GAP) module that integrates spatial features with frequency domain information across multiple receptive fields, significantly enhancing the representation of small targets. Concurrently, a learnable gating mechanism is designed to dynamically balance the fusion of frequency-domain and spatial-domain features, thereby improving the robustness and discriminative power of the model's feature representations. Experimental results demonstrate that OPGW-DETR significantly improves detection accuracy of OPGW cables and ground wires in UAV images while maintaining lightweight computational resource consumption and low power requirements, effectively meeting the real-time inspection demands of practical transmission lines. Our contributions are summarized as follows:

- **MC-GAP for Small-Target Enhancement:** We develop an MC-GAP module that aggregates features from different receptive fields via multi-scale convolutions, followed by concatenation, fusion, and global average pooling augmented with frequency-domain information. By jointly exploiting spatial and frequency cues in a multi-scale manner, MC-GAP strengthens the representation of fine textures and global context for small OPGW targets, leading to notably improved detection accuracy under complex backgrounds.
- **Hybrid Gating for Frequency-Spatial Feature Balancing:** We design a hybrid gating mechanism that combines global learnable scalars (α, β) with spatial-adaptive gate maps to dynamically weight frequency-enhanced and spatial-enhanced features. The global scalars provide coarse-grained balance control, while the gate maps enable pixel-wise adaptivity. Together with residual connections that preserve the original feature information, this scheme enriches feature diversity and robustness, and alleviates the limitations of conventional convolutional layers constrained by single receptive fields and limited feature expressiveness.

2. Related Work

2.1. Transmission Line Inspection Technologies

Traditional inspection methods rely on field personnel carrying auxiliary equipment to conduct close-range patrols, enabling timely detection of component defects and potential hazards [4]. The advantage of manual inspection lies in its detailed and intuitive detection capability, which is especially suitable for complex terrain and remote areas. However, its drawbacks include low inspection efficiency, unstandardized inspection data, and difficulties in achieving real-time supervision [5].

To improve inspection efficiency and safety, helicopter-based inspection technology has been gradually adopted. Equipped with high-resolution visible light and infrared thermal imaging devices, helicopters enable long-distance inspection of conductor connectors, insulators, and fittings [6]. Nevertheless, helicopter inspection faces several limitations: susceptibility to weather and airspace

 doi:10.20944/preprints202512.0582.v1

restrictions, limited flight duration, difficulty accessing obstructed areas and tower bases, and high operational costs.

Subsequently, owing to their agility and lower costs, UAVs have been applied for intelligent transmission line inspection. UAVs carry high-definition cameras, multispectral sensors, thermal imaging, and LiDAR systems, enabling the acquisition of high-resolution images and 3D point cloud data [6]. However, UAV inspection images often contain complex background information, and transmission line targets, such as optical cables and insulators, are slender structures prone to occlusion. This poses significant challenges to traditional image processing and recognition methods.

2.2. Learning-based Object Detection

Deep Convolutional Neural Networks (CNNs) have laid the foundation for object detection [7]. The main CNN-based detection frameworks can be categorized into two-stage and single-stage approaches. Two-stage detectors, exemplified by R-CNN and its variants, use selective search to generate candidate regions, followed by CNN-based feature extraction, classification, and bounding box regression for each region [8]. Fast R-CNN improved speed by sharing convolutional computations, and Faster R-CNN [9] introduced a Region Proposal Network (RPN) to enable end-to-end training, significantly boosting detection efficiency and accuracy. Single-stage detectors like the YOLO series [10]and SSD [11] employ a single neural network to directly predict class probabilities and bounding boxes on the image, achieving end-to-end and efficient inference. YOLOv3 [12] and later versions enhance small object recognition through multi-scale prediction, while RetinaNet [13] introduces focal loss to effectively address class imbalance.

Nevertheless, small and densely packed targets, large scale variations, and complex backgrounds in UAV inspection images present unique challenges to CNNs. To mitigate these issues, Feature Pyramid Networks (FPN) [14] were proposed to enhance multi-scale feature representation. Further improvements, such as Path Aggregation Networks (PANet) [15] and Bidirectional Feature Pyramid Networks (BiFPN) [16], strengthen cross-level feature fusion, enhancing the model's adaptability to complex scenes. Despite these advances, the intrinsic limited receptive field of CNNs hinders their ability to capture global contextual information adequately.

Transformers, with their powerful self-attention mechanisms and global information modeling capabilities, overcome the limited receptive field issue inherent in CNNs. Following this, Carion et al. [2] proposed the DETR, which applied a standard Transformer for object detection, realizing truly end-to-end detection by discarding traditional anchor designs and complex post-processing, such as non-maximum suppression. However, the original DETR suffered from slow convergence, requiring approximately 500 training epochs to achieve satisfactory accuracy, and limited context acquisition, which hindered performance on small objects.

To address these drawbacks, Deformable DETR [17] introduced deformable attention to perform sparse spatial sampling, reducing model complexity. Conditional DETR [18] learned conditional spatial queries from decoder embeddings, enabling multi-head cross-attention in the decoder. DAB-DETR [19] optimized query design with dynamic anchor boxes to accelerate convergence. Anchor DETR [20] encoded anchor points as object queries and designed attention variants, improving prediction accuracy and memory efficiency. Efficient DETR [21] proposed a simplified and efficient end-to-end detection pipeline further optimizing convergence speed. Additionally, slow convergence was linked to the discrete bipartite matching component, which is unstable in early training due to randomized optimization. Thus, denoising training strategies and deformable attention in decoder layers were introduced in DN-DETR [22] and DINO [23] to achieve faster convergence.

For real-time detection, RT-DETR [24] designed an efficient hybrid encoder and a minimum uncertainty query selection strategy to fuse multi-scale features and enhance initial query quality, becoming the first real-time end-to-end DETR detector. However, RT-DETR's limited dynamic perception hindered its ability to detect complex inputs effectively. In response, Dynamic DETR [25] incorporated adaptive dilation and deformable modules in the backbone stage along with dynamic upsampling operators in the neck stage to optimize perceptual capabilities. Moreover, to solve degra-

 doi:10.20944/preprints202512.0582.v1

dation caused by repeated downsampling and poor small object detection, UAV-DETR [26] designed modules to enhance feature perception, semantic representation, and spatial alignment, improving feature expressiveness. D-FINE [3] redefined bounding box regression and introduced an effective self-distillation strategy to optimize prediction methods and enable early model refinement, boosting overall performance. Most DETR models are limited during training by the single-object to single-positive-sample matching scheme, leading to sparse supervision signals that impact training efficiency and detection accuracy. To this end, DEIM-DETR [27] proposed a dense positive sample matching strategy and a novel matching-aware loss function, increasing sample diversity and encouraging the model to focus on high-quality matches.

