Tower Foreign Object Recognition Based on Large Model and Deeplabv3+

Xin Li, Yan Wang, Ying Wang, Ruizhi Zhang, and Guoliang Feng

***Abstract*—This study proposes an advanced foreign object detection system for power lines, integrating DeepLabv3+ with large language models (LLMs) to enhance both image segmentation and semantic understanding. Unlike traditional detection methods, our approach not only identifies foreign objects but also provides detailed textual descriptions, enabling maintenance personnel to make informed decisions. Experimental results demonstrate that our method improves overall accuracy from 94.34% to 94.75%, effectively reduces false positive rates, and maintains high detection precision. This hybrid approach significantly enhances operational efficiency, particularly in resource-constrained environments, and provides a scalable solution for intelligent power system maintenance.**

***Index Terms*—Large language model (LLM), Deeplabv3+, Semantic segmentation, Risk assessment, Image processing, Feature extraction.**

I. Introduction

POWER transmission towers serve as essential infrastructure for the stable operation of power grids, acting as both structural supports and junctions for transmission lines. In modern smart grids, shared towers that integrate 5G base stations further extend their role into communication networks. However, continuous exposure to harsh outdoor environments makes these towers vulnerable to various potential faults and external interferences, posing significant risks to the reliability of both power and communication systems. Ensuring their operational safety is therefore a critical challenge in power system maintenance.

Traditional inspection methods—including manual inspections [1], helicopter-based monitoring [2], drone surveys [3], and online video surveillance [4]—are often inefficient and suffer from high false positive rates, making them inadequate for modern power system maintenance. Moreover, these methods struggle to adapt to complex real-world conditions, where real-time and high-precision detection is essential for ensuring system reliability.However, relying solely on manual inspections is increasingly inadequate for modern power system maintenance due to the high demand for human and material resources, as well as lack of timeliness. Sun et al [5] proposed a combined inspection model using drones and manual efforts, significantly reducing labor consumption. Jiang et al [6] demonstrated that helicopter inspections are highly effective in extracting transmission lines from complex backgrounds, thus fulfilling circuit inspection needs. Meng et al [7] designed an automatic transmission line inspection system based on drone imagery. This system autonomously adjusts the drone's flight path and utilizes visual algorithms and convolutional neural networks (CNNs) to detect line defects, employing the ELU nonlinear activation function for automated inspections. Jin et al [8] applied machine vision technology to capture stereo images of high-voltage transmission lines and calculate the disparity to derive ice thickness, enabling real-time monitoring of ice accumulation on transmission lines.

In the domain of power transmission system safety, Zongqi et al [9] proposed an image recognition technique based on the YOLOv2 network, leveraging its powerful learning capabilities for accurate detection of transmission line equipment. Gulzar et al [10] designed a mobile robot capable of traveling along overhead transmission lines under live conditions, equipped with a camera to inspect and transmit data back to the ground. Faiyaz et al [11] proposed an autonomous visual detection method utilizing deep learning, with aerial images captured by drones as the primary data source. Their approach addresses challenges such as limited training data and insulator defect detection.

With the continuous expansion and upgrade of power systems, ensuring the safety and stability of transmission lines has become one of the foremost concerns within the power industry. The integration of advanced inspection technologies, including drone-based systems, deep learning, and machine vision, has paved the way for more efficient and accurate monitoring, offering promising solutions for the future of power transmission safety.

Foreign objects on power towers, such as bird nests, kites, and plastic bags, represent significant threats to the reliable operation of power systems. These objects can cause line failures or even lead to severe safety hazards, such as fires. Therefore, the timely and accurate detection of foreign objects on power towers is essential for maintaining the safety of the power grid. Traditional detection methods, such as manual inspections or automated techniques based on basic image processing [12][13], often face challenges such as low efficiency, high false-positive rates, and difficulty in handling diverse and complex scenarios. With the rapid advancement of deep learning and computer vision technologies [14][15][16], large language models (LLMs) have shown considerable potential for image recognition, offering significant advantages in extracting complex features and deep semantic information [21].

Manuscript received November 11, 2024; revised March 29, 2025. This research was funded by Heilongjiang Electric Co., LTD., supporting the state grid science and technology project (No. 522448240002).

