ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

Attention-Enhanced Multi-Scale Feature Modeling for Slender Power Facility Detection in Complex Outdoor Environments

Peiqi He[1,a,*], Jinhao Yuan[1,b]

[1]*School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai, China*
[a]*233421702@st.usst.edu.cn*, [b]*233421700@st.usst.edu.cn*
[*]*Corresponding author*

Abstract: *To address issues in outdoor power facility detection such as strong environmental interference, large variations in target scale, and slender structures being easily obscured by the background, this paper reconstructs a power facility detection dataset and proposes an improved framework for outdoor power facility detection based on YOLOv11n. The proposed method systematically redesigns the network architecture and feature modeling approach in terms of multi-scale feature modeling, orientation-aware modeling for slender targets, and robust feature modeling under complex backgrounds. In particular, the feature modeling strategy for handling complex background interference is further enhanced. Experimental results demonstrate that the proposed method significantly outperforms the baseline model on the self-constructed power facility dataset, with mAP@0.5 and mAP@0.5:0.95 increased by 4.3% and 10.3%, respectively. Meanwhile, it achieves comparable performance to YOLOv11n on several public datasets, demonstrating strong generalization capability and providing solid evidence that the proposed architecture can more effectively accomplish intelligent detection of outdoor power facilities.*

Keywords: *Power facility detection; YOLOv11n; Multi-scale features; Attention mechanism; Extreme outdoor environments*

1. Introduction

With the rapid development of China's economy, the scale of the power system has been continuously expanding, yet some early-stage facilities have entered an "aging" phase, resulting in weakened resilience. The security of power facilities is also an important component of national strategic security[1]. Prolonged exposure to the natural environment accelerates hardware aging and corrosion, making operation and maintenance increasingly difficult. If such hidden hazards trigger accidents, they could lead to significant economic losses and casualties [2], posing a severe threat to national and societal security. In the face of these challenges, it is particularly urgent and important to introduce efficient visual monitoring technologies for real-time inspection and hazard detection of power infrastructure.

Regarding research on power facility detection, numerous studies have been conducted. Traditional algorithms and machine learning methods include using Scale Invariant Feature Transform (SIFT) for feature extraction and support vector machines (SVM) for classification and recognition of power equipment [3], converting sound signals into images for subsequent recognition [4], and employing multi-task sparse representation-based methods for power equipment image recognition [5].

In today's intelligent inspection of outdoor power equipment, methods that integrate multimodal data, such as multimodal approaches, 3D point clouds [6,7], and LiDAR [8], have gained significant attention. Reference [9] proposed an efficient method for extracting 3D targets of transmission facilities from UAV-acquired point cloud data. Reference [10] introduced a segmentation method for substation LiDAR point clouds based on point neighborhood structures, while reference [11] presented a substation equipment recognition approach combining plane detection and point cloud registration. Considering the high cost and deployment difficulties of 3D point clouds and LiDAR, as well as the susceptibility of traditional machine vision algorithms to false positives, deep learning techniques—with their low cost and strong feature extraction capabilities—have gradually emerged as the mainstream choice in current object detection research.

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

Deep learning-based object detection can be mainly divided into two categories: two-stage algorithms represented by R-CNN [12] and Fast R-CNN [13], which offer high accuracy but are computationally intensive and thus challenging for real-time applications on mobile devices; and single-stage algorithms represented by the YOLO series [14,15] and SSD [16], which are simpler and easier to deploy but often suffer from limited native accuracy when dealing with dense occlusions and large scale variations in outdoor power facilities.

Research on power equipment recognition using deep learning has become a hotspot in the field. Zhou et al. [17] proposed a power equipment image recognition method using a dual-channel convolutional neural network (DC-CNN) combined with a random forest (RF) classifier, but the inference speed was limited by the classifier's performance. Xu et al. [18] integrated Long Short-Term Memory (LSTM) networks into the Faster R-CNN framework for power equipment detection, which, however, further increased model parameters and computational burden. Hao et al. [19] compared YOLOv3 and Faster R-CNN for high-voltage transmission tower detection and evaluated their performance under various metrics, revealing the shortcomings of early YOLO versions in localization accuracy. Wu et al. [20] improved YOLOv5 to address the challenge of small-object recognition in complex substation scenes by introducing deformable convolution modules into the backbone network and replacing PANet with a simple and effective BiFPN structure in the neck. Wang et al. [21] enhanced the YOLOv5m model by adding a dynamic head module to unify scale awareness, spatial awareness, and task awareness in the detection head and proposed a novel NWD-CIoU loss function. However, these improvements often trade inference speed for increased accuracy. Yang et al. [22] modified YOLOv7 for insulator defect detection under extreme weather.

Despite these optimizations, existing methods still lack robustness when faced with severe multi-view scale variations, high target-background integration, or extreme weather conditions such as heavy rain or snow. To address these challenges, this study first constructs and augments a power facility dataset with rain and snow-enhanced samples to mitigate the lack of extreme weather data. Based on the lightweight YOLOv11n as a baseline, we propose an improved framework for outdoor power facility detection with the following systematic strategies:

- To suppress strong interference from complex backgrounds such as trees and buildings, multi-scale cross-attention (MCA) [23] is deeply integrated into the feature extraction network, forming the C3k2_MCA structure, whose multi-scale channel interaction effectively reduces background noise.

- To address the challenges of extracting features from transmission towers and other equipment with varying shapes and scales, the C2PSA structure is reconstructed using deformable large kernel attention (D-LKA) [24], enhancing adaptability to geometric deformations under different viewpoints.

- To improve edge information capture for slender targets such as utility poles, we propose a modified scheme that combines standard deviation pooling with a stochastic dual-stream strategy, reconstructing the Monte Carlo attention (MCAttn) [25] mechanism to significantly enhance sensitivity to high-frequency features of elongated structures.

