Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3145

INTELLIGENT DETECTION METHOD FOR FOREIGN OBJECTS ON OVERHEAD LINES OF URBAN BUILDINGS

ZhiPeng Li[1], ZhiYong Liu[2*], ZhaoYun Liu[3], JiaXing Hao[4], ShuJun Yang[4]
[1]*School of Civil Engineering, Shijiazhuang Tiedao University, Shijiazhuang 050043, Hebei, China.*
[2]*Shijiazhuang Architectural Design Institute Co., Ltd., Shijiazhuang 050011, Hebei, China.*
[3]*Shenxing Cable Group Co., Ltd., Jinzhou 052260, Hebei, China.*
[4]*School of Electrical and Electronic Engineering, Shijiazhuang Tiedao University, Shijiazhuang 050043, Hebei, China.*
[*]*Corresponding Author: ZhiYong Liu*

Abstract: To address practical engineering challenges of overhead power lines near urban residential buildings, including low erection height, close proximity to building facades, lightweight litter such as plastic bags easily tangled on cables, complex background interference from building walls, vegetation and outdoor air-conditioning units, as well as variable shapes of tangled foreign objects, tiny target sizes and imbalanced scarce fault samples, this paper proposes an intelligent foreign object detection method for overhead lines combining improved EfficientDet and improved Vision Transformer (ViT). Firstly, the improved EfficientDet is adopted to locate regions containing bird nests and tangled plastic bags. Dilated convolutions are embedded into shallow backbone layers P1 and P2 to enlarge the receptive field, and the CBAM attention mechanism replaces the original SE module to strengthen the extraction of contour features of foreign objects under cluttered backgrounds. Secondly, the improved Vision Transformer performs refined discrimination on the presence or absence of foreign objects based on cropped local images. The rigid hard patch embedding of raw images is substituted by multi-layer small convolutions, the number of multi-head attention heads is optimized, and Focal Loss is introduced to alleviate sample imbalance. Actual test results demonstrate that the mean average precision (mAP) of foreign object detection reaches 96.41%, and the classification accuracy of foreign objects is 96.73%. The proposed method exhibits outstanding stability in typical urban scenarios including backlight, building occlusion and dense shrub interference.
Keywords: Urban overhead lines; Fault detection; Improved EfficientDet; Improved Vision Transformer (ViT)

1 INTRODUCTION

Low-voltage power and communication overhead lines supporting old urban communities and street-side buildings are generally erected at low heights and closely arranged along residential exterior walls. Plastic bags, plastic films and other lightweight domestic wastes discarded by residents are easily wrapped around line surfaces under the action of wind. Long-term attachment of foreign objects will block heat dissipation of cables, accelerate aging and cracking of insulating sheaths, and further induce electric leakage, short circuits, and even fires and power outages, which seriously threaten regional power supply safety and residents' personal and property safety[1]. Traditional operation and maintenance relies on manual climbing and ground visual inspection, which is restricted by floor occlusion and blind viewing areas, leading to prominent missed detection of tiny wrapped plastic bags[2]. In addition, high-altitude operations carry high safety risks, making manual inspection unable to meet the large-scale operation and maintenance demands of urban cables.

In recent years, target detection and image classification algorithms have been widely applied in power and railway industries. For small target localization, EfficientDet achieves a favorable balance between accuracy and inference speed via a lightweight backbone and BiFPN bidirectional feature pyramid. Nevertheless, the original network suffers from insufficient receptive fields in shallow layers, and channel-only attention fails to focus on slender cables. In terms of state recognition, Vision Transformer can capture long-distance global dependencies and is suitable for learning cable structural features. However, the original block embedding tends to lose detailed information, and cross-entropy loss cannot handle imbalanced fault samples.

Early overseas research on overhead line foreign object detection mostly adopted traditional handcrafted feature algorithms such as SIFT and HOG to extract line contours for foreign object screening. However, handcrafted features exhibit poor generalization and are only applicable to open wild transmission scenarios, failing to adapt to densely built urban environments. Recent foreign studies mostly improve line detection models based on YOLO and Transformer series[3-7]. Assad S et al. proposed an improved YOLOv8 algorithm adopting deformable convolutions to optimize the recognition of wild kites and bird nests, yet their datasets mainly cover open-field high-voltage lines without optimization for low-rise residential cables and floating wrapped plastic bags, resulting in high false detection rates of plastic bags under interference from building walls and dense green plants. Du et al. adopted a hybrid CNN+Transformer architecture to optimize small target detection of transmission lines, focusing on large floating protective nets in transmission corridors without special training samples for domestic waste plastic bags[8].

