A Multi-View Classification Method for Distribution Network Towers Based on Improved EfficientNet

Gao Liu[1], Changyu Li[2], Junsheng Lin[3], Xinzhe Weng[4], Qianming Wang[5]*, Zhenbing Zhao[6]
UAV Inspection Center, Guangdong Power Grid Corporation, Guangzhou, China[1, 2, 3]
Department of Automation, North China Electric Power University, Baoding, China[4, 5]
Department of Computer Science, North China Electric Power University, Baoding, China[6]

***Abstract*—View recognition of distribution network towers is a key technology in UAV intelligent inspection. To address the problem of low accuracy of existing deep learning methods in complex background interference, this paper proposes a tower view classification method based on EfficientNet that integrates foreground perception, multi-scale feature fusion, and dual-dimensional attention. First, a Mask-Guided Fusion Module (MGFM) is designed to extract tower foreground masks using the BiRefNet network, enhancing foreground representation and suppressing background interference through a two-stage fusion strategy. Second, a Multi-Scale Attention Aggregation Module (MSAA) is constructed to achieve efficient cross-layer feature fusion through parallel multi-scale convolution, fully integrating shallow details and deep semantic information. Finally, the Convolutional Block Attention Module (CBAM) is introduced to adaptively strengthen view-discriminative features through channel and spatial dual-attention mechanisms, significantly improving the recognition capability for small-sample categories such as top views. Ablation experiments on a self-built multi-view tower dataset show that the proposed method can effectively distinguish different views such as top view, front view, and side view, with significantly improved accuracy compared to other deep learning models, providing technical support for intelligent inspection of transmission lines.**

***Keywords*—Multi-view classification of power towers; mask-guided feature fusion; BirefNet; multi-scale feature fusion; convolutional block attention**

I. Introduction

Transmission lines constitute a critical component of modern power systems, and their safe and stable operation is essential for national economic development and public quality of life [1, 2]. Owing to the vast geographical coverage of transmission networks—often deployed across complex terrains such as mountainous regions and river crossings—traditional manual inspection methods are increasingly constrained by low efficiency, high operational costs, and potential safety risks. In recent years, the rapid development of unmanned aerial vehicle (UAV) technology has enabled vision-based intelligent inspection to emerge as a promising alternative for power infrastructure monitoring [3].

As the primary load-bearing structures of transmission lines, towers play a decisive role in ensuring system reliability. During UAV-based inspection, accurately recognizing the shooting viewpoint of a tower (e.g., front, side, or top view) provides essential spatial context for subsequent inspection tasks, including defect detection, component localization, and three-dimensional reconstruction [4, 5]. Moreover, reliable viewpoint information can facilitate inspection path planning and improve the overall efficiency and consistency of automated inspection workflows. Consequently, high-accuracy and robust tower viewpoint recognition is of significant practical importance for intelligent power inspection systems.

Existing inspection solutions have achieved notable progress by leveraging advances in computer vision and deep learning. Nevertheless, real-world UAV inspection scenes remain highly challenging. Images are frequently captured under complex natural backgrounds, varying illumination conditions, and occlusions, which can severely degrade recognition performance. In addition, inspection data often exhibit strong viewpoint imbalance, where certain viewpoints—such as top views—are underrepresented, further limiting the robustness of conventional recognition pipelines. These practical constraints highlight the need for viewpoint recognition approaches that are not only accurate but also resilient to background interference and data imbalance.

Recent research trends indicate that enhancing visual recognition performance often relies on improved foreground perception, multi-scale feature representation, and attention-based feature refinement. While these directions have demonstrated effectiveness in general vision tasks, they are typically explored in isolation or adopted as generic components without being explicitly tailored to the characteristics of UAV-based tower inspection. As a result, existing approaches struggle to jointly address background suppression, cross-scale feature interaction, and viewpoint-discriminative enhancement within a unified framework, particularly under complex inspection scenarios.

To bridge this gap, this paper investigates the following research question: how can foreground awareness, multi-scale feature interaction, and attention-driven discrimination be systematically integrated into a unified, task-oriented framework to improve tower viewpoint recognition under real UAV inspection conditions? To address these issues, this paper proposes a task-oriented and stage-wise tower viewpoint recognition framework based on EfficientNet-B4. The proposed approach organizes viewpoint recognition as a progressive enhancement process, explicitly strengthening foreground guidance, adaptive multi-scale interaction, and viewpoint-discriminative attention in a coordinated manner.

The main contributions of this work are summarized as follows: 1) A task-oriented, stage-wise viewpoint recognition

framework is developed to systematically integrate foreground-guided perception, adaptive multi-scale feature fusion, and attention-based discrimination for UAV-based tower inspection. 2) A Mask-Guided Fusion Module (MGFM) is introduced to enhance tower foreground perception and suppress complex background interference through a two-stage fusion strategy. 3) A Multi-scale Attention Aggregation Module (MSAA) and a dual-dimensional attention mechanism are incorporated to strengthen cross-scale feature interaction and improve recognition robustness for minority viewpoints. The remainder of this paper is organized as follows. Section II reviews related work on UAV-based inspection, viewpoint recognition, and feature enhancement strategies. Section III details the proposed framework and its constituent modules. Section IV presents the experimental setup and comprehensive performance evaluation is given in Section V. Finally, Section VI concludes the paper and discusses future research directions.

II. Related Work

A. UAV-Based Intelligent Inspection of Power Infrastructure

Unmanned aerial vehicles (UAVs) have been increasingly adopted in power infrastructure inspection due to their flexibility, low operational cost, and capability to access complex terrains that are difficult for manual inspection. Existing UAV-based inspection systems mainly focus on tasks such as transmission line detection, tower localization, and defect identification. Early studies relied on handcrafted visual features and traditional image processing techniques to extract tower or line structures from aerial images. However, these methods generally exhibit limited robustness under complex backgrounds, illumination variations, and occlusion.