- To improve localization convergence in complex scenes, the WIoU v3 [26] loss function is employed with its dynamic non-monotonic focusing mechanism to optimize bounding box regression, effectively balancing gradient contributions from samples of different quality and thereby improving detection accuracy

2. YOLOv11

YOLOv11 is the latest computer vision model released by Ultralytics in 2024. It represents an architectural upgrade based on YOLOv8 and supports multiple tasks, including object detection and instance segmentation. Its network architecture, shown in Figure 1, continues the end-to-end single-stage detection paradigm, completing localization and classification simultaneously within a unified framework:

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

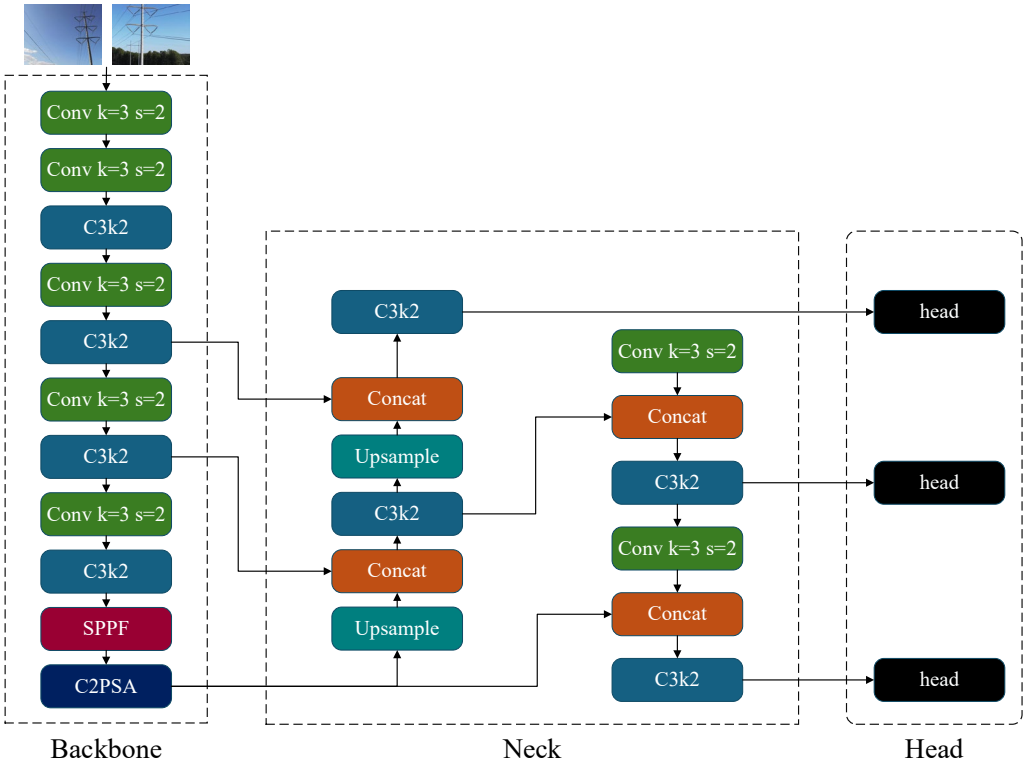

Figure 1: YOLOv11 network model.

Compared with its predecessors, YOLOv11 introduces the following optimizations: (1) replacing the C2f module in the backbone with the C3k2 module to enhance feature extraction while maintaining a lightweight design; (2) retaining the SPPF module for efficient multi-scale feature aggregation; (3) introducing the C2PSA module to improve attention to key regions; and (4) optimizing the Neck connections to reduce memory usage and accelerate inference. Overall, YOLOv11 achieves higher detection accuracy and better generalization performance through architectural reconstruction while reducing the number of parameters.

3. Improved YOLOv11 Algorithm

3.1. MCA Module

For outdoor scenarios, where power facilities such as transmission towers and power lines often exhibit large scale variations and prominent elongated geometries, this study introduces the Multi-Scale Cross-Axis Attention (MCA) mechanism to enhance the networks perception of spatial structures. Unlike traditional axial attention, which is limited to a single dimension, the MCA strategy constructs a dual-parallel architecture to aggregate contextual features along both horizontal and vertical directions, effectively capturing global dependencies. The mechanism employs two sets of directional strip convolutions (1×k convolution for the horizontal path and k×1 convolution for the vertical path) to adapt to the diverse extension directions of power facilities and extract morphological details across different receptive fields. The overall architecture of the module is shown in Figure 2.

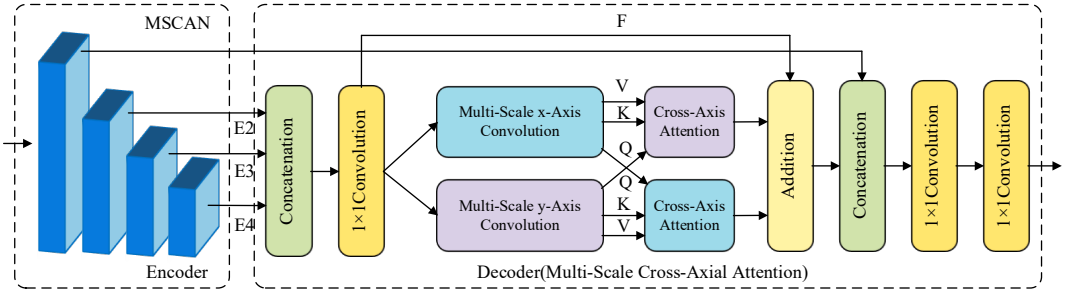

Figure 2: The overall architecture of the MCA module.