Domestic universities and power research institutes have conducted extensive research on aerial foreign object detection, while most existing achievements focus on high-voltage tower lines in wild areas. Xie Guobo et al. put

forward the AAGV-YOLOX model for floating foreign object detection on transmission lines, which is designed for large outdoor dust-proof nets and hanging nets without adaptation to tiny wrapped plastic bag targets in urban areas[9]. Xue Ang adopted an improved YOLOv5 algorithm with window self-attention to boost the precision of small channel foreign objects, but test samples were collected from suburban wild lines lacking real-scene data around old residential buildings, leading to limited recognition performance of thin plastic bags tightly attached to cables[10]. Song Liye realized the detection of insulators and bird nests based on improved EfficientDet to optimize defect recognition of high -voltage fittings, yet the training set contained no samples of plastic bags discarded by residents and cannot be directly applied to urban overhead line inspection[11]. In addition, some studies adopted the combination of original EfficientDet and ViT for line detection. The shallow receptive field of native EfficientDet is relatively small, and the SE attention only performs channel filtering, which cannot eliminate spatial clutter such as building walls and air conditioning outdoor units. Hard image block segmentation of ViT tends to lose fold details of plastic bags, resulting in severe missed and false detection when plastic bags are tiny and closely attached to cables.

In summary, most existing studies target large foreign objects on wild high-voltage lines, while few detection schemes specially address low-rise overhead lines in urban residential buildings and wrapped domestic plastic bags. Based on the unique characteristics of urban areas, this paper constructs a detection framework combining improved EfficientDet and improved ViT to solve the difficult recognition problem of plastic bag foreign objects under complex backgrounds.

2 ALGORITHM PRINCIPLE AND IMPROVEMENT DESIGN

2.1 Improved EfficientDet Localization Network

Proposed in 2019, EfficientNet breaks the limitation of single adjustment among network width, depth and input resolution. It integrates the three dimensions through Neural Architecture Search (NAS) and simultaneously explores their impacts on network performance, as illustrated in Figure 1.

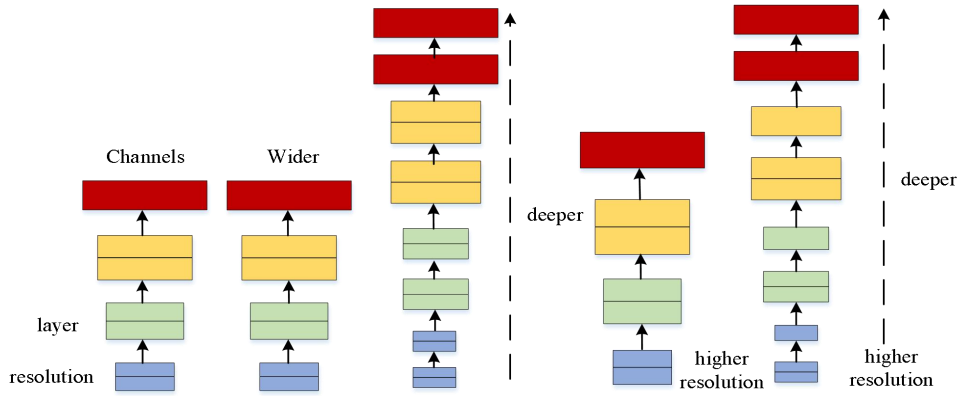

Figure 1 Strategies under Different Parameter Configurations of Width, Depth and other Hyperparameters

In EfficientNet, to obtain the optimal network model, three parameters α , β , γ are used to simultaneously adjust the network from three dimensions: depth, width, and resolution. A compound scaling factor is adopted to unify the configuration of α , β , γ , and the relationship among the four parameters is expressed in Equation (1). There exist certain constraints on α , β , γ : doubling the network depth will double the computational cost, while doubling either the width or resolution will quadruple the computational cost. These constraint relationships are given in Equation (2).

$$\begin{aligned} &depth: d = \alpha^{\phi} \\ &width: w = \beta^{\phi} \\ &resolution: r = \gamma^{\phi} \end{aligned} \tag{1}$$

$$\begin{aligned} &s.t. \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ &\alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned} \tag{2}$$

When the compound scaling factor is set to 1 in EfficientNet, the optimal combination of three parameters α , β , γ for peak network performance is searched out, and the corresponding parameter values $\alpha = 1.2$, $\beta = 1.1$, $\gamma = 1.15$ constitute the baseline model EfficientNet-B0. MBConv serves as the core component of EfficientNet, which is designed based on inverted residuals. Specifically, a 1×1 convolution is utilized to raise the feature dimension prior to the 3×3 or 5×5 depthwise separable convolution. After the depthwise separable convolution, the Squeeze-and-Excitation (SENet) attention module is embedded. Subsequently, another 1×1 convolution is applied for dimension reduction, followed by a residual shortcut connection. The MBConv structure in EfficientNet is illustrated in Figure 2.