With the advancement of deep learning, convolutional neural networks (CNNs) have become the dominant paradigm in UAV-based power inspection. Shajahan et al. [6] proposed a ROS-based computer vision framework to automatically locate tower regions of interest in UAV images, while Manninen et al. [7] developed a multi-stage deep learning network for automated assessment of transmission infrastructure using fly-by images. These approaches demonstrate the effectiveness of data-driven feature learning for inspection tasks. Nevertheless, in most existing works, viewpoint information is treated as an implicit factor rather than an explicit recognition objective. In practical inspection workflows, viewpoint recognition can provide reliable spatial priors for downstream tasks such as defect detection, component localization, and three-dimensional reconstruction. Despite its practical significance, tower viewpoint recognition has received comparatively limited attention as an independent and systematically studied problem in UAV inspection literature.

B. Viewpoint Recognition and Multi-View Classification

Viewpoint recognition and multi-view classification have been widely studied in computer vision, aiming to learn discriminative representations across different observation angles. Recent studies indicate that exploiting viewpoint diversity plays a crucial role in enhancing feature discrimination. Liu et al. [8] explored trusted multi-view deep learning frameworks and demonstrated that effective aggregation of multi-view information can significantly improve classification robustness and accuracy. Yan et al. [9] provided a comprehensive review of deep multi-view learning methods, highlighting the importance of self-attention and cross-attention mechanisms in aligning cross-view representations and modeling inter-view dependencies.

In the context of power infrastructure inspection, viewpoint information has also been recognized as an important factor for improving downstream perception tasks. Luo et al. [10] conducted a comprehensive survey of UAV-based intelligent transmission line inspection systems and emphasized that viewpoint awareness is critical for subsequent tasks such as defect detection and three-dimensional reconstruction. Furthermore, Faisal et al. [11] pointed out in their recent review that although traditional inspection methods may still be effective under constrained conditions, their generalization capability is significantly inferior to that of end-to-end deep learning models, particularly in complex real-world environments.

Despite these advances, existing multi-view learning and inspection-oriented approaches are mostly designed for generic scenarios or treat viewpoint as auxiliary information rather than an explicit recognition target. When directly applied to UAV-based tower inspection, such methods often struggle with severe background interference, viewpoint ambiguity, and class imbalance, especially for minority viewpoints. These limitations highlight the necessity of developing task-specific viewpoint recognition frameworks that explicitly model viewpoint discrimination under realistic UAV inspection conditions.

Meanwhile, generic image classification architectures have evolved rapidly, from classical convolutional neural networks to modern transformer-based models. Architectures such as Inception, ResNet, and VGG have achieved strong performance on large-scale benchmarks, while Vision Transformer (ViT) and Swin Transformer further improve global modeling capability through self-attention mechanisms. Although these models provide powerful feature extractors, they are primarily designed for generic object recognition tasks. When directly applied to UAV-based tower viewpoint recognition, their performance may be limited by severe background interference, viewpoint ambiguity, and class imbalance commonly encountered in real inspection scenarios. These characteristics distinguish tower viewpoint recognition from conventional multi-view benchmarks and call for task-specific modeling strategies.

C. Foreground Perception, Multi-Scale Feature Fusion, and Attention Mechanisms

Foreground perception and background suppression are essential for visual recognition in complex scenes. Segmentation-based methods have been widely used to obtain object-level priors that help reduce background interference, among which encoder–decoder architectures such as U-Net demonstrate the effectiveness of explicitly modeling foreground regions for complex visual environments [12]. Such priors are particularly valuable in UAV inspection images, where the target tower often occupies only a small portion of the image and is surrounded by cluttered natural backgrounds.

However, in many recognition pipelines, foreground extraction and classification are treated as separate stages, and the segmentation prior is not explicitly integrated into the feature learning hierarchy.

Multi-scale feature fusion is another key strategy for improving recognition robustness, as shallow features capture fine-grained spatial details while deep features encode high-level semantic information. Feature pyramid networks (FPN) explicitly construct hierarchical feature representations across multiple scales, highlighting the importance of structured cross-scale interaction for robust visual recognition [13]. In practice, various fusion strategies, including feature concatenation, weighted summation, and pyramid-based architectures, have been explored to exploit cross-scale complementarity. At the same time, attention mechanisms have been extensively investigated to enhance feature representation by emphasizing informative channels or spatial regions.

Despite these advances, existing fusion and attention mechanisms are typically applied as generic plug-and-play components. They are often optimized independently and lack a unified, task-oriented design that jointly considers foreground perception, cross-scale interaction, and viewpoint-discriminative enhancement. In the context of UAV-based tower inspection, this limitation becomes more pronounced under complex backgrounds and severe viewpoint imbalance. Consequently, there remains a gap in developing a coordinated and hierarchical framework that systematically integrates these techniques for robust tower viewpoint recognition.

III. Improved EfficientNet Algorithm

Unlike existing approaches that treat foreground perception, multi-scale fusion, and attention mechanisms as independent components, the proposed framework organizes these processes into a progressive enhancement pipeline specifically tailored for viewpoint recognition. The EfficientNet model [14] consists of three parts: input, backbone, and classification head. This paper selects EfficientNet-B4 as the baseline model with an input resolution of 380×380. The backbone starts with a Stem module (a stride=2, 3×3 standard convolution) followed by multiple MBConv modules. The MBConv module expands channels via a 1×1 convolution, extracts spatial features through depthwise separable convolution, integrates an SE module for adaptive channel recalibration, and finally restores the dimension via a 1×1 convolution with a residual connection to the input (see Fig. 1). The model employs a compound scaling strategy to jointly adjust network depth, width, and resolution, and applies stochastic depth regularization to enhance generalization, achieving an optimal balance between accuracy and efficiency.