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

For a given feature map $X \in R^{H\times W\times C}$, multi-scale strip convolutions are employed for encoding, and the formulation is as follows:

$$F_x = Conv_{1\times1}\left(\sum_{i=0}^{2} Conv1D_i^x(Norm(F))\right) \tag{1}$$

Here, $Conv1D_i^x$ denotes the strip one-dimensional convolution applied along the x-axis, *Norm* represents layer normalization, and F_x is the resulting output. The convolution kernels are set to different widths, such as 7, 11, and 22, to capture multi-scale features. Similarly, F_y can be computed using the same approach. For the upper branch, the cross-attention between F_x and F_y is calculated to obtain F_t, which is formulated as follows:

$$F_t = Attention_y(F_y, F_x, F_x) \tag{2}$$

Here, *Attention* represents the multi-head cross-axis attention computed along the x-axis, while the lower branch applies a similar computation along the y-axis to obtain F_b. The final output feature is given by the sum of the two:

$$F_{out} = Conv_{1\times1}(F_t) + Conv_{1\times1}(F_b) \tag{3}$$

This multi-scale cross-axis design endows the model with strong feature modeling capability. On the one hand, multi-scale convolutions jointly capture local textures and global semantics, mitigating the issue of slender targets being overwhelmed by complex backgrounds. On the other hand, the bidirectional cross-attention mechanism establishes robust spatial correlations, enabling the model to more flexibly handle power facilities with clear directional characteristics. To efficiently adapt this mechanism to outdoor power facility detection, an embedded fusion strategy is adopted in this study to deeply reconstruct the original C3k2 feature extraction unit in YOLOv11, resulting in the proposed C3k2_MCA structure. Through this tightly integrated feature fusion approach, the structure reduces input channel redundancy and maintains computational efficiency while significantly enhancing the model's sensitivity to slender targets such as transmission towers, thereby achieving high-precision detection in complex outdoor environments.

3.2. D-LKA Module

To address the severe scale variations and background interference in outdoor power facility detection, this study introduces the Deformable Large Kernel Attention (D-LKA) mechanism, which fundamentally combines the broad receptive field of large-kernel convolutions with the geometric adaptability of deformable convolutions. To achieve "deformable" perception, D-LKA breaks through the fixed grid limitation of standard convolution by introducing learnable offsets. This mechanism employs lightweight convolutions to dynamically predict the spatial offsets of sampling points in the (x) and (y) directions, and obtains the shifted feature values through bilinear interpolation. Such an adaptive sampling strategy enables the convolution kernel to flexibly adjust its shape, thereby closely fitting the physical structures of irregular targets such as transmission towers. On this basis, D-LKA adopts a combination of depthwise separable convolution (DW-Conv) and depthwise separable dilated convolution to efficiently construct a large receptive field.

The parameter count and computational complexity of D-LKA are calculated as follows. For a two-dimensional input, the kernel size of the depthwise separable convolution is given by:

$$DW = (2d-1)\times(2d-1) \tag{4}$$

The number of parameters is given by:

$$P(K,d) = C\left(\left\lceil \frac{K}{d} \right\rceil^2 + (2d-1)^2 + 3 + C\right) \tag{5}$$

The number of floating-point operations is given by:

$$F(K,d) = P(K,d)\times H\times W \tag{6}$$

The D-LKA module structure integrates residual connections to ensure effective feature propagation. Its computation is formulated as follows:

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

$$x_1 = D - LKA - Attn(LN(x_{in})) + x_{in} \tag{7}$$

$$x_{out} = MLP(LN(x_1)) + x_1 \tag{8}$$

$$MLP = Conv_1(GeLU(Conv_d(Conv_1(x)))) \tag{9}$$

Here, x_{in} represents the input features, *LN* denotes layer normalization, *D-LKA-Attn* is the deformable large kernel attention, *Convd* corresponds to the depthwise convolution, $Conv_1$ is the linear layer, and *GeLU* represents the activation function. The 2D implementation is illustrated in Figure 3:

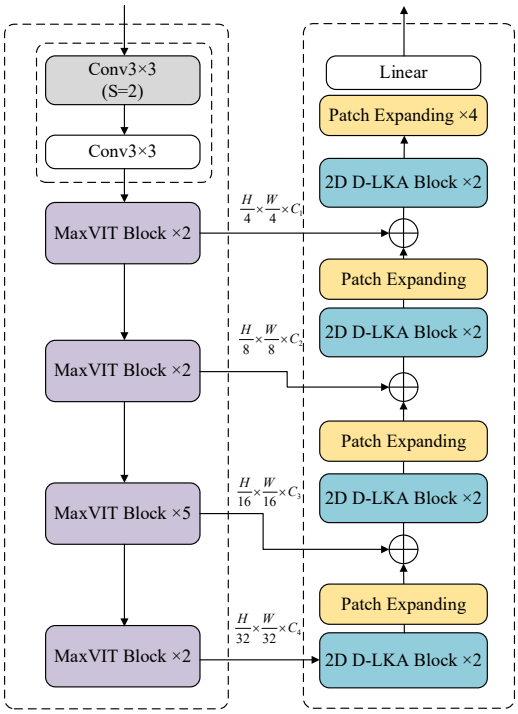

Figure 3: 2D data-Based D-LKA network architecture.

To address the original model's limited capability in perceiving large-scale targets in power facility recognition tasks, this study specifically reconstructs the C2PSA unit in YOLOv11, resulting in the C2PSA_DLKA structure. The original C2PSA unit is constrained by fixed convolution sizes, making it difficult to capture long-range dependencies. By embedding the D-LKA mechanism, the effective receptive field is significantly expanded through the advantages of large kernels, enabling the network to capture richer environmental context. At the same time, leveraging the adaptive sampling characteristics of deformable convolutions, the network can flexibly adjust its attention regions according to the actual shapes of targets such as transmission towers, thereby achieving more precise feature focusing and recognition in complex outdoor scenarios.