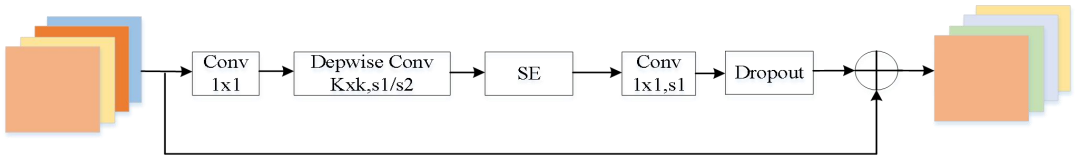

Figure 2 Structure of MBConv Module

The structure of the SE module is presented in Figure 3, where Swish and Sigmoid activation functions are applied after the fully connected layers respectively.

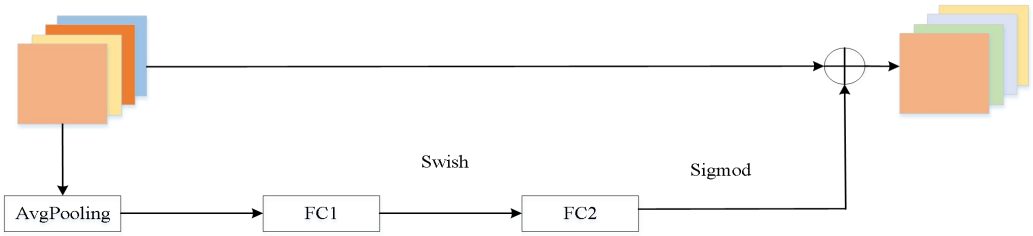

Figure 3 Structure of SE Module

In the detection of foreign objects on overhead lines around urban buildings, captured images contain complex backgrounds, requiring the model to focus more on target areas and edge information for complete feature extraction. The original EfficientDet only adopts SE channel attention, which merely weights feature channels and easily causes information loss during dimension compression, making it hard to capture global features. This paper introduces the CBAM attention mechanism in Figure 4, which generates attention weights from channel and spatial dimensions to adaptively enhance target features and suppress background interference, effectively improving localization accuracy under complex scenarios.

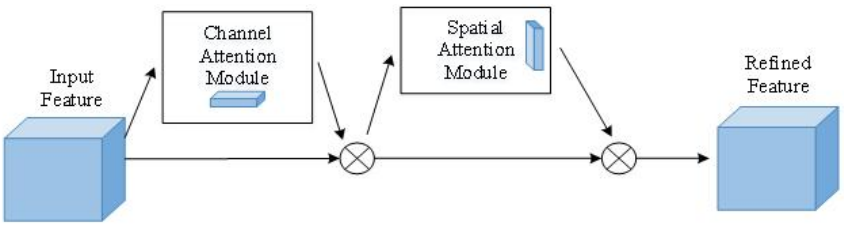

Figure 4 Structure of CBAM Module

The CBAM module consists of a channel attention submodule and a spatial attention submodule. The input feature map is first fed into the channel attention module to generate a channel attention map, which is multiplied with the original feature map by element-wise product to obtain a weighted feature map. The weighted feature map then passes through the spatial attention module for the same operation to output the final feature map, whose mathematical expression is given in Formula (3).

$$PE(pos,2i)=\sin\left(\frac{pos}{1000^{\frac{2i}{d_{\mathrm{mod}el}}}}\right) \tag{3}$$

When $\otimes$ denotes element-wise multiplication; F represents the input feature map; $M_c(F)$ is the channel attention map output by the channel attention module; $M_s(F')$ refers to the spatial attention map generated by the spatial attention module; F'' stands for the final feature map output from CBAM.

The detailed procedures are as follows: First, average pooling and max pooling are performed on feature map $M_c(F)$ to obtain two pooled feature maps F^s_{2avg} and $F^s_{2\max}$. Next, these two feature maps are concatenated to form a new feature map with dimension $H\times W\times 2$. The fused feature map is then fed into a convolutional layer to finally generate the spatial attention map $H\times W\times 1$ with dimension W_{ea}.

$$W_{sa}=M_s(F^{ca}_{1})=\sigma(W_{sa1}([F^s_{2avg};F^s_{2\max}])) \tag{4}$$

As a critical output of the model, the attention map assigns weighting coefficients to all pixel regions of the image. Under this mechanism, pixel regions that exert a substantial influence on classification decisions are assigned higher

weights, while regions with minor impacts are given lower weights. Such differentiation not only enhances the model's capability to capture key information but also optimizes the efficiency of feature representation. Finally, feature map F_1^{ca} is multiplied by attention map W_{sa} to output the weighted feature map $F_1^{CBAM} = F_1^{ca} \otimes W_{sa}$.

$$F_1^{CBAM} = F_1^{ca} \otimes W_{sa} \tag{5}$$

After feature map F_1^{CBAM} is weighted via the attention mechanism, it exhibits more distinct feature discrimination for classification decisions compared with the original feature map $M_c(F)$. This mechanism effectively improves the model's ability to capture critical information, enables more flexible and intelligent processing of complex input data, and boosts the model's performance and generalization capacity. It guarantees that classification decisions rely more accurately on salient features, thereby elevating the overall classification accuracy.