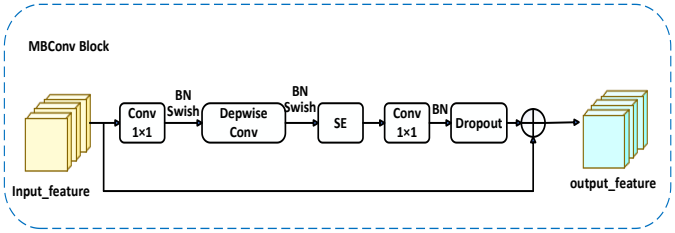

Fig. 1. The structure of MBConv.

Despite EfficientNet's strong performance in various vision tasks, it still faces challenges in the specific context of distribution network tower viewpoint recognition, including severe background interference, insufficient multi-scale feature fusion, and weak discriminative capability for minority viewpoints. To address this, a multi-module collaborative framework is proposed, in which three innovative modules are embedded at different stages of the backbone network to construct a progressive feature learning and enhancement mechanism. The overall framework adopts a "Foreground Guidance — Multi-scale Feature Fusion — Dual-dimensional Attention Modulation" three-stage enhancement strategy: First, the Mask-Guided Fusion Module (MGFM) is introduced at the shallow layer (Layer2). It uses a foreground mask extracted by BiRefNet [15] to spatially guide shallow features, enabling the network to focus on the tower body region early in feature learning and effectively suppress complex background noise. Second, the Multi-scale Attention Aggregation Module (MSAA) is deployed between the middle-deep layers (Layer3 and Layer5). It achieves intra-branch addition fusion and inter-layer feature concatenation through parallel multi-scale convolutions and channel attention weighting, fully integrating shallow details and deep semantic information. Finally, the Convolutional Block Attention Module (CBAM) is applied to the global features. Through dual channel and spatial attention mechanisms, it adaptively strengthens the feature response of viewpoint-discriminative regions, significantly enhancing the recognition robustness for minority classes like the top view. The overall architecture of the proposed model is illustrated in Fig. 2.

The three modules work synergistically to form a complete feature enhancement chain: MGFM solves the background interference problem at the source of feature extraction, providing high-quality foreground-enhanced features for subsequent modules; MSAA achieves efficient and complementary multi-scale feature fusion in the intermediate stage, breaking the limitations of traditional unidirectional propagation or simple concatenation; CBAM performs channel-spatial joint modulation at the decision front-end, ensuring the network can adaptively focus on the most discriminative regions and channels for different viewpoints. This hierarchical, progressive design philosophy closely aligns with the requirements of tower viewpoint recognition, enabling the model to receive targeted optimization at different abstraction levels, thereby achieving high-accuracy and robust multi-view classification performance.

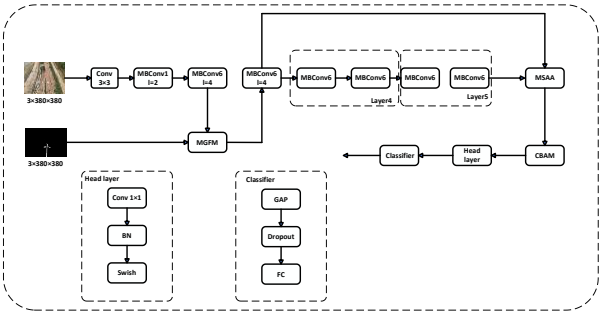

Fig. 2. The overall structure of the model.

A. MGFM Module

The Mask-Guided Fusion Module (MGFM) is designed to address the issue of foreground-background confusion in shallow feature extraction by traditional convolutional neural networks. In viewpoint classification tasks, objects under different viewpoints exhibit significant differences in foreground-background distribution. Conventional methods often treat foreground and background information equally, causing key foreground features to be corrupted by background noise and degrading subsequent viewpoint discrimination performance. Regarding the choice of the foreground segmentation network, this paper systematically evaluated several candidates and ultimately selected BiRefNet as the core component of the MGFM module. Compared to mainstream salient object detection networks such as U²-Net [16] and ISNet [17], BiRefNet demonstrates significant advantages in the power tower scenario. First, BiRefNet employs an innovative bidirectional refinement mechanism that leverages both top-down and bottom-up pathways in a collaborative manner, enabling it to simultaneously capture global semantic context and fine-grained local details. This characteristic is particularly crucial for power towers: as a structured object, a tower exhibits a clear topological hierarchy, which the top-down pathway effectively models through global contextual reasoning. Meanwhile, critical small-scale components (e.g., bolts, grading rings) possess strong discriminative power, and the bottom-up pathway ensures these fine details are preserved during segmentation. Second, BiRefNet achieves high segmentation accuracy while maintaining exceptional computational efficiency, with only 27.8 million parameters—significantly fewer than U²-Net or ISNet.

To leverage this high-quality prior, the MGFM module integrates the foreground mask generated by BiRefNet as spatial guidance. It is inserted after Layer2 of EfficientNet-B4 and employs a two-stage fusion strategy to enhance foreground representation and suppress background interference. The detailed structure of the MGFM is shown in Fig. 3.

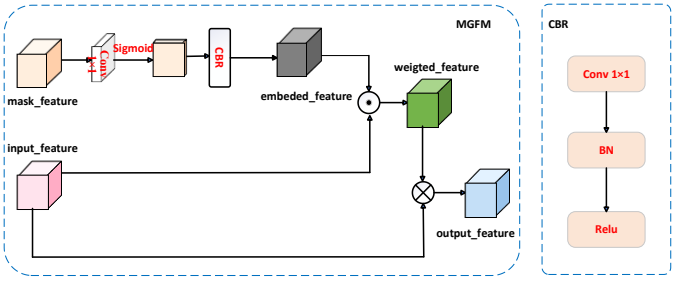

Fig. 3. The structure of MGFM.