D-FINE represents an example of a lightweight real-time DETR design. To improve detection accuracy of RT-DETR, D-FINE introduces fine-grained distribution refinement and globally optimal localization self-distillation components. This design redefines the bounding box regression task, enhances fine-grained intermediate representations, and encourages the model to learn better early adjustments, achieving a balance between speed and high accuracy. These innovations allow RT-DETR to leverage its non-post-processing operations better, adapting more effectively to real-time detection scenarios. The D-FINE model architecture is illustrated in Figure 1.

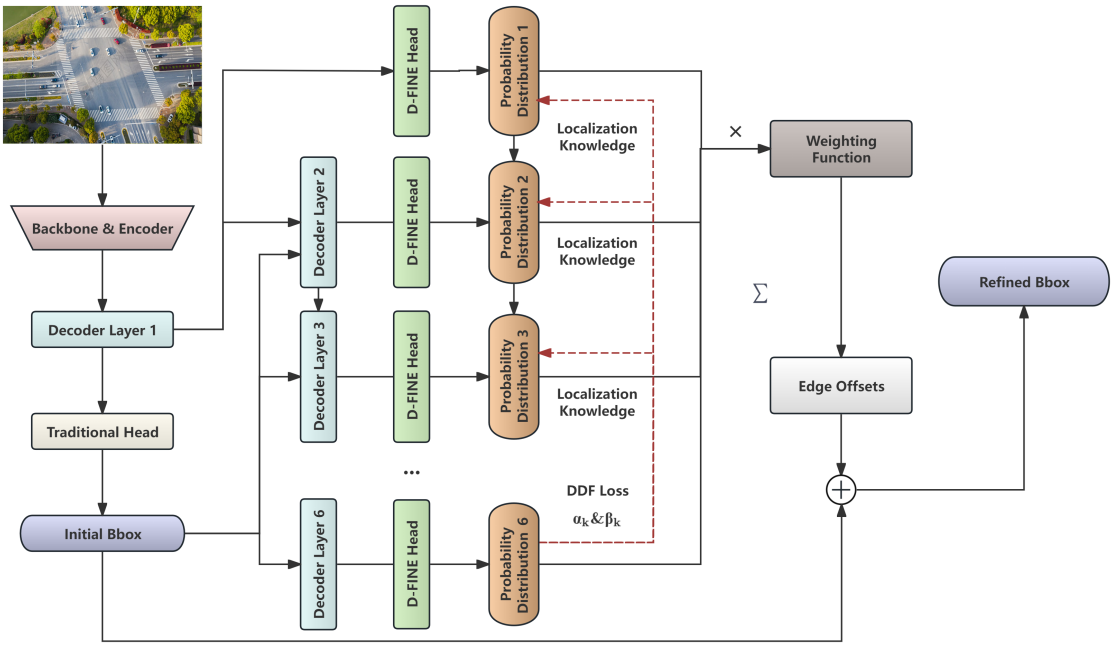


Figure 1. Architecture of D-FINE Model.

Compared with the above DETR variants, our work focuses on the unique challenges of small and thin OPGW cables in UAV transmission-line inspection images, where targets exhibit weak textures, elongated shapes, and strong background interference. Building on the lightweight and real-time D-FINE framework, we introduce a frequency-spatial feature enhancement paradigm rather than redesigning the detection head or matching strategy. Specifically, we develop an MC-GAP module that jointly aggregates multi-scale spatial features and frequency-domain information to strengthen fine-grained texture and global context representation for small targets. In addition, we design a hybrid gating mechanism that dynamically balances frequency-enhanced and spatial-enhanced features in a globally guided yet spatially adaptive manner, effectively mitigating the limited dynamic perception of existing real-time DETR models. These designs are tailored to the characteristics of OPGW targets and UAV inspection scenarios, and thus complement prior DETR improvements that mainly emphasize convergence speed, query design, or generic small-object detection.

 doi:10.20944/preprints202512.0582.v1

3. OPGW-DETR with Frequency-Spatial Feature Fusion for UAV Inspection Images

OPGW optical cables in power transmission lines captured by UAVs present practical challenges characterized by slender, twisted shapes and frequent interference from complex backgrounds and occlusions. In response to these challenges, this paper proposes an improved detection model, OPGW-DETR, based on the D-FINE framework. The architecture of the model is illustrated in Figure 2. The model emphasizes enhanced multi-scale perception of small-object features and the auxiliary utilization of frequency-domain information, thereby enriching feature representation and discriminative capability. These improvements aim to meet the real-time and high-precision detection requirements inherent to UAV inspections.

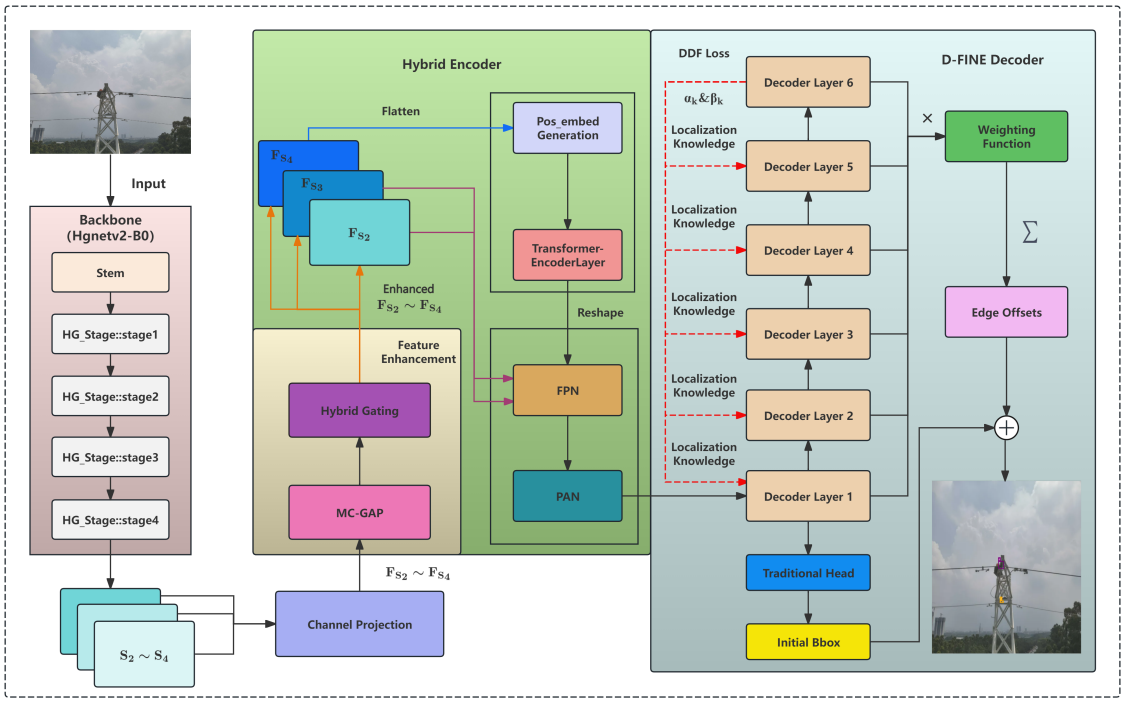


Figure 2. Architecture of OPGW-DETR Model.