This paper selects lightweight EfficientDet-B0 for experiments. Combined with the feature layer distribution characteristics of the network, CBAM replaces the original SE attention mechanism on layers P1~P5 to optimize feature extraction from both channel and spatial dimensions, expand the receptive field without increasing parameter quantity, and fully mine detailed information from high-resolution feature maps. The structure of the improved EfficientDet network is shown in Figure 5.

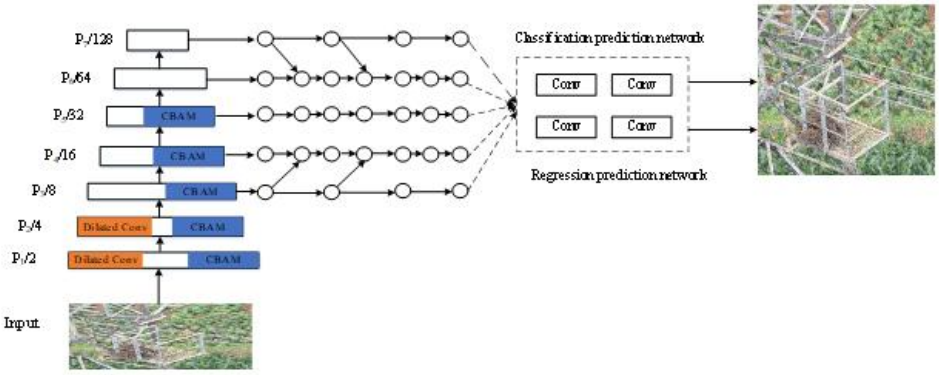

Figure 5 Network Structure of the Improved EfficientDet

2.2 Improved Vision Transformer Network

Proposed by Alexey Dosovitskiy et al. in 2020, Vision Transformer (ViT) is a classification network different from CNN series. It breaks the traditional cognition of attention mechanism application and directly migrates Transformer encoder modules without relying on CNNs, realizing pure transformer deployment on image patch sequences and achieving high accuracy in image classification tasks. The structure of ViT is illustrated in Figure 6.

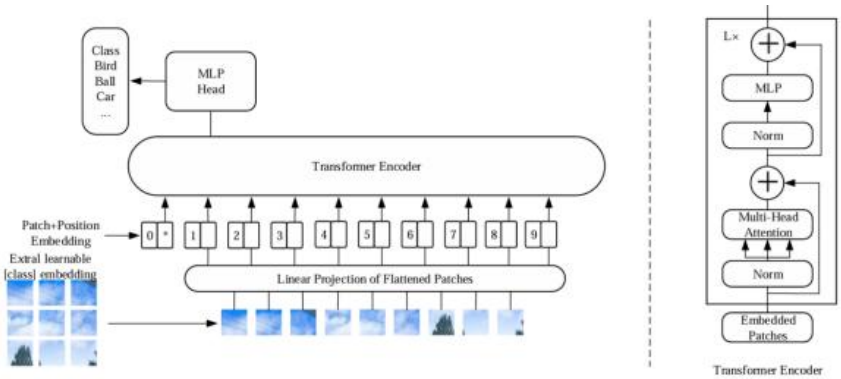

Figure 6 Network Structure of ViT

ViT mainly comprises three components: embedding layer, Transformer encoder and classification head. The network first divides a 224×224 image into 16×16 image patches, maps them into feature vectors via linear projection, and superimposes position encoding and class tokens. Multi-head attention within the encoder extracts multi-dimensional features, and the classification head outputs recognition results.

The attention mechanism of Transformer cannot perceive the positional information of sequences, and disrupting the input order will lead to errors in feature parsing. Therefore, sinusoidal positional encoding is introduced in this paper to supplement positional information for feature vectors and guarantee the inference accuracy of the model.

$$PE(pos,2i)=\sin\left(\frac{pos}{1000^{\frac{2i}{d_{\mathrm{mod}el}}}}\right) \tag{6}$$

$$PE(pos,2i+1)=\cos\left(\frac{pos}{1000^{\frac{2i}{d_{\mathrm{mod}el}}}}\right) \tag{7}$$

In the formula, pos denotes the absolute position of the element in the vector; $pos=0,1,2,...$; $d_{\mathrm{mod}el}$ represents the dimension of the vector; $2i$ and $2i+1$ indicate odd and even indices respectively; i refers to the index of the dimension of the element vector.