The complete MGFM pipeline is as follows: 1) Mask Channel Compression: The three-channel RGB-format mask is first passed through a 1×1 convolutional layer to compress it into a single-channel grayscale map, which is then constrained to the [0,1] range via a Sigmoid activation function to generate a spatial attention prior. 2) Mask Resolution Alignment: The single-channel mask is upsampled/downsampled via bilinear interpolation to match the spatial resolution (H'×W') of the backbone feature map. 3) Mask Channel Embedding: The aligned single-channel mask is passed through another 1×1 convolutional layer, combined with Batch Normalization (BN) and ReLU activation, to map it to the same C-dimensional channel space as the backbone feature, generating the embedded mask Embedded_Mask. 4) Spatial Guidance Fusion: Embedded_Mask is element-wise multiplied (broadcasting) with the original feature Input_Feature to obtain the foreground-enhanced Weighted_Feature. 5) Adaptive Weighted Fusion: A learnable scalar parameter $\alpha \in [0,1]$ (constrained by Sigmoid) is introduced to linearly weight Weighted_Feature and the original Input_Feature.

The core advantage of the MGFM module lies in its efficient foreground perception capability. The mask embedding mechanism accurately identifies foreground regions, avoiding the negative impact of background interference on feature learning. The single learnable fusion parameter α allows the network to adaptively adjust the intensity of foreground enhancement based on task requirements, providing sufficient flexibility while maintaining model lightweightness.

B. MSAA Module

The Multi-scale Attention Aggregation Module (MSAA) is designed to address the limitations of traditional convolutional neural networks in cross-layer feature fusion, specifically the insufficient information interaction and limited multi-scale modeling capability. In deep networks, shallow features contain rich spatial details, while deep features encode stronger semantic information. However, these two types of features are often difficult to effectively collaborate due to their differing receptive fields and levels of abstraction. To this end, the MSAA module employs parallel multi-scale convolutional pathways combined with a channel attention mechanism to achieve efficient and complementary cross-layer feature fusion. Fig. 4 depicts the architecture of the proposed MSAA module. It is deployed after the outputs of Layer3 and Layer5 in EfficientNet-B4 to enhance the model's joint representation capability for both local details and global structures of the power tower.

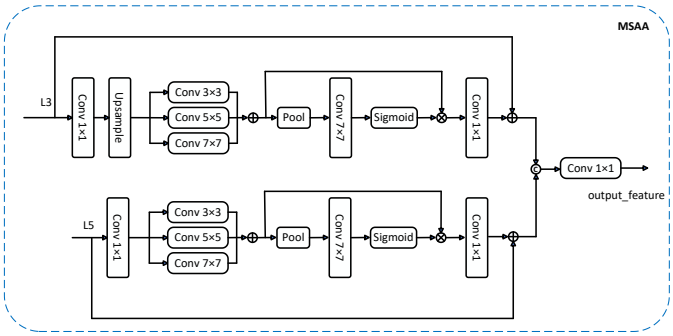

Fig. 4. The structure of MSAA.

The MSAA module's core pipeline consists of two main branches processing features from Layer3 and Layer5, respectively. For each input feature map, it is first mapped to a unified intermediate channel dimension via a Conv1×1, then replicated into three copies and fed into three parallel convolutional branches using Conv3×3, Conv5×5, and Conv7×7 kernels to cover a range of receptive fields from small to large. Within each parallel branch, channel attention

weights are generated through global average pooling, a 1×1 convolution (implemented as 7×7 due to 1×1 input), and Sigmoid activation, which are then used to modulate the original features. The channel-modulated feature maps are element-wise added to the original features to obtain the final output feature map for that branch. Finally, the final output feature maps from the Layer3 and Layer5 branches are concatenated along the channel dimension and compressed to produce the final fused feature map.

C. CBAM Module

To achieve finer feature selection and viewpoint-discriminative enhancement at the decision front-end, this paper introduces the Convolutional Block Attention Module (CBAM) [18] after global feature fusion. CBAM is a lightweight, plug-and-play attention mechanism that sequentially applies Channel Attention Mechanism (CAM) and Spatial Attention Mechanism (SAM) to doubly modulate the input feature map, thereby adaptively enhancing the response intensity of key regions and channels. The serial structure of CBAM is presented in Fig. 5.

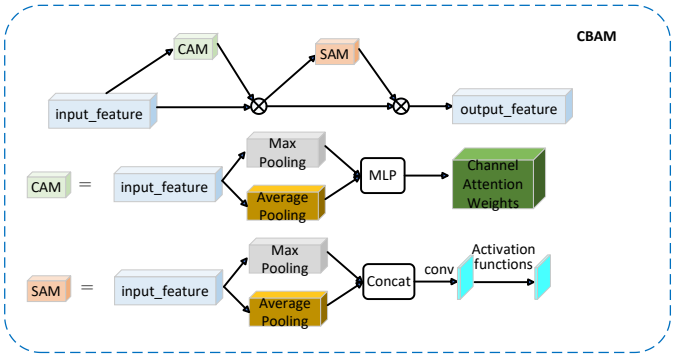

Fig. 5. The structure of CBAM.

The CBAM module adopts a serial structure of CAM and SAM to doubly modulate the input feature map. Its processing flow is as follows: First, the channel attention mechanism compresses the input feature map through Global Average Pooling (GAP) and Global Max Pooling (GMP) to generate two channel descriptors. These descriptors are then processed by a shared Multi-Layer Perceptron (MLP) for feature interaction and nonlinear transformation. Their outputs are summed and passed through a Sigmoid activation function to obtain the final channel attention weight map, which is element-wise multiplied with the original feature map to complete channel-wise adaptive recalibration. Next, the spatial attention mechanism takes the output from the previous step, performs max and average pooling along the channel dimension, concatenates the resulting two spatial maps, fuses them through a 7×7 convolutional layer, and applies a Sigmoid activation function to generate the spatial attention weight map. Finally, this spatial weight map is element-wise multiplied with the current feature map to achieve spatial focusing enhancement, yielding the final feature after dual channel-spatial attention modulation.