Xin Li is a graduate student at the School of Automation Engineering, Northeast Electric Power University, Jilin City, Jilin 132012, China (e-mail: 2037264571@qq.com).

Yan Wang is is an engineer at the Economic and Technical Research Institute of Harbin State Grid Heilongjiang Electric Power Co., Ltd., Harbin, Heilongjiang 150010, China (e-mail: 909075959@qq.com).

Ying Wang is a senior engineer at the Economic and Technical Research Institute of Harbin State Grid Heilongjiang Electric Power Co., Ltd., Harbin, Heilongjiang 150010, China (e-mail: bo831129@126.com).

Ruizhi Zhang is a senior engineer at the Economic and Technical Research Institute of Harbin State Grid Heilongjiang Electric Power Co., Ltd., Harbin, Heilongjiang 150010, China (e-mail: 344053532@qq.com).

Guoliang Feng is a professor at the School of Automation Engineering, Northeast Electric Power University, Jilin, Jilin 132012, China (corresponding to e-mail: fengguoliang@neepu.edu.cn).

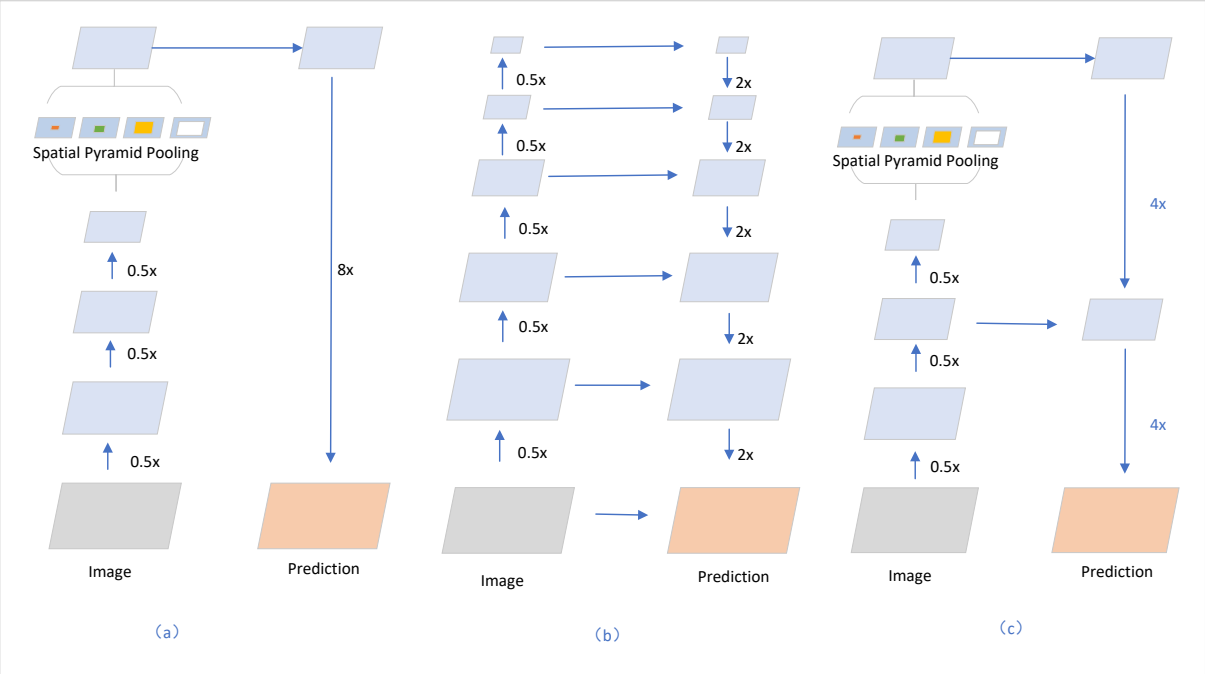

Fig. 1. Deeplabv3+structure diagram

This study proposes a novel foreign object detection framework that integrates the advanced feature extraction capabilities of large-scale language models (LLMs) with the high-precision segmentation power of DeepLabv3+. By combining deep semantic understanding with detailed image segmentation, this system aims to improve the efficiency, accuracy, and interpretability of power tower foreign object detection. By leveraging the strengths of large-scale models in extracting complex scene features and deep semantic information, along with the accuracy and multi-scale informationfusion capabilities of DeepLabv3+[17] in semantic segmentation, the system can effectively detect and classify foreign objects on and around power towers. Through the training and optimization of neural network models, the system not only automates the detection of foreign objects but also accurately identiffes their types, shapes, and locations. This approach is designed to address the challenges of foreign object detection in complex environments, providing support for routine maintenance of power systems and rapid response to abnormal situations. The system's[18][19][20] efffciency and precision signiffcantly enhance the operational performance of power systems, reduce the risk of failures caused by foreign object interference, and lower the cost and difffculty of manual inspections, thereby providing critical technical support for the intelligent development of the power industry.

II. METHODS FOR POWER TOWER FOREIGN OBJECT DETECTIO

A. *Deeplabv3+ Algorithm*