3.3. MCAttn Module and Its Improvement

Traditional attention mechanisms, such as the SE method [27], have limited capability for cross-scale feature interaction and often struggle to meet the fine-grained detection requirements of very small-scale power facilities in outdoor scenarios. The Monte Carlo Attention (MCAttn) mechanism draws inspiration from Monte Carlo integration and uses a random sampling strategy to approximate the global attention distribution while maintaining computational efficiency.

In the MCAttn module, the attention generation process is based on a randomly sampled pooling strategy to obtain an attention map that is robust to variations in input feature scales. The method first aggregates features across three different pooled tensors and then randomly selects one 1×1 attention map as the final scale-invariant attention representation. Given an input tensor, the computation of the attention map x is as follows:

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

$$A_m(x) = \sum_{i=1}^{n} P_1(x,i) f(x,i) \tag{10}$$

Here, $A_m(x)$ denotes the output attention map, i represents the size of the attention map, $f(x,i)$ is the average pooling function, and n indicates the number of pooled output tensors. $P_1(x,i)$ satisfies conditions $\sum_{i=1}^{n} P_1(x,i) = 1$ and $\prod_{i=1}^{n} P_1(x,i) = 0$ to generate the attention map.Its basic architecture is shown in Figure 4:

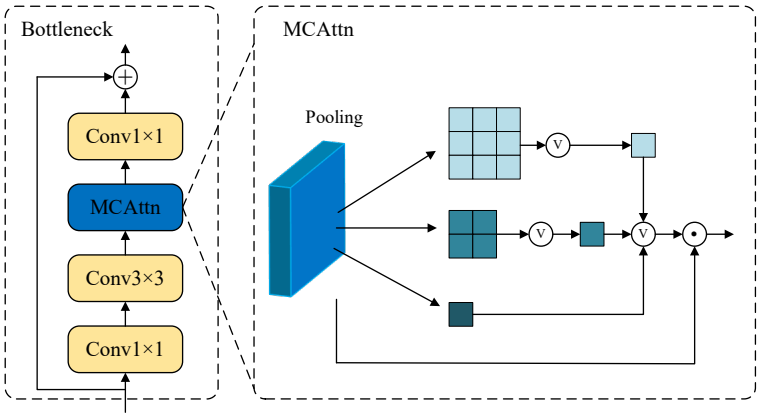

Figure 4: The working architecture of the MCAttn module.

However, directly applying this original design to power facility detection reveals significant shortcomings. On the one hand, the average pooling operation at its core tends to "smooth out" slender features such as power lines amidst large background noise. On the other hand, the original random pixel shuffle disrupts the critical geometric continuity of targets like transmission lines. Therefore, this study reconstructs the MCAttn module and proposes a dual-stream improved architecture incorporating standard deviation awareness. The detailed improvement process is illustrated in Figure 5:

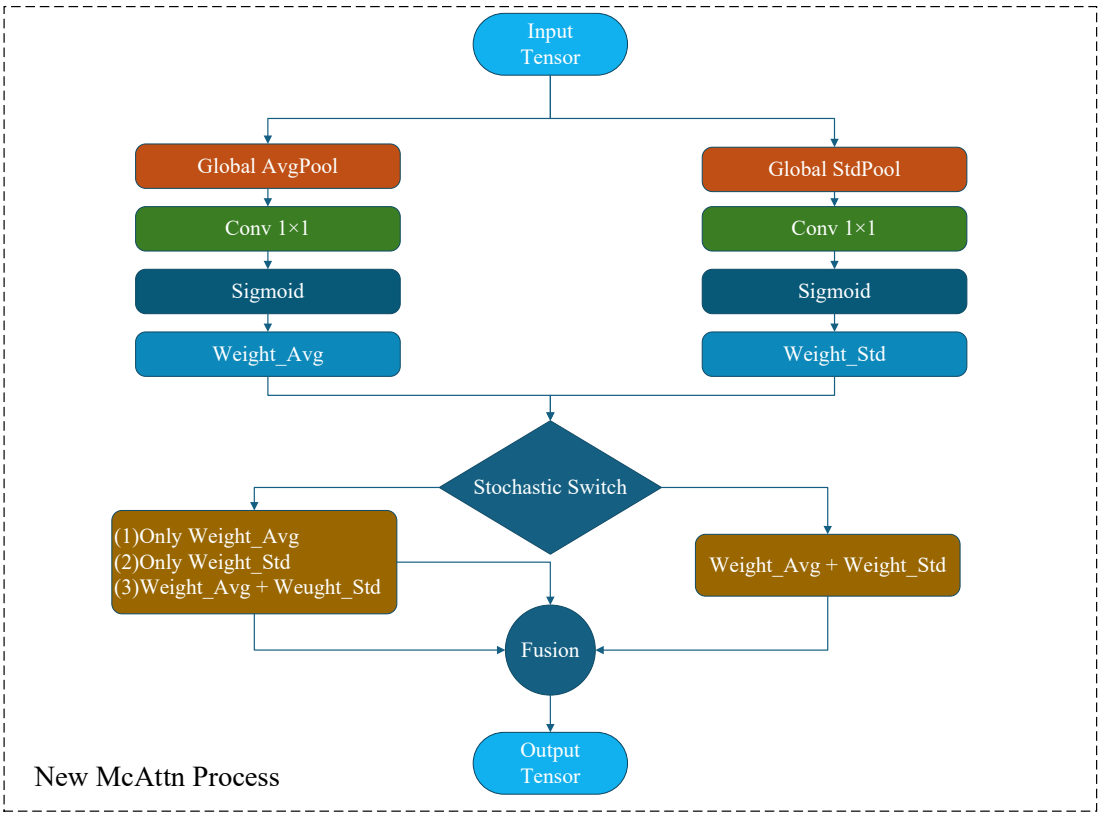

Figure 5: Flowchart of the improved MCAttn module.