The overall framework inherits the lightweight and efficient Transformer-based detection architecture of D-FINE and introduces a novel MC-GAP feature enhancement module, as depicted in Figure 3. This module captures rich spatial receptive field features through parallel convolutions with multiple kernel sizes and employs global average pooling as an effective low-frequency information aggregation method to supplement the global semantic context of spatial features. Subsequently, a hybrid gating mechanism is designed to dynamically fuse spatial and low-frequency domain information, improving robustness in recognizing targets with intricate details. Meanwhile, residual connections preserve the original feature information to prevent degradation.

 doi:10.20944/preprints202512.0582.v1

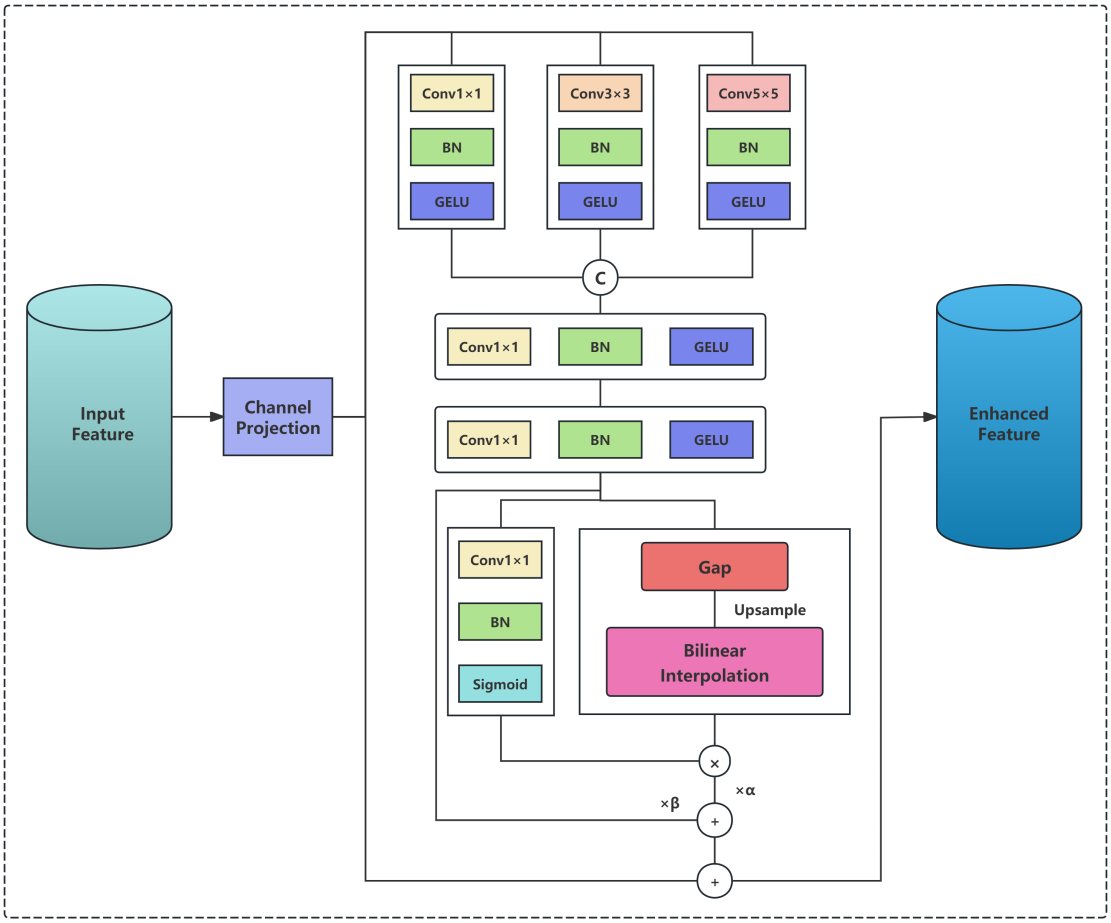


Figure 3. MC-GAP Architecture with Integrated Gating Mechanism.

Targeting the challenges posed by the small size, diverse morphology, and complex backgrounds of cable targets seen from UAV perspectives, this architecture effectively integrates the complementary advantages of spatial and frequency features. It enhances the model's fine-grained representation of curved, slender shapes while balancing real-time inference efficiency with detection accuracy.

3.1. Multi-scale Convolution-based Global Average Pooling (MC-GAP)

The model takes an RGB input image of size $3 \times H \times W$. The backbone network employs the pre-trained HGNetv2 to extract feature maps from N stages, where N varies with model scale. In this paper, we focus on stages $i \in \{2,3,4\}$, which output multi-scale feature maps $X_{S_i} \in \mathbb{R}^{B \times C_i \times \frac{H}{s_i} \times \frac{W}{s_i}}$, $i = 2,3,4$, where B denotes the batch size, C_i represents the number of channels at stage i, and s_i is the downsampling factor. These multi-stage feature maps capture hierarchical information from fine details to high-level semantics.

To unify the feature dimensions across different stages and facilitate subsequent multi-scale fusion and encoder processing, three independent 1×1 convolution layers with BatchNorm modules are employed to project all feature maps to a unified hidden dimension D:

$$F_{S_i} = \mathrm{BN}(\mathrm{Conv}_{1\times1}(X_{S_i})), \quad i = 2,3,4, \tag{1}$$

yielding projected features $F_{S_i} \in \mathbb{R}^{B \times D \times \frac{H}{s_i} \times \frac{W}{s_i}}$, where D is the unified channel dimension, which varies with the model scales.

 doi:10.20944/preprints202512.0582.v1

3.1.1. Multi-scale Parallel Convolution Processing

OPGW cables and ordinary ground wires (denoted as optical cables and ground wires respectively) appear as slender objects with varying sizes and rich texture details in UAV images. To effectively capture these characteristics, the MC-GAP module performs uniform enhancement across stages S_2, S_3, and S_4 after feature projection on the N extracted feature maps F_{S_i} ($i = 2, 3, 4$).