DeepLabv3+ is a semantic segmentation model that combines deep convolutional networks with dilated convolutions to capture multi-scale context information, making it highly efficient for small object detection and complex scenarios like foreign object detection on power lines. As the latest version in the DeepLab series, it combines the strengths of deep convolutional networks with dilated convolutions. The model introduces an Atrous Spatial Pyramid Pooling (ASPP) module and a decoder module, which together enable the model to capture multi-scale context information, improving performance and effectively mitigating the issue of inconsistent object scales in segmentation tasks.

Figure 1 Explanation:

(a) This illustrates one of the core modules of DeepLabv3+, known as the Spatial Pyramid Pooling (SPP) module. The SPP module extracts multi-scale information from images by applying parallel convolutional kernels of varying sizes (with different receptive fields), allowing the model to capture global semantic information. As shown in the figure, multiple convolutional kernels are applied in parallel to the input feature map. After processing, the feature maps at different scales are combined to form a more comprehensive feature map that incorporates global information. The final feature map undergoes 8x downsampling to generate the prediction results. This structure excels at capturing multi-scale information in images, though it may lack precision in segmenting fine object boundaries.

(b) The encoder-decoder structure is a key enhancement in DeepLabv3+, significantly improving segmentation performance. The encoder progressively downsamples the input image to extract rich semantic information, while the decoder performs stepwise upsampling, gradually restoring the image to a higher resolution. This iterative upsampling process in the decoder allows the model to capture finer object details, particularly at the boundaries. The encoder-decoder structure effectively preserves semantic information while enhancing object boundary precision, resulting in more refined segmentation outcomes.

(c) This depicts the modified version of DeepLabv3+ with Atrous Convolution. This structure employs atrous (or dilated) convolution for feature extraction, expanding the receptive field without increasing computational complexity.

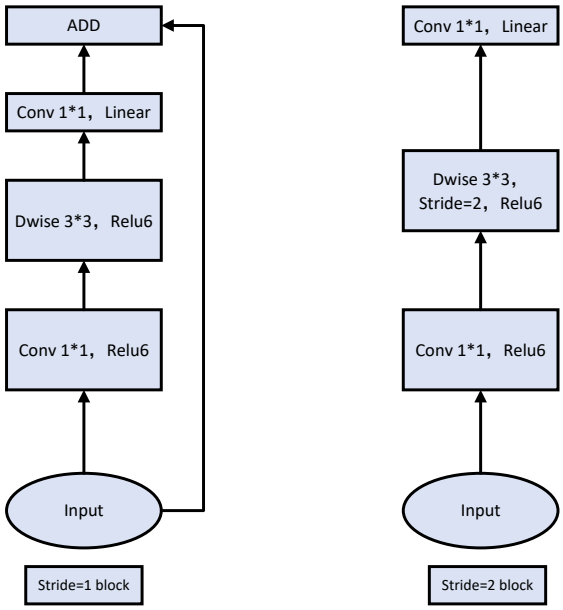

Fig. 2. MobileNet-V2 architecture diagram

Atrous convolution enables the model to extract features over a larger area while maintaining high resolution. This version also integrates the Spatial Pyramid Pooling (SPP) module, further enhancing the model's ability to extract multi-scale information. The segmentation prediction is produced through 4x downsampling, striking a balance between accurate boundary detection and the extraction of global semantic information.