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

First, to preserve the complete spatial structure of power facilities, we discarded the destructive random pixel shuffle used in the original mechanism. More importantly, to address the challenge that "thin lines are prone to disappearing" this study constructs a dual-stream parallel pathway based on "mean-standard deviation": while retaining the original mean branch to perceive the overall background distribution, an additional global standard deviation pooling branch is introduced. Because pixel values along the edges of power lines and insulators fluctuate sharply compared to the relatively flat background, standard deviation statistics can sensitively capture these high-frequency edge signals, thereby effectively "extracting" small targets from complex backgrounds.

This improved design not only preserves the lightweight advantage of the original mechanism, making it suitable for deployment on edge devices such as UAVs, but also fundamentally compensates for its shortcomings in maintaining the geometric integrity of slender targets. Experiments show that after integrating this improved architecture, the model achieves significantly enhanced robustness and accuracy in detecting small-scale power facilities under conditions of drastic illumination changes and complex outdoor backgrounds.

3.4. WIoU v3 Loss Function

In constructing datasets for outdoor power facility detection, the training set inevitably contains a certain proportion of low-quality samples—such as tower bases with blurred boundaries or insulators occluded by power lines—due to motion blur from UAV imaging, tree occlusions, and subjective annotation biases. Traditional object detection regression losses, such as CIoU and SIoU, typically employ static weighting strategies, which tend to overemphasize these difficult-to-regress low-quality samples. This can generate harmful gradient interference that misguides the model's convergence and simultaneously suppresses feature learning from high-quality samples.

To address this, this study introduces WIoU v3 (Wise-IoU v3) as an optimization strategy for bounding box regression. The core of this approach lies in its dynamic non-monotonic focusing mechanism, designed to intelligently evaluate anchor quality and allocate gradient gains in a more balanced and effective manner.

The basic IoU is defined as follows: given a predicted bounding box B_p and a ground-truth box B_g, their intersection over union is:

$$IoU = \frac{|B_p \bigcap B_g|}{|B_p \bigcup B_g|} \tag{11}$$

The total loss of WIoU v3 is defined as:

$$L_{WIoU-v3} = r \times L_{IoU} \tag{12}$$

Here, (r) denotes the dynamic regression weight, which adjusts the contribution of different samples during training. The weight is computed as follows:

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \tag{13}$$

Here, β is the outlier degree metric used to evaluate the quality of the anchor box, and α and δ are key hyperparameters that control the distribution of gradient gains.

In this study, the dynamic focusing parameters are set to $\alpha = 1.9$ and $\delta = 3$. This configuration constructs a robust "non-monotonic" gradient gain curve: it maintains strong attention to high-quality samples to improve regression accuracy while allowing the gradient gains of low-quality samples (e.g., blurred or anomalously annotated instances) to decay rapidly. This mechanism aligns well with the heterogeneous quality of outdoor power facility datasets, effectively blocking harmful gradient propagation from low-quality data and ensuring that the model focuses on extracting geometric features from valid samples. As a result, it significantly enhances detection performance and convergence stability in complex background scenarios.

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

4. Experiments

4.1. Dataset

Based on multiple power facility-related data sources, this study constructs a dataset for power facility detection. The original dataset contains a total of 1,242 images, covering typical targets such as transmission towers, utility poles, and insulators under different scenarios, shooting angles, and scale variations. To address the issue of inconsistent and non-standard annotations in the original data that are not fully suitable for object detection tasks, all images were re-annotated in a unified format, forming the base power facility dataset.

To enhance scene diversity under complex weather and environmental conditions, 228 additional images of rainy scenes were synthesized based on the original samples, resulting in a base dataset of 1,470 images. On this basis, various data augmentation strategies were applied to further expand the dataset, including horizontal and vertical flipping, center cropping and scaling, random brightness adjustment, random contrast adjustment, and Gaussian noise injection. These six augmentation methods effectively enriched the sample distribution across different viewpoints and environmental conditions. After augmentation, the dataset size increased from 1,470 to 8,820 images. Finally, the dataset was split into training, validation, and test sets in an 7:1:2 ratio for subsequent model training and performance evaluation. Examples of the dataset images are shown in Figure 6:

Figure 6: Sample images from the dataset.

4.2. Experimental Environment and Parameter Configuration

4.2.1. Experimental Environment Configuration

This experiment was conducted on the Windows 11 operating system. The detailed hardware configuration is shown in the Table1 below:

Table 1: Experimental environment configuration.

Parameters	Configuration
CPU	i7-13700H
GPU	NIVIDIA RTX 4060 8GB
Operating System	Windows11 64
Deep Learing Framework	Pytorch2.0.0
Python	3.8.0

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

4.2.2. Experimental Parameter Configuration

This experiment uses the following parameters: (1) epochs: the total number of training iterations; (2) batch: the number of samples used to compute gradients for each parameter update; (3) imgsz: the input image size; (4) workers: the number of threads used for data loading; (5) device: the device specified for training/inference, where 0 indicates a single GPU; (6) optimizer: the type of optimizer used for parameter updates; and (7) lr0: the initial learning rate at the start of training.

The detailed experimental parameter settings are shown in the Table2 below:

Table 2: Experimental parameter configuration.