The core advantage of the CBAM module lies in its simple structure, high computational efficiency, and no requirement for additional annotations. In this work, the module is applied to the final feature map of the backbone network (before global average pooling). Its functions are: 1) highlighting key channel features relevant to viewpoint discrimination through channel attention; 2) focusing on the most discriminative regions in the image through spatial attention, thereby effectively suppressing background interference; 3) being particularly suitable for solving the recognition problem of minority viewpoints like the top view, as it can adaptively amplify the discriminative feature responses of these rare viewpoints.

IV. Experimental Setup

A. Experimental Environment and Dataset

The experimental hardware environment is shown in Table I. All experiments were conducted on a single server equipped with an NVIDIA GeForce RTX 3090, with CUDA 12.1 driver installed. The software environment uses Ubuntu 20.04 OS, Python 3.10, and deep learning frameworks PyTorch 2.2.1 and Torchvision 0.17.1. The ablation experiments used the AdamW optimizer with momentum parameters beta1 and beta2 set to 0.9 and 0.999, respectively. The model was trained for 30 epochs with a cosine annealing learning rate scheduler (maximum learning rate: 0.001), a batch size of 8, and a weight decay of 0.01.

TABLE I. Experimental Hardware Environment

Hardware	Model	Quantity
CPU	Intel Xeon 6148 Processor	1
Memory	Samsung DDR4 16GB	8
GPU	NVIDIA GeForce RTX 3090	1
Storage	Samsung SSD 980 PRO 2TB	1

The dataset was collected during routine UAV inspections conducted by a power grid company in real operational environments. All images were captured under clear weather conditions with sufficient natural illumination, which is representative of typical inspection scenarios. The UAV flight altitude ranged approximately from 11 m to 50 m, covering different observation distances and view scales. The dataset mainly includes two common structural types of distribution network towers, namely reinforced concrete towers and angle steel towers. In addition, multiple functional tower categories are involved, including straight-line towers, strain towers, and angle towers. Such diversity in acquisition conditions and tower configurations contributes to the robustness and generalization ability of the proposed method. The dataset used in this paper is a distribution network tower dataset captured by a power grid company, containing 5,268 multi-view tower images. There are three viewpoint types: front/back, side, and top. The dataset was randomly split into training and validation sets at a 7:3 ratio, with the validation set also serving as the test set, resulting in 3,686 training images and 1,582 validation/test images.

B. Evaluation Metrics

1) Top-1 accuracy: Accuracy is the most intuitive and commonly used evaluation metric in classification tasks, measuring the proportion of samples correctly predicted by the model. In viewpoint classification, accuracy reflects the

model's overall performance in recognizing different viewpoints (front/back, side, top). The formula is:

$$Accuracy = \frac{\sum_{i=1}^{C} TP_i}{N} \quad (1)$$

where C is the number of classes, N is the total number of samples, and TP_i is the number of true positives for class i. While accuracy provides an intuitive measure of overall performance, it can be misleading under class imbalance.

2) Macro-F1 score: To fairly evaluate the model's performance on each class (especially minority classes), this paper adopts the Macro-F1 Score. The F1 score is the harmonic mean of Precision and Recall:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \quad (2)$$

Where, FP and FN are the false positives and false negatives for class, respectively. The Macro-F1 Score is the arithmetic mean of the F1 scores of all classes, independent of the number of samples per class:

$$F1_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} F1_i \quad (3)$$

V. Experiments and Results Analysis

A. Ablation Study on Individual Modules

To validate the effectiveness of each proposed module, a systematic ablation study is conducted on a self-built multi-view tower dataset. The results are summarized in Table II.

TABLE II. Ablation Study on Individual Modules

Model	Top-1/%	$F1_{\text{macro}}$/%
EfficientNet-B4	88.56	66.40
EfficientNet-B4+MGFM	89.82	69.97
EfficientNet-B4+MGFM+MSAA	90.71	71.75
EfficientNet-B4+MGFM+ MSAA+CBAM	92.38	72.26

The results demonstrate that each module contributes positively to model performance, and their combination yields strong synergistic effects. The MGFM module, serving as the foundation for foreground-aware perception, leverages high-quality foreground masks generated by BiRefNet to effectively suppress complex background interference. It improves Top-1 accuracy by 1.26% and boosts Macro-F1 by 3.57%, confirming the efficacy of foreground guidance in addressing background noise. Building upon this, the MSAA module further refines multi-scale feature fusion. Although it only increases Top-1 accuracy by 0.91%, it significantly raises Macro-F1 by 1.78%, indicating that its parallel multi-scale convolutions and channel-wise attention weighting effectively enhance the complementary representation between shallow details and deep semantics, thereby improving the model's joint modeling of local structures and global layouts of towers. Finally, the CBAM module, acting as a dual-modulation mechanism at the decision front-end, further elevates Top-1 accuracy by 1.65%, achieving a final accuracy of 92.38%. Its channel-spatial dual attention adaptively focuses on viewpoint-discriminative regions, substantially enhancing robustness for minority classes (e.g., top view) while maintaining class balance. Together, these three modules form a complete feature enhancement chain following the hierarchical design principle of "Foreground Guidance — Multi-scale Feature Fusion — Dual-dimensional Attention Modulation", effectively addressing the core challenges in distribution network tower viewpoint classification and validating the superiority of progressive feature enhancement in power vision tasks.

B. Comparison of Different Attention Mechanisms

To investigate the impact of attention mechanisms on classification performance, four representative attention modules—SE [19], ECA [20], EMA [21], and CBAM—are integrated before the classification head of the baseline model. The results are shown in Table III.

TABLE III. Comparison of Different Attention Mechanisms

Attention	Top-1/%	$F1_{\text{macro}}$/%
SE	87.73	66.38
ECA	88.59	64.46
EMA	88.25	65.15
CBAM	89.41	68.19

As shown, CBAM outperforms all other attention mechanisms in both Top-1 accuracy and Macro-F1. While SE, ECA, and EMA primarily operate in the channel dimension, CBAM's incorporation of spatial attention enables more comprehensive feature recalibration. This dual-dimensional modulation not only improves overall performance but also significantly enhances recognition of minority classes. The results confirm that, in tower viewpoint classification, spatial localization is as critical as channel-wise feature selection, and their joint optimization leads to more discriminative feature representations.