DeepLabv3+ utilizes an encoder-decoder architecture, where the encoder's deep convolutional neural network (DCNN) part can be tailored to meet specific task requirements. For this study, the focus is on ensuring both efficiency and accuracy in system performance. Therefore, I have implemented MobileNet as the encoder backbone due to its lightweight and efficient design. The architecture and adjustments are optimized to meet the specific demands of the current task.

MobileNet is a highly efficient image classification model developed by Sandler et al., designed to minimize computational cost. MobileNet comes in three versions: MobileNet-V1, MobileNet-V2, and MobileNet-V3.

In this study, MobileNet-V2 is selected as one of the pre-trained models for transfer learning. MobileNet-V2 is the smallest model trained on the ImageNet dataset and has the second-fastest GPU processing time, surpassed only by MobileNet-V1. MobileNet-V2 consists of two stride-1 residual blocks and a second block with a stride of 2. Each block contains three layers: a 1×1 pointwise convolution layer, a depthwise convolution layer, and a 1×1 linear convolution layer, all activated by ReLU6. The architecture of MobileNet-V2 is shown in Figure 2.

DeeplabV3+ is a convolutional neural network architecture developed by the Google research team for image semantic segmentation. Its core feature is the use of the Atrous Spatial Pyramid Pooling (ASPP) module, which captures multi-scale contextual information to improve the recognition of object boundaries and small objects. Additionally, DeeplabV3+ incorporates a decoder structure to further enhance the accuracy of segmentation results. The overall architecture of the DeeplabV3+ network is illustrated in Figure 3.

The structure of DeeplabV3+ follows an encoder-decoder architecture. In the encoder, the input image undergoes downsampling through the backbone network, producing high-level semantic feature maps. These feature maps are then passed to the ASPP module, which consists of three atrous convolutions with rates of 6, 12, and 18, a 1×1 convolution, and a global average pooling layer. The resulting five feature maps are concatenated along the channel dimension, completing the multi-scale sampling process. A subsequent 1×1 convolution is applied to reduce the number of channels.

In the decoder, the low-level semantic feature maps obtained from 4x downsampling in the backbone network are first processed by a 1×1 convolution to reduce the number of channels. These feature maps are then concatenated with the upsampled feature maps from the encoder, which merges low-level and high-level semantic information. This fusion enhances the network's ability to segment objects accurately. The fused features undergo further processing with a 3×3 convolution, followed by another 4x upsampling to produce the final predicted segmentation map.

B. Model Compression

In this study, the Taylor pruning method is used as the core model compression technique to optimize the performance of the deep learning model. Taylor pruning leverages Taylor series expansion to estimate the impact of each channel or weight in a neural network on the final output. Specifically, it assumes that the network's output f(x) can be expressed as a function of the network weights W, and under the condition of maintaining performance, it removes weights or channels that have minimal impact on the output.

The implementation of Taylor pruning follows a systematic process. First, for each convolutional kernel or channel, we calculate the gradient of the network output f(x) with respect to that weight W using backpropagation. This gradient represents the degree of influence that weight has on the output. We use the first-order approximation of the Taylor series expansion to estimate the contribution of each channel or weight to the loss function. The estimation formula is:

$$\Delta L \approx \left| \frac{\partial L}{\partial W} \cdot W \right| \tag{1}$$

where $\frac{\partial L}{\partial W}$ represents the gradient of the loss with respect to the weight, and W is the current weight value. This calculation yields an approximation of how much the loss would change if the weight were removed.

Based on these estimated values, we calculate an importance score for each channel or convolutional kernel. We then sort these scores in ascending order and select channels or kernels with lower scores for pruning, as their impact on the model's output is minimal. In our implementation, we examined different pruning ratios (15%, 25%, and 40%) and determined that a 25% pruning ratio offered the optimal balance between model size reduction and performance preservation.