Parameters	Settings
Epochs	120
batch	16
imgsz	640
workers	10
device	0
optimizer	SGD
lr0	0.01

4.3. Model Evaluation Metrics

To verify the effectiveness and stability of the model, the trained model was evaluated in terms of performance. By adopting multiple commonly used evaluation metrics, the overall performance of the model was quantitatively analyzed from aspects such as detection accuracy, prediction consistency, and recognition performance across different object categories. The specific evaluation metrics are as follows:

▪ Mean Average Precision (mAP): mAP is the most commonly used comprehensive performance metric in object detection, measuring the overall detection accuracy of the model across all categories. It reflects the model's predictive capability under different evaluation criteria. There are two standards for mAP: mAP@0.5 and mAP@0.5:0.95. mAP@0.5 indicates that a detection is considered correct when the overlap (IoU) between the predicted box and the ground-truth box exceeds 50%. In contrast, mAP@0.5:0.95 represents the average mAP over IoU thresholds ranging from 0.5 to 0.95, providing a more comprehensive assessment of detection accuracy.

▪ Precision: Precision reflects the reliability of the model's predictions, indicating the proportion of true positive samples among all samples predicted as positive. It is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

▪ Recall: Recall measures the model's detection capability, indicating the proportion of true positive samples correctly identified by the model among all actual positive samples. It is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

▪ F1-Score: The F1-Score is the harmonic mean of precision and recall, used to balance the trade-off between the two. By using the harmonic mean, it reflects the model's overall performance in both accurately identifying positive samples and completely detecting all positive samples. The F1-Score is more sensitive to low values; when either precision or recall is low, the F1-Score decreases significantly. Therefore, it effectively represents the model's balance between these two dimensions. It is calculated as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{16}$$

4.4. Training Process Visualization

To analyze the convergence behavior and stability of the proposed model during training, its performance variations throughout the training process were visualized and recorded. During training, the changes in key performance indicators were continuously monitored and plotted against the training epochs, as shown in Figure 7. This visualization provides a dynamic perspective on the model's feature

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

learning, gradient updating, and overall convergence process, offering an intuitive reference for subsequent performance analysis.

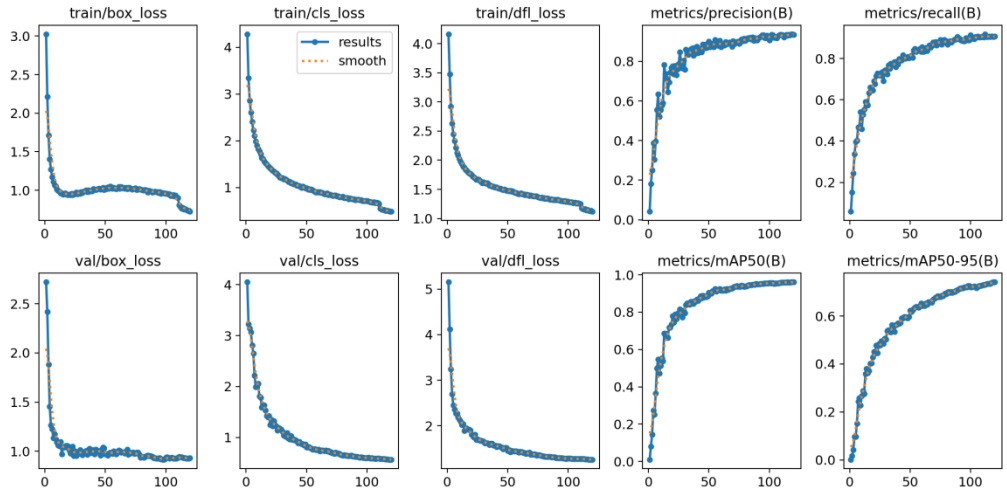

Figure 7: Curves of the model's performance metrics over training epochs during the training process

4.5. Ablation Study

In object detection tasks, improvements in model performance arise from the synergistic operation of various modules. To investigate the specific impact of each proposed improvement on the performance of the YOLOv11n model, a series of ablation studies were designed and conducted. By selectively adding or removing different components based on the baseline model, the effects of each module on the model's performance metrics can be explicitly analyzed. The detailed experimental results are shown in Table 3:

Table 3: Results of the ablation experiment.

Base	C3k2_MCA	MCAttn	C2PSA_DLKA	mAP50@ /%	mAP50-90@ /%	P /%	R /%	F1 /%
✓				91.8	63.8	92.4	82.6	87.2
✓	✓			93.3	66.4	92.7	80.2	86.0
✓		✓		94.8	69.2	93.3	83.9	88.4
✓			✓	92.7	67.7	92.2	83.3	87.5
✓	✓	✓		94.9	70.7	94.3	87.5	90.8
✓	✓		✓	93.4	67.9	91.1	85.6	88.3
✓		✓	✓	94.5	70.5	93.9	87.2	89.3
✓	✓	✓	✓	96.1	74.1	93.8	91.9	92.8

As shown in the table, the baseline model already achieved high detection performance on the power facility dataset, with mAP50 and mAP50-90 reaching 91.8% and 63.8%, respectively. Upon introducing the C3k2_MCA module, both metrics improved by 1.5% and 2.6%, indicating that enhanced multi-scale feature interaction contributes significantly to higher detection accuracy. The individual addition of the MCAttn module yielded even more substantial gains, with mAP50 rising to 94.8%, particularly effective in reducing missed detections in complex scenarios.