2.2.1 Improved multi-head attention mechanism

The number of multi-head attention heads (Num-head) affects the correlation extraction of spatial information, position information and feature points within Transformer Encoder. Multi-head attention can be interpreted as mapping original image information to different subspaces, enabling the encoder to capture feature information from multiple subspaces. Different Num-head values correspond to different focused subspaces, exerting influences on the detection accuracy of small targets.

Multi-head attention integrates multiple single-head attention modules to simultaneously identify and decide target feature information from multiple perspectives, and the final decision is obtained via weighted summation of individual results.

$$MultiHead(Q,K,V)=Concat(head_1,...head_n)W^0 \tag{8}$$

$$WhereHead_i=Attention(QW_i^Q,KW_i^K,VW_i^V) \tag{9}$$

In the formula, K and V are matrices generated by the encoder module; Q denotes the matrix generated by the decoder; W represents the weight coefficient matrix.

2.2.2 Improved loss function

Cross-entropy loss is widely adopted in image classification tasks, assigning equal loss weights to all training samples, and it also serves as the original loss function of ViT. The expression of cross-entropy loss is given in Formula (10):

$$CE(p,y)=\begin{cases}-\log(p) & ify=1\\ -\log(1-p) & ify=0\end{cases} \tag{10}$$

In the formula, $y=1$ denotes positive samples and $y=0$ denotes negative samples.

Let the probability of correct prediction p for a sample be defined, the binary classification loss function is Formula (11)

$$CE(p,y)=\begin{cases}-\log(p) & ify=1\\ -\log(1-p) & ify=0\end{cases} \tag{11}$$

The loss function for binary classification is shown in Equation (12):

$$L=\frac{1}{N}\sum_i L_i(y_i,p_i)=\frac{1}{N}\sum_i -y_i\log p_i-(1-y_i)\log(1-p_i) \tag{12}$$

The multi-classification loss function is Formula (13):

$$L=\frac{1}{N}\sum_i L_i=-\frac{1}{N}\sum_i\sum_{c=1}^{M}y_{ic}\log(p_{ic}) \tag{13}$$

In the formula, y_i takes the value of 1 for positive samples and 0 for negative samples; p_i represents the probability that sample i is predicted as a positive sample.

Images of urban overhead lines collected in practice contain far more normal cable samples than foreign object samples, resulting in severe sample imbalance. Meanwhile, some foreign objects possess tiny shapes with subtle feature differences from normal cables, belonging to hard-to-identify samples. Traditional cross-entropy loss allocates identical weights to all samples, easily causing missed and false detection. This paper replaces the original loss function with Focal Loss, which dynamically adjusts sample weights, reduces weights of easily-classified samples and elevates the training proportion of hard samples to alleviate sample imbalance and strengthen the model's learning ability of foreign object features, as shown in Formula (15).

$$FL(p_t)=-(1-p_t)^{\gamma}\log(p_t) \tag{14}$$

It can be seen from Equation (15) that unlike the original cross-entropy loss function, Focal Loss introduces two additional parameters p_t and γ. The function of parameter p_t is as follows: if the prediction result of a sample is nearly accurate (i.e., p_t approaches 1), the loss contributed by this sample will converge to zero. Parameter γ controls the convergence rate of the loss; different values of γ lead to different decay speeds of the loss. Meanwhile, a balance weight factor α_t is introduced to address the imbalance between positive and negative samples. Accordingly, the final formula of Focal Loss is defined in Equation (15).

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{15}$$

The overall structure of the improved ViT is presented in Figure 7.