- **Step 1 - Parallel Multi-scale Convolution**: Three convolutional layers with kernel sizes 1×1, 3×3, and 5×5 are applied in parallel to extract multi-scale spatial features. All convolutions include Batch Normalization and GELU activation:

$$F_{S_i}^{(k)} = \text{GELU}(\text{BN}(\text{Conv}_{k\times k}(F_{S_i}))), \quad k \in \{1, 3, 5\}. \tag{2}$$

  The 1×1 convolution captures point-wise information, the 3×3 convolution aggregates local fine textures enhancing the cable's fibrous details, and the 5×5 convolution integrates semantic and textural continuity over a larger spatial range.
- **Step 2 - Channel Concatenation**: The three branches are concatenated along the channel dimension:

$$F_{\text{concat}}^{(i)} = \text{Concat}(F_{S_i}^{(1)}, F_{S_i}^{(3)}, F_{S_i}^{(5)}) \in \mathbb{R}^{B\times 3D\times \frac{H}{s_i}\times \frac{W}{s_i}}. \tag{3}$$

- **Step 3 - Feature Fusion**: A 1×1 convolution is used to fuse and reduce the dimensionality back to D channels:

$$F_{\text{fused}}^{(i)} = \text{GELU}(\text{BN}(\text{Conv}_{1\times 1}(F_{\text{concat}}^{(i)}))) \in \mathbb{R}^{B\times D\times \frac{H}{s_i}\times \frac{W}{s_i}}. \tag{4}$$

This parallel multi-scale convolution design enables the model to simultaneously capture spatial features at different receptive field scales, which is particularly important for detecting cable targets with varying appearances and sizes.

3.1.2. Approximate Extraction of Low-Frequency Information

Traditional frequency domain processing methods (e.g., FFT [28]) are computationally expensive and significantly impact real-time performance. This paper employs global average pooling (GAP) [29] as an efficient alternative for extracting approximate low-frequency information.

For the fused feature $F_{\text{fused}}^{(i)}$, we first apply a 1×1 convolution for feature adjustment:

$$F_{\text{adj}}^{(i)} = \text{GELU}(\text{BN}(\text{Conv}_{1\times 1}(F_{\text{fused}}^{(i)}))) \in \mathbb{R}^{B\times D\times \frac{H}{s_i}\times \frac{W}{s_i}}. \tag{5}$$

Subsequently, global average pooling is performed to extract approximate low-frequency information:

$$F_{\text{gap}}^{(i)} = \text{GAP}(F_{\text{adj}}^{(i)}) = \frac{1}{HW}\sum_{h=1}^{H/s_i}\sum_{w=1}^{W/s_i} F_{\text{adj}}^{(i)}[:, :, h, w] \in \mathbb{R}^{B\times D\times 1\times 1}, \tag{6}$$

where GAP denotes the global average pooling operation. From a frequency domain perspective, this spatial smoothing operation acts as a low-pass filter that effectively suppresses high-frequency components (detail noise and abrupt changes) while preserving low-frequency components (overall structure and global texture information) of the features.

Finally, bilinear interpolation upsampling is used to restore the pooled result to the original spatial size:

$$F_{\text{freq}}^{(i)} = \text{Upsample}_{\text{bilinear}}(F_{\text{gap}}^{(i)}, \text{size} = (\frac{H}{s_i}, \frac{W}{s_i})) \in \mathbb{R}^{B\times D\times \frac{H}{s_i}\times \frac{W}{s_i}}. \tag{7}$$

This achieves spatial alignment between the low-frequency enhanced features and spatial features for subsequent fusion.

 doi:10.20944/preprints202512.0582.v1

3.2. Hybrid Gating Mechanism and Residual Connection

To adaptively balance the contributions of spatially fused features and frequency domain low-frequency information, this paper introduces a hybrid learnable gating mechanism that combines global balance control with spatial adaptivity. The structure is illustrated in Figure 4.

First, a 1×1 convolution followed by batch normalization and Sigmoid activation generates a spatial-adaptive gate map:

$$\text{Gate}^{(i)} = \sigma(\text{BN}(\text{Conv}_{1\times1}(F^{(i)}_{\text{fused}}))) \in \mathbb{R}^{B\times D\times \frac{H}{s_i}\times \frac{W}{s_i}}, \tag{8}$$

where $\sigma(\cdot)$ denotes the Sigmoid function that constrains values to the range $(0, 1)$, and $\text{BN}(\cdot)$ represents batch normalization. Meanwhile, two global learnable scalar parameters $\alpha, \beta \in \mathbb{R}$ are introduced to control the overall contribution balance, initialized as $\alpha_0 = 0.1, \beta_0 = 0.9$. Unlike traditional complementary gating mechanisms where $\beta = 1 - \alpha$, our design allows α and β to be learned independently, providing greater flexibility in balancing frequency and spatial information contributions.

This hybrid mechanism enables adaptive fusion strategy adjustment according to both global statistics and local spatial content:

- **Spatial Adaptivity**: The $\text{Gate}^{(i)}$ map varies across spatial locations, enabling the model to emphasize frequency features in regions where global structural information is crucial while preserving spatial details where fine textures dominate;
- **Global Balance**: The scalar parameters α and β provide coarse-grained control over the relative importance of frequency versus spatial pathways across the entire feature map, allowing the model to learn task-specific optimal weighting strategies.

Finally, the enhanced output feature is obtained through hybrid weighted fusion and residual connection:

$$F_{\text{out}} = F_{S_i} + \alpha \cdot (\text{Gate}^{(i)} \odot F^{(i)}_{\text{freq}}) + \beta \cdot F^{(i)}_{\text{fused}}, \tag{9}$$

where $\odot$ denotes element-wise (Hadamard) product and $\cdot$ represents scalar multiplication with broadcasting. Due to the associativity of multiplication, $\alpha \cdot (\text{Gate}^{(i)} \odot F^{(i)}_{\text{freq}})$ is equivalent to $(\alpha \cdot \text{Gate}^{(i)}) \odot F^{(i)}_{\text{freq}}$. The residual connection (adding back F_{S_i}) preserves the complete information of the original projected features, preventing network degradation, gradient vanishing, and training instability issues.