The synergy between these modules is evident when they are combined. The integration of C3k2_MCA and MCAttn further boosted the mAP50-90 to 70.7%. Ultimately, incorporating all proposed improvements—C3k2_MCA, MCAttn, and C2PSA_DLKA—resulted in the best overall performance. The mAP50 and mAP50-90 rose to 96.1% and 74.1%, respectively, while the Recall (R) saw a significant increase from 82.6% to 91.9%. These results demonstrate that the structural enhancements complement each other effectively, significantly improving power facility detection performance and robustness while maintaining high model efficiency.

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

4.6. Comparison Experiments

In the comparison experiments, relevant analyses were conducted by comparing different methods' models with the improved model proposed in this study. Several versions of object detection models were selected for evaluation to reveal their performance differences in the power facility detection task and to analyze their applicability and advantages. The detailed experimental results are presented in Table 4:

Table 4: Results of the method comparison experiment.

Methodology	mAP50@ /%	mAP50-90@ /%	P /%	R /%	F1 /%
YOLOv3-tiny	93.4	69.8	94.1	89.1	91.5
YOLOv6	93.2	70.0	91.0	87.3	89.1
YOLOv8n	95.0	73.5	93.9	90.5	92.2
YOLOv9t	94.7	72.4	90.6	93.1	91.8
YOLOv10n	92.4	70.9	83.6	86.4	85.0
SSD	88.3	66.5	87.4	83.2	85.3
Ours	96.1	74.1	93.8	91.9	92.8

As shown in the table, the proposed method achieves the best performance across all key metrics, with mAP50 and mAP50-90 reaching 96.1% and 74.1%, respectively. It maintains a well-balanced precision and recall (93.8% and 91.9%), indicating a significant advantage in detecting power facilities under multi-scale variations and complex background conditions. Among the comparison models, YOLOv8n and YOLOv9t demonstrate strong competitiveness, with YOLOv8n achieving the second-best mAP50 of 95.0%. YOLOv3-tiny and YOLOv6 also perform respectably, both maintaining mAP50 scores above 93%.

In contrast, SSD performs significantly worse than the YOLO series across all metrics, with an mAP50 of only 88.3%, highlighting its limited adaptability to the intricate requirements of outdoor power facility detection tasks. These results further validate that our proposed structural enhancements lead to superior feature extraction and localization capabilities.

4.7. Visualization Analysis

To intuitively demonstrate the performance superiority of the proposed method in outdoor power facility detection, a comparative visual analysis was conducted involving the baseline model and our improved model. To evaluate the detection capabilities across diverse and challenging environments, three representative images were selected from the dataset for visualization. The comparison results are illustrated in Figure 8:

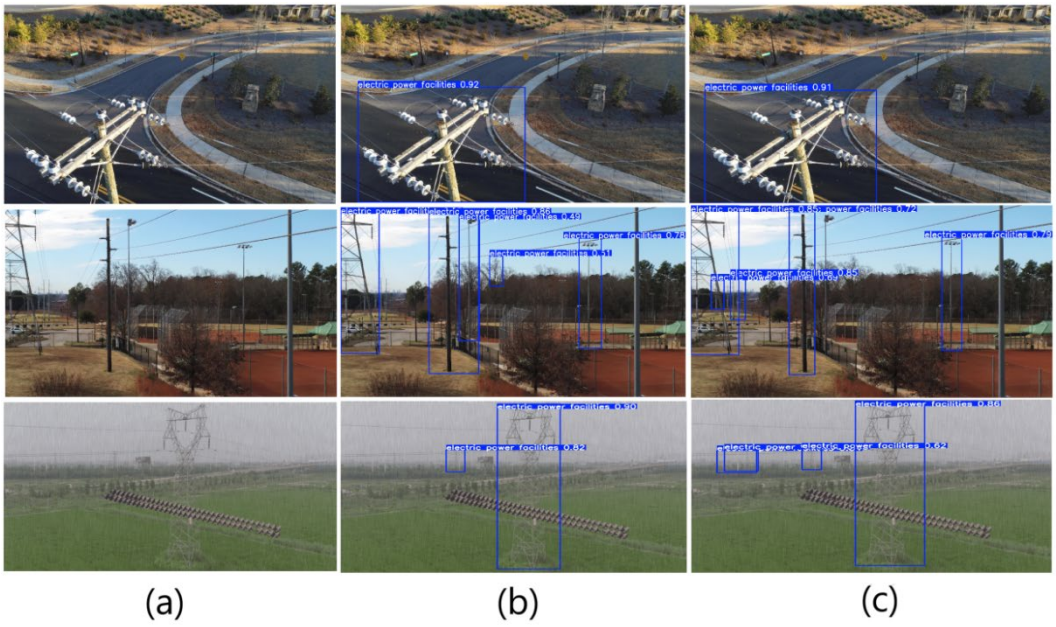

Figure 8: Comparison of results. (a) Oringal Image; (b) YOLOv11n; (c) Ours.

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

4.8. Validation on public datasets

To evaluate generalization ability, experiments were conducted on the KITTI and VisDrone datasets compared with YOLOv11n, and the results are shown in Table 5. The results show that the proposed method achieves performance comparable to the baseline on KITTI, and slightly improves mAP@0.5, precision, and recall on VisDrone while maintaining the same mAP@0.5:0.95. Although optimized for outdoor power facility detection, the method still maintains performance comparable to YOLOv11n on public datasets, demonstrating good generalization capability.

Table 5: Results of different datasets.