C. Comparison of Experiments

The full model is further compared with several mainstream deep learning architectures, including classical CNNs (Inception V3 [22], ResNet50 [23], VGGNet [24]) and modern vision transformers (ViT [25], Swin Transformer [26]). The results are presented in Table IV.

TABLE IV. Comparison of Experiments

Model	Top-1/%	$F1_{\text{macro}}$/%
EfficientNet-B4	88.56	66.40
MobileNetV3	88.75	77.08
ShuffleNetV2	85.30	63.57
Inception V3	83.43	64.69
Resnet 50	85.24	67.20
VggNet	83.18	64.72
Vision Transformer	88.57	67.42
Swin Transformer	90.42	70.35
Ours	92.38	72.26

The proposed method achieves a Top-1 accuracy of 92.38%, significantly outperforming all baseline approaches. Notably, it surpasses the second-best model (Swin Transformer) by 1.96% in accuracy and 1.91% in Macro-F1. This consistent improvement across both metrics—particularly the higher

Macro-F1—demonstrates that the proposed approach not only attains superior overall performance but also exhibits stronger robustness under class imbalance, which is a critical requirement for real-world power inspection scenarios.

To further evaluate the effectiveness of the proposed method under resource-constrained settings, a comparison is conducted with representative lightweight architectures, including MobileNetV3 and ShuffleNetV2. As shown in Table Ⅳ, MobileNetV3 achieves a Top-1 accuracy of 88.75% with a Macro-F1 of 77.08%, while ShuffleNetV2 attains 85.30% Top-1 accuracy and 63.57% Macro-F1. These results indicate that although lightweight models offer advantages in computational efficiency, their discriminative capability for viewpoint recognition—especially under class imbalance—remains limited.

D. Scalability Analysis Across Different EfficientNet Variants

To evaluate the scalability of the proposed modules, experiments are further conducted by integrating them into different EfficientNet variants. The results are shown in Table V.

TABLE V. PERFORMANCE COMPARISON ACROSS DIFFERENT EFFICIENTNET VARIANTS

Backbone	Baseline		Ours	
	Top-1(%)	*$F1_{macro}$(%)*	*Top-1(%)*	*$F1_{macro}$(%)*
Efficient-B0	90.77	70.47	92.58	74.68
Efficient-B1	87.62	66.37	89.35	68.13
Efficient-B2	87.41	66.29	89.93	67.82
Efficient-B3	89.43	68.37	93.16	70.46
Efficient-B4	88.56	66.40	92.38	72.26
Efficient-B5	87.27	67.38	88.15	68.29
Efficient-B6	89.49	69.41	91.24	71.31
Efficient-B7	86.43	66.15	86.19	65.32

As shown in Table V, the proposed method consistently improves performance over the corresponding baselines across most backbone scales. Notable gains in both Top-1 accuracy and Macro-F1 are observed for EfficientNet-B0 to B6, indicating that the proposed foreground-guided and multi-scale attention mechanisms generalize well across different model capacities.

In particular, the improvements in Macro-F1 demonstrate enhanced class-balanced recognition performance, suggesting that the proposed framework effectively alleviates viewpoint imbalance regardless of backbone depth. Although marginal performance saturation or slight fluctuation is observed on the largest backbone (EfficientNet-B7), the overall trend confirms that the proposed method is not limited to a specific EfficientNet configuration. These results validate the scalability and robustness of the proposed framework across a wide range of network capacities.

E. Confusion Matrix

To gain deeper insight into the model's classification performance for each individual category, particularly its error patterns, this section employs confusion matrices for visual analysis. In addition to the visual analysis, quantitative confusion-matrix-derived statistics are reported to further clarify class-wise misclassification behavior. Table VI summarizes the class-wise precision and recall of the baseline model and the proposed method, providing a quantitative comparison of classification performance across different viewpoints.

TABLE VI. CLASS-WISE PRECISION AND RECALL COMPARISON BETWEEN BASELINE AND PROPOSED METHOD

Class	Baseline		Ours	
	Precision(%)	*Recall(%)*	*Precision(%)*	*Recall(%)*
Front/Back	90.44	93.96	93.52	96.23
Side	92.26	89.07	96.76	90.86
Top	17.24	15.63	24.71	32.81

The confusion matrix clearly illustrates the model's classification and misclassification situations for each viewpoint category. The diagonal elements represent the proportion of correctly classified samples, while the off-diagonal elements reveal the category pairs that the model tends to confuse. The confusion matrices of the baseline model (EfficientNet-B4) and the models with the proposed modules incrementally integrated on the test set are compared, as shown in Fig. 6 to Fig. 9.

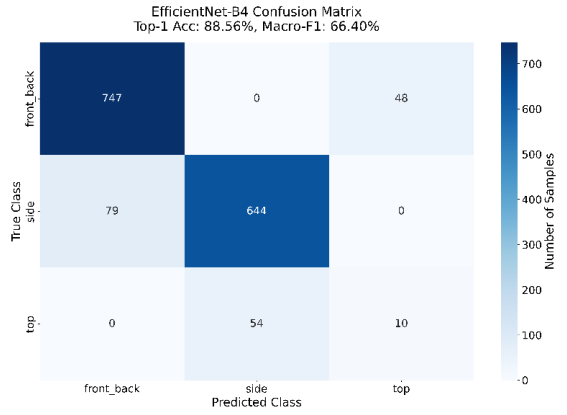

Fig. 6. Confusion matrix of the baseline model (EfficientNet-B4).