After processing through the MC-GAP module, spatial domain features and low-frequency enhanced features (obtained via the low-pass filtering effect of GAP) achieve effective cooperative enhancement through the hybrid gating mechanism, significantly improving the model's representation capability for small twisted cable targets and their contextual environment. This frequency-spatial collaborative enhancement mechanism improves the hierarchical representation, discriminability, and robustness against high-frequency noise interference.

 doi:10.20944/preprints202512.0582.v1

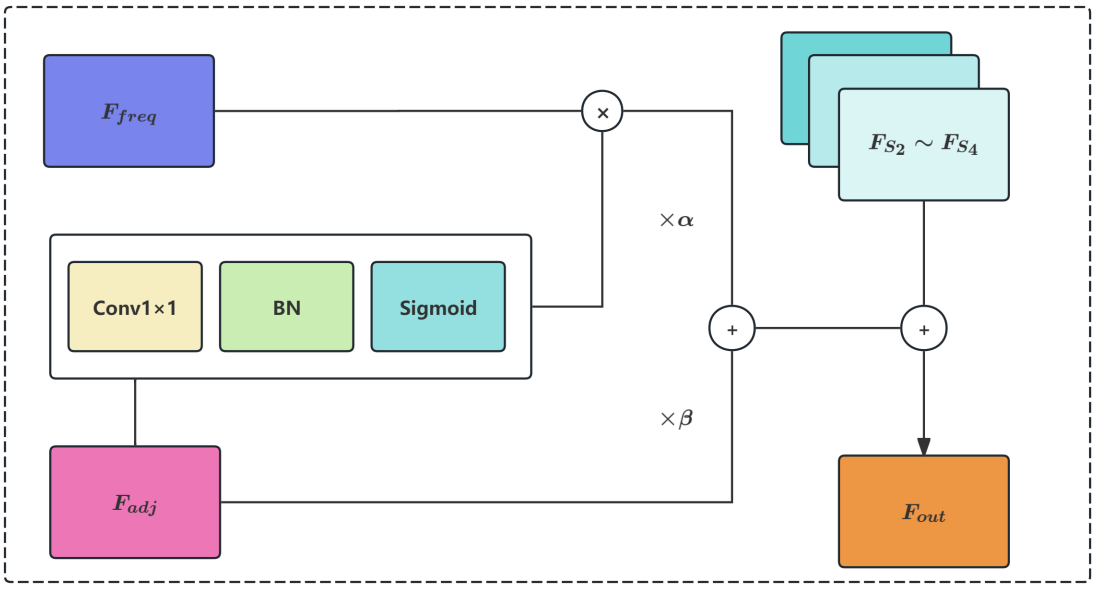


Figure 4. Hybrid Gating Mechanism for Frequency-Spatial Balancing.

4. Numerical Results

4.1. Experimental Setup

We collected a total of 3,137 images of two types of transmission towers captured by UAVs under different environments and angles. Manual annotation was performed using the Labeling image annotation tool. Among these, the OPGW optical cable of the steel towers is labeled as O1, the ordinary ground wire as D1; the OPGW cable of the straight-line towers is labeled as O2, and the ordinary ground wire as D2. The annotated dataset is saved in COCO format and split into training, validation, and test sets with a ratio of 3:1:1.

All models were trained on Linux systems equipped with NVIDIA Quadro RTX 8000 GPUs. The Python version used is 3.11.9, and PyTorch version is 2.8.0+cu126. Our model is based on D-FINE [3], employing HGNetv2 as the backbone consistent with D-FINE. We implemented our proposed method and trained the model for 280 epochs with a batch size of 16. Early stopping was applied, monitoring the metric Bounding Box Average Precision (AP), calculated as the mean AP over 10 IoU thresholds ranging from 0.5 to 0.95. The patience was set to 35 epochs. The OPGW-DETR network is optimized by AdamW [30]. Learning rates were set to 0.0004, 0.0002, 0.0002 and 0.00025 for backbones of sizes N, S, M and X, respectively. Input images are resized to 640 × 640, and COCO standard metrics are reported, including mAP_{50} and $mAP_{50\sim95}$.

We compares four model variants of different scales: N, S, M, and X, designed for various application scenarios and computational resource constraints. These variants differ in backbone network configuration, feature extraction stages, and channel dimensions to balance detection accuracy with computational efficiency. All variants adopt HGNetv2 as the feature extraction backbone, but utilize pre-trained models of different depths. HGNetv2 comprises 4 stages, with returned stage features controlled by parameter adjustment. The specific configurations are shown in Table 1, where C_i denotes the number of output channels at the i-th returned stage of the backbone network, s_i represents the corresponding downsampling stride, and D is the hidden dimension after channel projection.

 doi:10.20944/preprints202512.0582.v1

Table 1. Detailed Configuration Parameters for Different Model Sizes.

Model Scale	Backbone Size	Returned Stages	(C_1, C_2, C_3)	(s_1, s_2, s_3)	D
N	B0	S_3, S_4	(512, 1024, –)	(16, 32, –)	128
S	B0	S_2, S_3, S_4	(256, 512, 1024)	(8, 16, 32)	256
M	B2	S_2, S_3, S_4	(384, 768, 1536)	(8, 16, 32)	256
X	B5	S_2, S_3, S_4	(512, 1024, 2048)	(8, 16, 32)	384

4.2. Detection Performance and Efficiency Comparison

As shown in Table 2, on the collected OPGW dataset, our OPGW-DETR-S (with an HGNetv2-B0 backbone) improves AP_{50} by 2.5% and AP by 3.9% over the baseline D-FINE-S. Despite achieving these gains, OPGW-DETR-S maintains a computational cost below 100 GFLOPs and uses fewer than 20 million parameters. This demonstrates that the proposed detection framework can significantly enhance performance without sacrificing efficiency.

Under the lightweight configuration, OPGW-DETR-S attains the highest accuracy among all compared methods, while still satisfying the stringent constraints of low power consumption and low computational complexity required by real-time embedded platforms (e.g., edge devices deployed on transmission lines or in substations). This balance between accuracy and efficiency is particularly important for long-term online monitoring scenarios, where hardware resources and energy budgets are limited.

Furthermore, we compare our model against other lightweight detectors with similar computational budgets in Table 3. Even under roughly comparable FLOPs and parameter counts, our method consistently achieves superior precision, indicating that the proposed architectural design and hybrid frequency-spatial enhancement strategy provide more effective feature utilization than conventional lightweight backbones and detection heads.

Table 2. Performance Comparison of OPGW-DETR Variants and D-FINE.

Model	InputSize	mAP_{50}	$mAP_{50:95}$	Param(M)	GFLOPs
D-FINE-N [3]	640 × 640	69.0	41.0	3.6	7.1
D-FINE-S [3]	640 × 640	76.0	46.7	9.7	24.8
D-FINE-M* [3]	640 × 640	76.1	47.7	18.3	56.4
D-FINE-X* [3]	640 × 640	78.5	48.5	58.7	202.2
OPGW-DETR-N(Ours)	640 × 640	71.3	42.2	4.8	9.7
OPGW-DETR-S(Ours)	640 × 640	78.5	50.6	17.2	68.9
OPGW-DETR-M*(Ours)	640 × 640	78.3	49.8	25.8	100.4
OPGW-DETR-X*(Ours)	640 × 640	77.3	48.8	75.6	301.3

* The M and X scale models of D-FINE and OPGW-DETR are trained without using pre-trained weights.

 doi:10.20944/preprints202512.0582.v1

Table 3. Performance Comparison of Various DETR and YOLO Models on OPGW Dataset.