Dataset	Models	mAP50@ /%	mAP50-90@ /%	P /%	R /%	F1 /%
KITTI	YOLOv11n	84.2	57.3	89.7	77.9	83.4
	Ours	83.8	56.5	88.5	77.2	82.5
VisDrone	YOLOv11n	23.3	12.1	33.0	25.3	28.7
	Ours	23.9	12.1	35.2	25.8	29.8

5. Conclusions

This study addresses key challenges in outdoor power facility detection, such as strong environmental interference, large variations in target scale, and slender structures being easily obscured by the background. A reconstructed and augmented power facility detection dataset was developed, and an improved YOLOv11n-based detection framework was proposed. By systematically enhancing feature modeling and network architecture, the proposed method demonstrates notable advantages in multi-scale feature modeling, orientation-aware detection of elongated targets, and robust feature extraction under complex backgrounds. Experimental results show that the proposed approach significantly outperforms the baseline model on the self-constructed dataset and maintains comparable performance to YOLOv11n on several public datasets, validating its effectiveness and strong generalization capability. Overall, the improved framework provides a high-precision, real-time solution for detecting outdoor power facilities, offering a practical tool for intelligent inspection and safety assurance in power systems.

References

[1] Korolev I , Zakrevsky A , Vasileva N V .Protection of power facilities personnel from electric fields of industrial frequency[J].2023 5th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE), 2023, 5:1-5.DOI:10.1109/REEPE57272.2023.10086848.

[2] CHEN Jiaqi, SHENG Shuang, LIN Jiaxi, et al. Zero sample power grid equipment ontology defect grade idetification method based on knowledge enhanced large language model[J].Chinese High Technology Letters,2025,35(04):429-439.doi:10. 3772/ j. issn. 1002-0470. 2025. 04. 010.

[3] WANG H, MENG F. Research on power equipment recognition method based on image processing[J/OL]. EURASIP Journal on Image and Video Processing, 2019.

[4] BAI K, ZHOU Y, CUI Z, et al. HOG-SVM-Based Image Feature Classification Method for Sound Recognition of Power Equipments[J/OL]. Energies, 2022, 15(12): 4449.

[5] LEI L, WU J, ZHENG S, et al. Recognition of Power Equipment Based on Multitask Sparse Representation[J/OL]. Scientific Programming, 2021, 2021: 1-7.

[6] ARASTOUNIA M, LICHTI D. Automatic Object Extraction from Electrical Substation Point Clouds [J/OL]. Remote Sensing, 2015: 15605-15629.

[7] ENGELCKE M, RAO D, WANG D Z, et al. Vote3Deep: Fast object detection in 3D point clouds using efficient convolutional neural networks[C/OL]//2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, Singapore. 2017.

[8] YU W, XI J, WU Z, et al. A METHOD FOR EXTRACTING SUBSTATION EQUIPMENT BASED ON UAV LASER SCANNING POINT CLOUDS[J/OL]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2020, XLIV-4/W3-2020: 413-419.

[9] ZHANG R, YANG B, XIAO W, et al. Automatic Extraction of High-Voltage Power Transmission Objects from UAV Lidar Point Clouds[J/OL]. Remote Sensing, 2019: 2600.

[10] ARASTOUNIA M, LICHTI D D. Segmentation of planar surfaces in LiDAR point clouds of an electrical substation by exploring the structure of points neighbourhood[J/OL]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2014, XL-5: 55-62.

[11] YUAN Q, LUO Y, WANG H. 3D point cloud recognition of substation equipment based on plane

ISSN 2616-5775 Vol. 9, Issue 3: 17-29, DOI: 10.25236/AJCIS.2026.090303

detection[J/OL]. Results in Engineering, 2022: 100545.
[12] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C/OL]//2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA. 2014.
[13] GIRSHICK R. Fast R-CNN[C/OL]//2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile. 2015.
[14] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection[C/OL]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2016.
[15] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger[C/OL]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI. 2017.
[16] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[M/OL]//Computer Vision – ECCV 2016,Lecture Notes in Computer Science. 2016: 21-37.
[17] ZHOU F, MA Y, WANG B, et al. Dual-channel convolutional neural network for power edge image recognition[J/OL]. Journal of Cloud Computing, 2021, 10(1).
[18] XIONG X, XU S, WU W, et al. Identification of Electrical Equipment Based on Faster LSTM-CNN Network[C/OL]//2020 IEEE International Conference on Networking, Sensing and Control (ICNSC), Nanjing, China. 2020:1-6.
[19] WANG H, YANG G, LI E, et al. High-Voltage Power Transmission Tower Detection Based on Faster R-CNN and YOLO-V3[C/OL]//2019 Chinese Control Conference (CCC), Guangzhou, China. 2019.
[20] Wu Y , Xiao F , Liu F ,et al.A Visual Fault Detection Algorithm of Substation Equipment Based on Improved YOLOv5[J].Applied Sciences (2076-3417), 2023, 13(21).
[21] Wang Q , Yang L , Zhou B ,et al.YOLO-SS-Large: A Lightweight and High-Performance Model for Defect Detection in Substations[J].Sensors (14248220), 2023, 23(19).
[22] Yang Y , Yang S , Li C ,et al.Insulator defect detection under extreme weather based on synthetic weather algorithm and improved YOLOv7[J].High Voltage, 2025, 10(1)
[23] Shao, H., Zeng, Q., Hou, Q. et al. MCANet: Medical Image Segmentation with Multi-scale Cross-axis Attention. Mach. Intell. Res. 22, 437–451 (2025).
[24] AZAD R, NIGGEMEIER L, HUTTEMANN M, et al. Beyond Self-Attention: Deformable Large Kernel Attention for Medical Image Segmentation[J]. Computer Vision and Pattern Recognition,2023.
[25] Dai W , Liu R , Wu Z ,et al.Exploiting Scale-Variant Attention for Segmenting Small Medical Objects[J]. Image and Video Processing,2024.
[26] TONG Z, CHEN Y, XU Z, et al. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism[J]. Computer Vision and Pattern Recognition,2023.
[27] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-Excitation Networks[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020: 2011-2023.