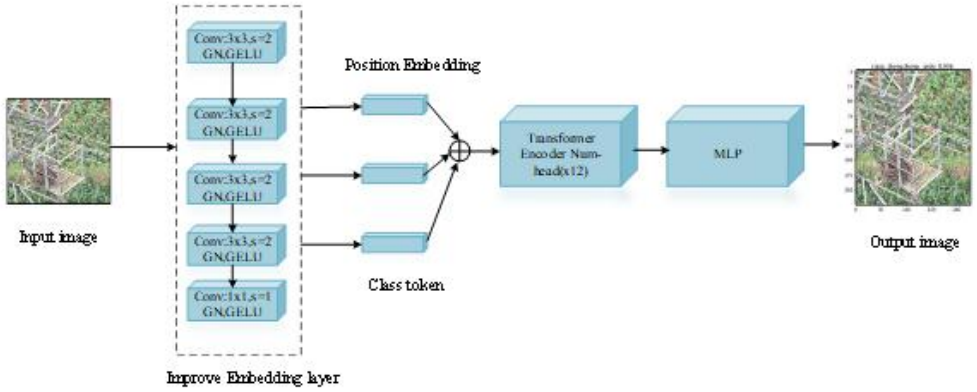

Figure 7 Structure Diagram of Improved ViT

3 ALGORITHM IMPLEMENTATION

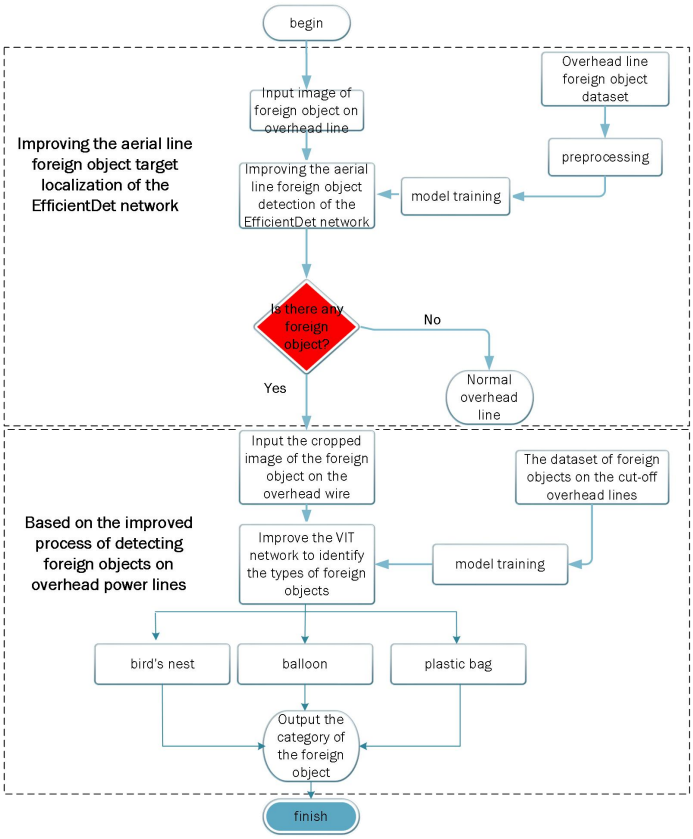

Figure 8 Detection Flow of Foreign Objects on Overhead Lines of Urban Buildings Based on Improved EfficientDet and Improved ViT

The detection flow of foreign objects on urban building overhead lines based on improved EfficientDet and improved ViT proposed in this paper is displayed in Figure 8, with specific steps as follows:

(1) Preprocess input images and feed them into the improved EfficientDet localization network;
(2) Judge whether foreign objects exist in the frame. Terminate the flow if no foreign objects are detected; otherwise, crop foreign object regions according to bounding boxes;
(3) Resize cropped foreign object images uniformly to 224×224;
(4) Normalize images and input them into the improved Vision Transformer for foreign object discrimination;
(5) Extract feature vectors via the optimized embedding layer, superimpose position encoding and class tokens, and feed into the Transformer Encoder;
(6) Fuse multi-subspace features through multi-head attention to focus on effective foreign object features and filter background interference;
Output the detection result of plastic bag foreign objects via the MLP head to complete the recognition task.
To evaluate the effectiveness of the proposed insulator fault detection method in this paper, Recall (R), Precision (P), Frames Per Second (FPS), mean Average Precision (mAP), and F1-score (the harmonic mean of Precision and Recall) are adopted as model evaluation metrics. The corresponding calculation formulas are shown in Equations (16), (17), (18), (19) and (20).

$$p = \frac{TP}{TP + FP} \tag{16}$$

$$R = \frac{TP}{TP + FN} \tag{17}$$

$$AP = \int_0^1 p(R)d(R) \tag{18}$$

$$mAP = \sum_{i=1}^{class} \frac{AP(i)}{class} \tag{19}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{20}$$

In the formula, AP represents the precision of a single category; TP denotes the number of true positive targets correctly predicted by the classifier; FP refers to the number of false positive targets mistakenly predicted as positive by the classifier; FN stands for the number of true negative targets correctly identified by the classifier.

4 EXPERIMENT AND RESULT ANALYSIS

All experiments are built on a deep learning framework with the following hardware configurations: Intel Core i7-12700H CPU, 32 GB RAM and NVIDIA RTX3060 12 GB discrete graphics card. Training hyperparameters are set as follows: Batch Size = 16, Epochs = 100, initial learning rate = 0.001, cosine annealing learning rate decay strategy, Adam optimizer, weight decay coefficient = 5e-4. Data augmentation and gradient freezing training are enabled during training to prevent overfitting and improve model generalization. All comparative experiments are completed under identical configurations to guarantee reliable experimental results.
To verify the performance advantages of the improved ViT for overhead line foreign object detection, four classic baseline models including original ViT, VGG16, AlexNet and ResNet-50 are selected for comparative tests. Accuracy, Precision, Recall, F1-score and loss value are taken as core evaluation indicators to quantitatively analyze the detection performance and fitting effect of each model comprehensively. The loss curves and accuracy curves of all networks are shown in Figure 9 and Figure 10 respectively, and quantitative evaluation results of five models are summarized in Table 1 to intuitively verify the comprehensive superiority of the improved ViT.