Model	InputSize	mAP_{50}	$mAP_{50\sim95}$	Param(M)	GFLOPs
RT-DETR-R50 [24]	640 × 640	70.9	45.2	40.8	130.5
RT-DETR-R101 [24]	640 × 640	72.8	46.7	58.9	191.4
RT-DETR-X [24]	640 × 640	74.1	46.3	65.5	222.5
UAV-DETR-Ev2 [26]	640 × 640	71.2	43.0	12.6	44.0
UAV-DETR-R18 [26]	640 × 640	72.6	44.7	20.5	73.9
UAV-DETR-R50 [26]	640 × 640	75.4	47.1	43.4	166.4
YOLOv9-T [31]	640 × 640	68.5	41.7	1.9	7.9
YOLOv9-S [31]	640 × 640	72.8	46.8	6.9	27.4
YOLOv10-S [32]	640 × 640	69.6	44.3	7.7	24.8
YOLOv10-N [32]	640 × 640	64.0	39.0	2.6	8.4
OPGW-DETR-N(Ours)	640 × 640	71.3	42.2	4.8	9.7
OPGW-DETR-S(Ours)	640 × 640	78.5	50.6	17.2	68.9
OPGW-DETR-M*(Ours)	640 × 640	78.3	49.8	25.8	100.4
OPGW-DETR-X*(Ours)	640 × 640	77.3	48.8	75.6	301.3

4.3. Performance Gains from MC-GAP and Hybrid Gating

We conduct ablation studies on the collected OPGW dataset using OPGW-DETR-S to quantify the contribution of each architectural component to the overall detection performance. Table 4 summarizes the results for different model configurations. In this analysis, MC-GAP denotes the multi-scale convolution-based global average pooling feature enhancement module, while Gate represents a hybrid gating mechanism that employs sigmoid activation and batch normalization to adaptively balance the stage-wise enhanced spatial features F_{S_2}-F_{S_4} with the corresponding approximate low-frequency features.

In the configuration that uses only the MC-GAP module (i.e., without Gate), the enhanced spatial features and low-frequency features F_{freq} are fused with a fixed equal ratio ($\alpha = 0.5, \beta = 0.5$). This simple, non-learnable fusion ignores image content and scene complexity, and thus cannot fully exploit the complementary nature of frequency-spatial information. In contrast, the hybrid gating mechanism learns content-aware weights, allowing the model to dynamically emphasize spatial details or low-frequency context depending on the local characteristics of twisted cable targets and backgrounds. This is particularly beneficial for complex field environments with varying illumination, background clutter, and cable appearances.

The baseline D-FINE-S achieves 76.0% mAP_{50} and 46.7% $mAP_{50\sim95}$. After introducing the MC-GAP module, the detector gains a clear improvement, with mAP_{50} increasing to 78.4% and $mAP_{50\sim95}$ rising to 49.4%. These gains demonstrate that explicitly enhancing multi-scale features with approximate low-frequency information is effective for small, elongated OPGW targets that are easily affected by high-frequency noise.

When the gating mechanism is further integrated into MC-GAP, OPGW-DETR-S achieves the best overall performance, reaching 78.5% mAP_{50} and 50.6% $mAP_{50\sim95}$. At the same time, the model maintains a computational cost of only 68.9 GFLOPs and 17.2M parameters, indicating that the proposed frequency-spatial hybrid gating introduces negligible overhead relative to the obtained accuracy gains. This balance between performance and efficiency confirms that the OPGW-DETR architecture is well suited for real-time deployment on low-power, resource-constrained embedded platforms, such as edge devices mounted on inspection robots or unmanned aerial vehicles for online transmission line monitoring.

 doi:10.20944/preprints202512.0582.v1

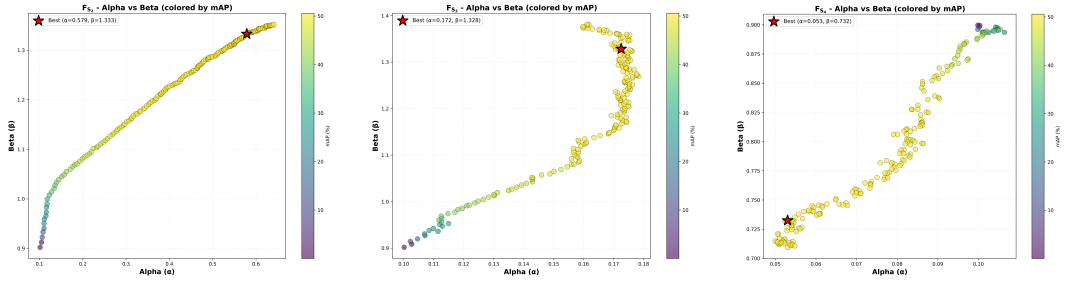
Table 4. Results of The Ablation Study.

Baseline	MC-GAP	Gate	Params(M)	GFLOPs	mAP_{50}	$mAP_{50\sim95}$
✓			9.7	24.8	74.6	46.3
✓	✓		17.1	67.8	78.4	49.4
✓	✓	✓	17.2	68.9	78.5	50.6

We analyzed the optimal hybrid gating parameter configurations (α^*, β^*) that achieve peak detection performance. Figure 5(a–c) visualizes the parameter space, where the color intensity of each point reflects the corresponding mAP, and star markers indicate the optimal configurations. Table 5 summarizes these optimal values.

Table 5. Optimal Hybrid Gating Parameter Configurations.

Features	Level	α^*	β^*	α^*/β^*
F_{S_2}	Low	0.58	1.33	0.43
F_{S_3}	Mid	0.17	1.33	0.13
F_{S_4}	High	0.05	0.73	0.07



(a) **(b)** **(c)**

Figure 5. Hybrid Gating Mechanism Parameter Learning Dynamics. **(a-c)** Exploration of optimal configurations of parameters α and β for different channel-projected features (representing F_{S_2}, F_{S_3} and F_{S_4}, respectively) with respect to mAP.