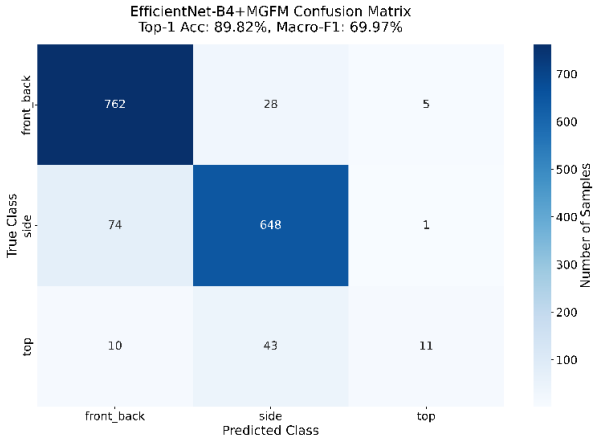

Fig. 7. Confusion matrix of the baseline model with MGFM.

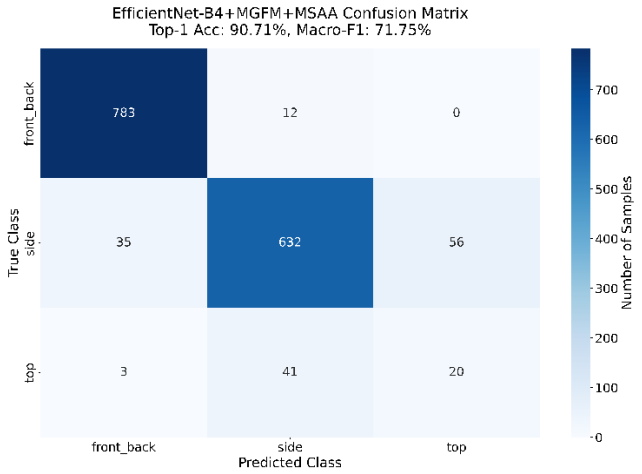

Fig. 8. Confusion matrix of the baseline model with MGFM and MSAA.

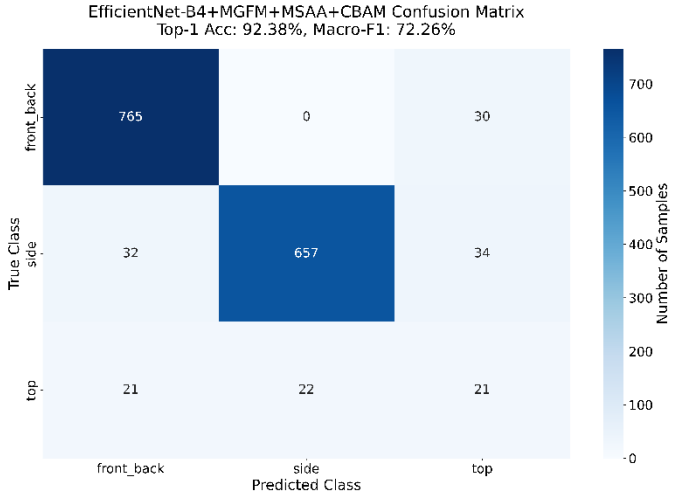

Fig. 9. Confusion matrix of the proposed full model (with MGFM, MSAA, and CBAM).

Fig. 6 illustrates the confusion matrix of the baseline EfficientNet-B4 model. Although satisfactory performance is achieved for the front/back and side viewpoints, severe misclassification is observed for the top view. A large proportion of top-view samples are incorrectly predicted as side view, resulting in extremely low recall for the top category. This indicates that the baseline model struggles to capture discriminative features for minority viewpoints under complex backgrounds, which directly contributes to the low Macro-F1 score. As shown in Fig. 7, after introducing the Mask-Guided Fusion Module (MGFM), the classification performance for front/back and side views is moderately improved, and background-induced misclassification is partially suppressed. However, confusion between the top and side viewpoints remains prominent, suggesting that foreground guidance alone is insufficient to resolve viewpoint ambiguity caused by scale variation and limited semantic discrimination. Fig. 8 presents the confusion matrix of the model with both MGFM and MSAA integrated. Compared with the previous configurations, the number of top-view samples correctly classified increases noticeably, while misclassification into the side category is reduced. This improvement demonstrates that multi-scale feature aggregation effectively enhances the representation of viewpoint-sensitive structural cues, particularly for small-scale and underrepresented viewpoints. Finally, Fig. 9 shows the confusion matrix of the full model incorporating MGFM, MSAA, and CBAM. The results reveal a substantial reduction in cross-view confusion across all categories. In particular, the recall of the top view is significantly improved, indicating that the attention-based modulation further strengthens viewpoint-discriminative regions in both channel and spatial dimensions. The overall confusion pattern becomes more compact along the diagonal, which is consistent with the highest Top-1 accuracy and Macro-F1 score achieved by the proposed method.

Overall, the progressive reduction of misclassification errors from the baseline to the full model validates the complementary roles of foreground guidance, multi-scale feature fusion, and dual-dimensional attention modulation. These results confirm that the proposed framework effectively alleviates viewpoint imbalance and enhances classification robustness in real UAV-based tower inspection scenarios.

F. Overall Experimental Discussion

Overall, the experimental results demonstrate that the proposed method achieves consistent and substantial performance improvements across different evaluation metrics and backbone configurations. Compared with the baseline model, the proposed framework improves Top-1 accuracy by up to 3.82% and Macro-F1 by more than 6%, with particularly notable gains observed on the minority top-view category. These improvements are most evident under challenging conditions, such as complex backgrounds and severe class imbalance, which commonly occur in real UAV-based power inspection scenarios. Consequently, the proposed method provides a more reliable and robust solution for tower viewpoint recognition, facilitating downstream inspection tasks and supporting practical deployment in real-world power grid applications.

VI. Conclusion

To address the key challenges in multi-view classification of distribution network towers—such as severe background interference, difficulty in recognizing minority viewpoints, and insufficient cross-scale feature interaction—this paper presents an improved EfficientNet-B4-based viewpoint recognition framework. A progressive three-stage enhancement mechanism, consisting of foreground-guided perception, adaptive multi-scale feature fusion, and dual-dimensional attention modulation, is constructed to systematically enhance viewpoint-discriminative representations.