After pruning, the network usually experiences a temporary decrease in accuracy. We address this issue by fine-

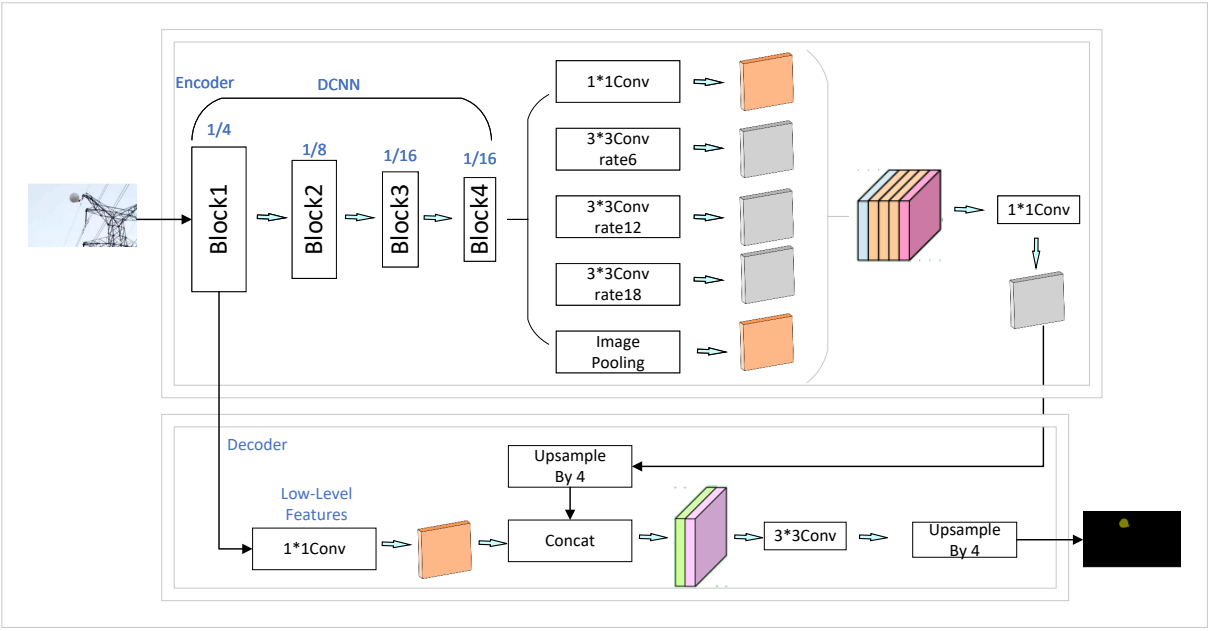

Fig. 3. DeepLabV3+Network Architecture Diagram

tuning the process to restore and optimize its performance on specific task datasets.

Taylor pruning offers several advantages over traditional sparsity-based or weight magnitude-based methods:

1) It directly leverages gradient information to assess the importance of channels or weights without requiring complex pre-training or evaluation steps.

2) It can be applied to different layers and types of weights, offering good flexibility.

3) It provides a theoretical basis for the pruning process, making it more systematic and well-founded.

In our experimental evaluation, we found that Taylor pruning not only significantly reduced the number of models, but also maintained comparable detection performance after pruning, while significantly reducing computational requirements, making it very suitable for deployment in resource constrained power system monitoring environments.

C. Power Tower Foreign Object Detection System

In this study, we propose a power line foreign object detection system based on multi-model fusion to efficiently detect and classify potential hazards on power towers. The system primarily consists of two core components: a large language model and the DeepLabv3+ semantic segmentation model. These two components work collaboratively through innovative feature complementarity and information fusion mechanisms, overcoming the limitations of single-model approaches.

The system workflow is as follows: First, the system leverages the powerful text generation capabilities of the large language model to perform an initial analysis and description of the input power tower image. The large language model can identify important features in the image amidst complex backgrounds and generate detailed descriptions. This process helps in identifying potential hazards and marking them within the image. The analysis provided by the large language model offers contextual information, which aids subsequent processing, particularly in identifying potential locations or abnormal areas where foreign objects may be present. This description forms the foundation for the following image processing steps.

Simultaneously, the image is passed to the DeepLabv3+ semantic segmentation model. DeepLabv3+ is a robust semantic segmentation network that excels in complex scenarios, performing detailed segmentation of input images and accurately locating foreign objects on power towers. Through semantic segmentation, DeepLabv3+ divides the image into distinct regions, categorizing them as tower structures, background, and various foreign objects (e.g., bird nests, kites, plastic bags). This process allows the system to precisely determine the shape, location, and size of foreign objects, providing essential data for subsequent risk assessment and mitigation efforts.