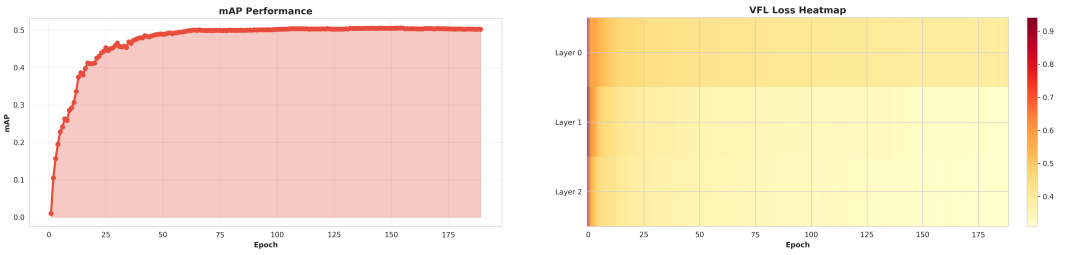
It can be observed that as the semantic hierarchy ascends, the contribution of frequency-domain enhancement must progressively decrease relative to spatial-domain processing. This hierarchical trade-off reflects the fundamental differences among features at different levels:

- **Low-level features (F_{S_2})** require substantial frequency-domain enhancement ($\alpha^* > 0.5$) to amplify fine-grained textures and edge information. The relatively balanced ratio indicates that low-level representations benefit nearly equally from both domains. The high absolute values ($\beta^* > 1.3$) suggest that both frequency and spatial enhancements are necessary to compensate for the inherently limited semantic discriminability of low-level features.
- **Mid-level features (F_{S_3})** exhibit a pronounced shift toward spatial-domain dominance. While maintaining a high β^*, the substantial reduction in α^* indicates that mid-level semantic abstractions are susceptible to frequency-domain perturbations. At this level, features encode partially hierarchical patterns and compositional structures, relying on spatial coherence rather than high-frequency details. Excessive frequency enhancement ($\alpha > 0.2$) may introduce artifacts that compromise semantic information acquisition.
- **High-level features (F_{S_4})** demonstrate an extreme spatial bias, with frequency enhancement nearly eliminated. The minimal α^* value reflects that high-level semantic representations exhibit minimal dependence on frequency-domain information while being particularly sensitive to high-frequency noise therein. Notably, β^* also falls below 1.0, indicating that high-level features obtained from the pretrained backbone already possess sufficient representational capacity, requiring only conservative spatial refinement.

 doi:10.20944/preprints202512.0582.v1

4.4. Analysis of Accuracy Improvement and Loss Stabilization in OPGW-DETR

In Figure 6, the mAP curve exhibits a clear two-stage learning behavior. In the first 50 epochs, the model undergoes a rapid learning phase, with mAP increasing from 0.02 to about 0.48, indicating that the backbone and detection head quickly capture the dominant discriminative features of OPGW targets and background. Afterward, training enters a refinement phase in which mAP gradually converges to 0.49-0.50. This slow saturation suggests that the model is performing fine-grained feature calibration, such as improving localization accuracy for partially occluded cables and reducing confusion with visually similar background structures.



(a) **(b)**

Figure 6. OPGW-DETR-S model training visualization dashboard. **(a)** mAP performance showing model accuracy improvement over training epochs. **(b)** VFL loss heatmap across three decoder layers.

In Figure 7, All primary loss components show stable and consistent convergence, reflecting healthy optimization dynamics. The classification loss (VFL) decreases from 0.80 to 0.40, while the bounding box regression loss drops from 0.22 to 0.03, corresponding to an 86% reduction. Meanwhile, the GIoU loss falls from 1.2 to 0.38. These losses stabilize after approximately epoch 50, indicating that both category discrimination and spatial localization have largely adapted to the data distribution, with no signs of divergence or overfitting. The VFL loss heatmap further reveals relatively uniform learning patterns across the three decoder layers, with all layers exhibiting similar convergence trends. This suggests that the multi-layer decoding structure contributes consistently to classification performance, without any single layer dominating or lagging behind. A comparison between encoder and decoder VFL losses shows similar final values (both around 0.42), but the decoder converges slightly faster in the early epochs. This behavior implies that the decoder benefits more directly from the supervision signal and quickly learns task-specific representations, while the encoder gradually refines the global feature encoding. The total loss decreases from 2.2 to 0.82 (a 63% reduction) and remains stable in the late training stage without noticeable oscillations, indicating that the learning rate schedule and regularization settings are appropriate. The DDF loss converges rapidly from 0.4 to 0.15, demonstrating that the dynamic decomposition features are learned efficiently and do not introduce optimization instability. In contrast, the FGL loss remains stable around 1.15-1.2 throughout training, reflecting its role as a relatively steady regularization or guidance term rather than a sharply decreasing objective. Together, these trends confirm that the proposed OPGW-DETR-S training pipeline is well-conditioned and that both the main detection branch and auxiliary modules contribute effectively to the final performance.

 doi:10.20944/preprints202512.0582.v1

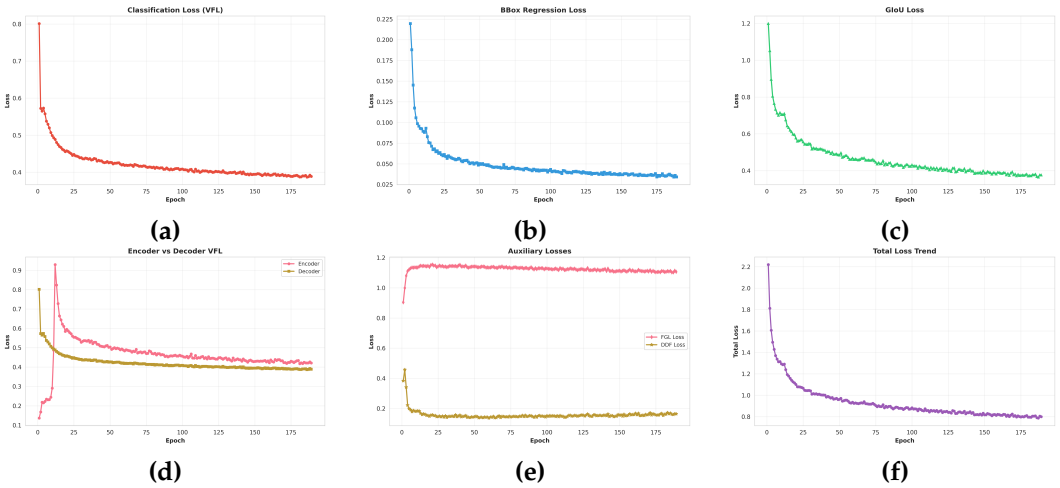


Figure 7. **(a)** Classification loss (VFL) convergence trend. **(b)** BBox regression loss evolution. **(c)** GIoU localization loss trajectory. **(d)** Comparison of Encoder vs Decoder VFL losses. **(e)** Auxiliary losses including FGL and DDF components. **(f)** Total combined loss trend throughout training.

5. Discussion on UAV Images Detection with OPGW-DETR

Compared with existing UAV-based power transmission line object detection models, OPGW-DETR exhibits two notable advantages. First, built upon the improved D-FINE framework, OPGW-DETR successfully eliminates the need for complex anchor box design and post-processing steps, achieving an end-to-end efficient detection pipeline. This greatly simplifies model deployment and maintenance complexity. Second, OPGW-DETR introduces the MC-GAP feature enhancement module and designs a learnable gating mechanism to dynamically balance and fuse spatial domain and low-frequency domain features. As demonstrated in Tables 2 and 3, our model attains favorable detection accuracy while maintaining low computational cost and model parameter count.