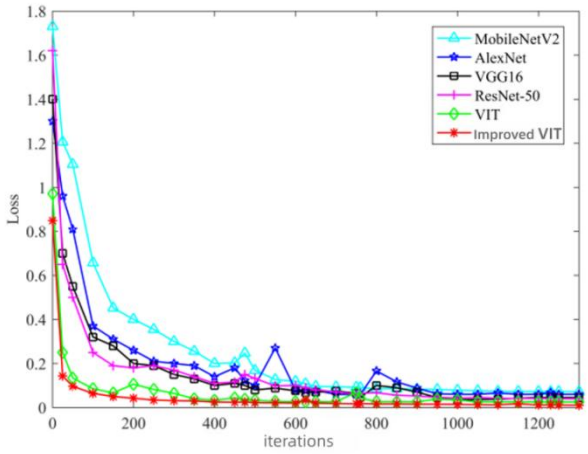

Figure 9 Loss Comparison Diagram of Different Networks

As shown in Figure 9, AlexNet and VGG16 possess an initial loss value of approximately 1.4 with obvious fluctuations during training. AlexNet exhibits more severe oscillation and slower convergence, requiring nearly 1000 iterations to stabilize with final loss values of 0.060 and 0.045 respectively. The final stable loss values of MobileNetV2 and ResNet-50 are 0.072 and 0.040. The original ViT starts with an initial loss of around 1 and converges rapidly in the early stage, yet the training curve contains minor fluctuations with poor smoothness, achieving a final loss of 0.025. The proposed improved ViT has an initial loss of only 0.9, dropping sharply within the first 200 iterations and converging smoothly afterwards without obvious oscillation. It achieves superior training stability and fitting performance with a final loss as low as 0.010.

The classification accuracy curves of improved ViT, original ViT, VGG16, AlexNet and ResNet-50 are plotted in Figure 10.

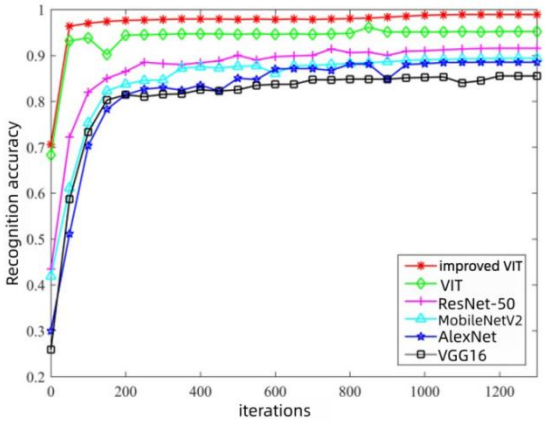

Figure 10 Accuracy Comparison Chart of Different Networks

It can be observed from the training loss curves in Figure 10 that the initial loss values of AlexNet and VGG16 are both around 1.4, accompanied by severe oscillation throughout the entire training process. Among them, the curve of AlexNet fluctuates most drastically with low convergence efficiency. Both models require nearly 1000 iterations to converge, with final converged loss values of 0.06 and 0.045 respectively. After training, MobileNetV2 and ResNet-50 achieve stable final loss values of 0.072 and 0.04 separately. The original ViT has an initial loss of 1. Although it converges rapidly in the early training stage, its curve suffers minor oscillations during training, leading to insufficient convergence stability, and the final converged loss reaches 0.025. In contrast, the improved ViT proposed in this paper reduces the initial loss to 0.9. The loss drops sharply within the first 200 iterations, and the curve remains smooth without obvious fluctuations in subsequent iterations. It delivers superior training stability and fitting performance, with a final converged loss as low as 0.01.

The recognition results of various foreign objects are shown in Figure 11. It can be seen that the method proposed in this paper can effectively detect the positions of foreign objects.

(a) Bird's Nest (b) Plastic Bag

Figure 11 Recognition Results of Various Foreign Objects

Table 1 Quantitative Performance Comparison of Different Network Models

Model	Accuracy (%)	Precision (%)	Map(%)	F1-score (%)	Loss
AlexNet	89.24	88.87	88.15	88.51	0.060

VGG16	85.57	86.21	86.96	87.08	0.045
ResNet-50	87.82	88.53	87.27	88.40	0.040
Original ViT	93.15	92.89	92.62	92.75	0.025
Improved ViT	96.73	96.56	96.41	96.48	0.010

Compared with the original ViT, the improved ViT gains 3.58% higher Accuracy, 3.67% higher Precision, 3.79% higher Recall and a 0.037 higher F1-score. Against VGG16, the increments of Accuracy, Precision, Recall and F1-score reach 11.16%, 10.35%, 9.45% and 0.094 respectively. Compared with AlexNet, the four metrics are promoted by 7.49%, 7.69%, 8.26% and 0.080. Relative to ResNet-50, Accuracy, Precision, Recall and F1-score increase by 8.91%, 8.03%, 9.14% and 0.081 separately. In general, the improved ViT outperforms all comparative models on all detection metrics and loss convergence performance. The performance improvement originates from replacing hard image patch embedding with multi-layer small convolutions to optimize the fusion of shallow and deep features and strengthen local feature correlation.