Extensive experimental results on a self-built multi-view tower dataset demonstrate that the proposed method achieves consistent and significant performance improvements under challenging UAV inspection conditions, including complex natural backgrounds and pronounced class imbalance. Specifically, foreground-guided feature extraction effectively suppresses background noise and emphasizes tower body regions, while adaptive multi-scale fusion fully exploits the complementary characteristics of shallow structural details and deep semantic information. In addition, the dual attention modulation mechanism further strengthens viewpoint-sensitive

regions, leading to substantial robustness gains for minority viewpoints, particularly the top view.

Beyond the specific application of tower viewpoint recognition, this work provides a generalizable framework for multi-view visual classification under challenging conditions. By organizing foreground guidance, multi-scale feature interaction, and attention-based discrimination into a unified progressive pipeline, the proposed method establishes a reusable design paradigm that can be extended to other UAV inspection tasks and viewpoint-sensitive recognition problems. Owing to its stage-wise and task-oriented design, the framework exhibits strong generality and transferability across different backbone configurations.

From a practical perspective, the consistent improvements in both Top-1 accuracy and Macro-F1—especially under complex backgrounds and severe class imbalance—directly translate into more reliable perception in real inspection scenarios. The robust viewpoint recognition results provide stable spatial priors for downstream tasks such as tower defect detection, component localization, inspection path planning, and three-dimensional reconstruction, thereby supporting real-world deployment beyond performance-driven architectural refinements. Future work will focus on three directions: (1) expanding the dataset scale and viewpoint diversity to further improve model generalization; (2) investigating lightweight deployment strategies to enable real-time inference on edge devices; and (3) exploring end-to-end joint optimization of viewpoint recognition and defect detection to construct a more integrated and efficient inspection perception pipeline.

ACKNOWLEDGMENT

This paper is supported by the Science and Technology Project of Guangdong Power Grid Co., Ltd. (Project Name: Research on Image Defect Detection and Deduplication Technology Integrating 3D Space; Project No.: 030000KC23120086).

REFERENCES

[1] T. Aziz, Z. Lin, M. Waseem, and S. Liu, "Review on optimization methodologies in transmission network reconfiguration of power systems for grid resilience," *Int. Trans. Electr. Energy Syst.*, vol. 30, no. 12, p. e12704, Dec. 2020.

[2] B. Lu, Z. Yang, M. Ding, J. Ma, D. Sun, W. Li, and X. Zhao, "Vision-based transmission tower strain monitoring using industrial camera module and slider strain sensing mechanism," *J. Civ. Struct. Health Monit.*, vol. 15, no. 3, pp. 2419–2432, Mar. 2025.

[3] S. Zhang, Z. Li, L. Huang, K. Jia, and D. Zhang, "UAV transmission line transfer-acceptance technology based on vision," in *Proc. Int. Conf. Algorithm, Imaging Process., Mach. Vis. (AIPMV)*, Qingdao, China, 2023, pp. 129692S.

[4] Y. He, Z. Liu, Y. Guo, Q. Zhu, Y. Fang, and Y. Yin, "UAV based sensing and imaging technologies for power system detection, monitoring and inspection: a review," *Nondestruct. Test. Eval.*, vol. 40, no. 1, pp. 1–25, Nov. 2024.

[5] E. Pastucha, E. Puniach, A. Ścisłowitz, P. Ćwiąkała, and W. Niewiem, "3D Reconstruction of Power Lines Using UAV Images to Monitor Corridor Clearance," *Remote Sens.*, vol. 12, no. 22, p. 3698, Nov. 2020.

[6] N. M. Shajahan, A. Sasikumar, T. Kuruvila, and D. Davis, "Automated inspection of monopole tower using drones and computer vision," in *Proc. 2nd Int. Conf. Intell. Auton. Syst. (ICoIAS)*, Kollam, India, 2019.

[7] H. Manninen, C. J. Ramlal, A. Singh, J. Kilter, and M. Landsberg, "Multi-stage deep learning networks for automated assessment of electricity transmission infrastructure using fly-by images," *Electr. Power Syst. Res.*, vol. 209, p. 107948, Aug. 2022.

[8] W. Liu, X. Yue, Y. Chen, and T. Denoeux, "Trusted multi-view deep learning with opinion aggregation," in *Proc. 36th AAAI Conf. Artif. Intell. (AAAI)*, Vancouver, Canada, 2022, pp. 1–8.

[9] X. Yan, S. Hu, Y. Mao, Y. Ye, and H. Yu, "Deep multi-view learning methods: A review," *Neurocomputing*, vol. 448, pp. 106–129, Aug. 2021.

[10] Y. Luo, X. Yu, D. Yang, and B. Zhou, "A survey of intelligent transmission line inspection based on unmanned aerial vehicle," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 173–201, Apr. 2022.

[11] M. A. A. Faisal, I. Mecheter, Y. Qiblawey, J. Hernandez Fernandez, M. E. H. Chowdhury, and S. Kiranyaz, "Deep learning in automated power line inspection: A review," *Appl. Energy*, vol. 385, p. 125507, May 2025.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 2015, pp. 234–241.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2117–2125.

[14] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, vol. 97, pp. 6105–6114, 2019.

[15] P. Zheng, D. Gao, D.-P. Fan, L. Liu, J. Laaksonen, W. Ouyang, and N. Sebe, "Bilateral reference for high-resolution dichotomous image segmentation," *CAAI Artif. Intell. Res.*, vol. 3, p. 9150038, 2024.

[16] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U²-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, p. 107404, Oct. 2020.

[17] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape matters for infrared small target detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 3–19.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.

[20] Q. Wang, B. Wu, P. Zhu, et al., "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11534–11542, 2020.

[21] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Vancouver, Canada, 2023, pp. 1–5.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2818–2826, 2016.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[26] Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 10012–10022, 2021.