This performance improvement primarily stems from several factors. First, the multi-scale convolutions capture rich spatial receptive field features at different scales, which, combined with low-frequency components extracted through the low-pass filtering effect of global average pooling, effectively reinforce the model's capability to represent slender, small-sized, and morphologically diverse cable targets. In traditional convolutional fusion, global structural information tends to be progressively weakened due to the accumulation of high-frequency noise; however, the MC-GAP module leverages GAP's spatial smoothing property to suppress high-frequency noise components while preserving low-frequency structural features, thereby aiding preservation of the global structure and crucial textures of targets, and improving robustness for recognizing subtle objects amid complex backgrounds.

Second, the introduction of the hybrid gating mechanism addresses the dynamic weighting balance between spatial domain features and low-frequency enhanced features (obtained through low-pass filtering). This mechanism combines global balance control (via learnable scalars α and β) with spatial adaptivity (via the Gate map), allowing the model to adjust contribution ratios at both coarse-grained and fine-grained levels according to scene and target characteristics. Specifically, the global parameters provide overall pathway balance, while the spatial-adaptive gate enables pixel-wise emphasis of frequency features in structurally important regions and spatial features in texture-rich areas. This dual-level adaptive mechanism not only reduces semantic conflicts and feature misalignment during feature fusion but also enhances the model's ability to discriminate between similar slender-line targets. The residual connection further prevents feature degradation and preserves the integrity of the original image information.

We also observed that OPGW-DETR achieves high detection accuracy while ensuring low computational resource consumption and real-time inference, meeting the stringent requirements of low power consumption, efficiency, and precision in power line inspection systems.

 doi:10.20944/preprints202512.0582.v1

Nevertheless, despite its overall excellent performance, OPGW-DETR may occasionally be misled by irrelevant textures in complex backgrounds, resulting in attention to non-target regions. Addressing this limitation will be an important direction of our future work.

6. Conclusions

This paper presented OPGW-DETR, a real-time, high-precision detector for OPGW optical cables and conventional ground wires in UAV-based power transmission line inspections, built on an improved D-FINE framework. By integrating a MC-GAP module and a hybrid gating mechanism, the model dynamically fuses spatial and low-frequency features, substantially improving the representation of slender, twisted, and small cable targets. Experimental results show that our method surpasses existing approaches with comparable cost under low power and computational budgets, meeting real-time deployment requirements for power line inspection. Future work will focus on enhancing robustness to noise and occlusion in complex environments and exploring the integration of multimodal sensor data and more advanced contextual modeling techniques.

References

1. Song, J.; Sun, K.; Wu, D.; Cheng, F.; Xu, Z. Fault Analysis and Diagnosis Method for Intelligent Maintenance of OPGW Optical Cable in Power Systems. *Proc. Int. Conf. Appl. Tech. Cyber Intell. (ATCI)* **2021**, pp. 186–191.
2. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. *Proc. Eur. Conf. Comput. Vis. (ECCV)* **2020**, pp. 213–229.
3. Peng, Y.; Li, H.; Wu, P.; Zhang, Y.; Sun, X.; Wu, F. D-FINE: Redefine Regression Task in DETRs as Fine-grained Distribution Refinement, 2024, [arXiv:cs.CV/2410.13842].
4. Alhassan, A.B.; Zhang, X.; Shen, H.; Xu, H. Power transmission line inspection robots: A review, trends and challenges for future research. *Int. J. Electr. Power Energy Syst.* **2020**, *118*, 105862.
5. Ahmed, F.; Mohanta, J.C.; Keshari, A. Power Transmission Line Inspections: Methods, Challenges, Current Status and Usage of Unmanned Aerial Systems. *J. Intell. Robot. Syst.* **2024**, *110*, 1–25.
6. Ferbin, F.J.; Meganathan, L.; Esha, M.; Thao, N.G.M.; Doss, A.S.A. Power Transmission Line Inspection Using Unmanned Aerial Vehicle – A Review. *Proc. Innov. Power Adv. Comput. Technol. (i-PACT)* **2023**, pp. 1–5.
7. Zhou, Y.; Wei, Y. UAV-DETR: An Enhanced RT-DETR Architecture for Efficient Small Object Detection in UAV Imagery. *Sensors* **2025**, *25*, 4582.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149.
9. Girshick, R. Fast R-CNN. *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)* **2015**, pp. 1440–1448.
10. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2016**, pp. 779–788.
11. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. *Proc. Eur. Conf. Comput. Vis. (ECCV)* **2016**, pp. 21–37.
12. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement, 2018, [arXiv:cs.CV/1804.02767].
13. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)* **2017**, pp. 2980–2988.
14. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2017**, pp. 2117–2125.
15. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2018**, pp. 8759–8768.
16. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2020**, pp. 10778–10787.
17. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection, 2021, [arXiv:cs.CV/2010.04159].
18. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional DETR for Fast Training Convergence. *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)* **2021**, pp. 3631–3640.
19. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR, 2022, [arXiv:cs.CV/2201.12329].
20. Wang, Y.; Zhang, X.; Yang, T.; Sun, J. Anchor DETR: Query Design for Transformer-Based Object Detection, 2022, [arXiv:cs.CV/2109.07107].

 doi:10.20944/preprints202512.0582.v1

21. Yao, Z.; Ai, J.; Li, B.; Zhang, C. Efficient DETR: Improving End-to-End Object Detector with Dense Prior, 2021, [arXiv:cs.CV/2104.01318].
22. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2022**, pp. 13619–13627.
23. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection, 2022, [arXiv:cs.CV/2203.03605].
24. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2024**, pp. 16965–16974.
25. Wei, J.; Shen, W.; Hu, T.; Liu, Q.; Yu, J.; Huang, J. Dynamic-DETR: Dynamic Perception for RT -DETR. *Proc. Int. Conf. Robot. Comput. Vis. (ICRCV)* **2024**, pp. 28–32.
26. Zhang, H.; Liu, K.; Gan, Z.; Zhu, G.N. UAV-DETR: Efficient End-to-End Object Detection for Unmanned Aerial Vehicle Imagery, 2025, [arXiv:cs.CV/2501.01855].
27. Huang, S.; Lu, Z.; Cun, X.; Yu, Y.; Zhou, X.; Shen, X. DEIM: DETR with Improved Matching for Fast Convergence. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2025**, pp. 15162–15171.
28. Brigham, E.O.; Morrow, R.E. The fast Fourier transform. *IEEE Spectr.* **1967**, *4*, 63–70.
29. Lin, M.; Chen, Q.; Yan, S. Network In Network, 2014, [arXiv:cs.NE/1312.4400].
30. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization, 2019, [arXiv:cs.LG/1711.05101].
31. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *Proc. Eur. Conf. Comput. Vis. (ECCV)* **2024**, pp. 1–21.
32. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2024**, *37*, 107984–108011.