5 CONCLUSION

Aiming at the problems of low detection accuracy and poor training convergence stability of overhead line foreign objects under complex urban building backgrounds, multiple comparative experiments are conducted between the improved ViT and four baseline models including AlexNet, VGG16, ResNet-50 and original ViT. Combined with quantitative indicators and loss convergence curves, the improved ViT proposed in this paper achieves significantly better detection accuracy, training stability and anti-interference ability than traditional CNNs and original ViT, delivering optimal comprehensive performance and satisfying the detection demands of tiny power targets under complex urban backgrounds. The core improvements and innovations supported by experimental data are summarized as follows:

(1) The improved ViT achieves Accuracy of 96.73%, Precision of 96.56%, Recall of 96.41% and F1-score of 96.48%. Compared with the original ViT, the four core metrics increase by 3.58%, 3.67%, 3.79% and 0.037. Relative to ResNet-50 (the best-performing traditional CNN), Accuracy, Precision, Recall and F1-score are improved by 8.91%, 8.03%, 9.14% and 0.081, effectively solving the low recognition accuracy of tiny overhead line targets under complex backgrounds existing in AlexNet and VGG16.

(2) Multiple optimization strategies are integrated to boost model feature extraction and generalization capacity. Four stacked 3×3 small convolutions replace the original hard patch embedding to realize smoother feature extraction and sufficient fusion of shallow details and deep semantic features. The multi-head attention mechanism with optimized head quantity is tailored for overhead line foreign object detection to extract target features accurately and suppress background noise from building facades and dense vegetation. Focal Loss is introduced to mitigate sample imbalance and hard sample identification difficulties, greatly enhancing the model's adaptability to occluded, blurry tiny foreign objects on overhead lines.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Zang Peng. Research on Fault Location Method of Cable-Overhead Line Hybrid Transmission Line. Xi'an University of Science and Technology, 2022.

[2] Hu Jiajun. Research and Implementation of Target Detection System Based on Bidirectional Feature Fusion EfficientDet. University of Electronic Science and Technology of China, 2023.

[3] Liang Yubin, Shen Rongfeng, Cheng Chengmao, et al. Weed Detection Model of Young and Middle-Aged Mountain Forest Based on Improved YOLOv8. Forest Engineering: 1-12 [2026-06-23].

[4] Aulton C, Chiu Y C, Chiou Y S. Advancing functional reach assessments: a comparison of YOLOv8 human pose estimation and 3D motion capture in a young healthy cohort. Journal of Biomechanics Open, 2026, 1(2): 100007.

[5] Mela L J, Sánchez G C. Integrating dynamic convolution and channel attention into YOLOv8 for enhanced maritime vision. Digital Signal Processing, 2026, 182: 106258.

[6] Talu H M, Baybars C S, Aboalqaraya R, et al. Automated Detection of Taurodontism in Panoramic Radiographs Using a YOLOv8-Based Deep Learning Model. Journal of Imaging Informatics in Medicine, 2026: 1-14 (Prepublish).

[7] Zeng Zhi, Li Xiaofeng, Wu Jialu, et al. Foreign Object Detection Method for Transmission Lines Based on Improved YOLOv8. Software Guide: 1-16[2026-06-23].

[8] Assad S, Isa M A N, Saleh M A S. Hybrid CNN-Transformer models for industrial defect detection: A systematic review. Results in Engineering, 2026, 29: 109457.

[9] Xie Guobo, Xia Wei, Lin Zhiyi, et al. Research on Transmission Line Foreign Object Detection Based on AAGV-YOLOX Model. Journal of Guangdong University of Technology, 2026, 43(2): 81-90.
[10] Xue Ang, Jiang Enyu, Zhang Wentao, et al. Foreign Object Detection in Transmission Line Corridors Based on Fusion of Window Self-Attention Network and YOLOv5. Journal of Shanghai Jiao Tong University, 2025, 59(3): 413-423.
[11] Song Liye, Liu Shuai, Wang Kai. Power Grid Component and Defect Recognition Method Based on Improved EfficientDet. Transactions of China Electrotechnical Society, 2022, 37(9): 2241-2251.