In the overall system, the large language model and the DeepLabv3+ semantic segmentation model complement each other: the descriptions generated by the large language model serve as a reference for identifying potential foreign objects, while DeepLabv3+ performs refined detection and segmentation. This dual-model fusion approach enhances the system's accuracy in detecting foreign objects on power towers and significantly reduces false-positive rates. In practice, this approach enables the system to maintain both high efficiency and accuracy when dealing with complex backgrounds and diverse scenarios.

III. EXPERIMENTAL STUDIES

A. Evaluation Metrics

We comprehensively evaluated the performance of our proposed foreign object detection system using four widely-used metrics: Mean Intersection over Union (mIoU), Precision, Mean Pixel Accuracy (MPA), and Recall. These metrics were carefully selected to provide a thorough assessment of the model's performance in detecting various types of foreign objects across different scenarios and environmental conditions. Our ablation experiments compared the standalone

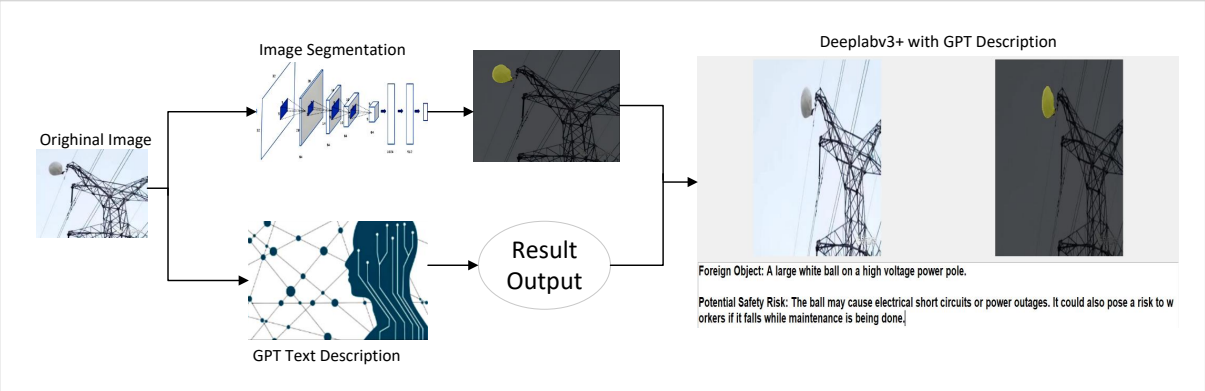

Fig. 4. Structure diagram of tower foreign object recognition system

DeepLabv3+ model with our integrated approach combining DeepLabv3+ and large language models. Additionally, we benchmarked our method against traditional approaches such as manual inspection and basic image processing methods, demonstrating superior performance in terms of both precision and recall. Below are the evaluation metrics along with their principles and formulas:

1) Mean Intersection over Union (mIoU)

mIoU is an important metric for measuring the performance of semantic segmentation models, representing the prediction accuracy for each class. It is calculated by comparing the overlap between the predicted and ground truth regions with their union.

For each class i , the formula for mIoU is:

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i} \tag{2}$$

Where: TP_i (True Positive): The number of pixels correctly predicted as class i, FP_i (False Positive): The number of pixels incorrectly predicted as class i, FN_i (False Negative): The number of pixels that actually belong to class i but were not predicted as such.

The mIoU is calculated as the average IoU value across all classes:

$$mIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i \tag{3}$$

Where N is the total number of classes.

2) Precision

Precision is a metric used to measure the accuracy of the model's predictions, representing the proportion of true positive samples among those predicted as positive. A higher precision indicates that a larger proportion of the model's positive predictions are correct.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

3) Mean Pixel Accuracy (MPA)

Mean Pixel Accuracy (MPA) is an important evaluation metric in semantic segmentation tasks, measuring the accuracy of the model in classifying each pixel. The calculation of MPA is based on the confusion matrix, which is an N times N matrix where N is the number of classes. Each element ij in the confusion matrix represents the number of pixels that actually belong to class i but are predicted as class j.

Table 1 Dataset Division

Data set	Nest	Balloon	Kite	Rubbish
Training set	90	90	90	90
Test set	10	10	10	10
Summary	100	100	100	100

$$MPA = \frac{1}{N}\sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i} \tag{5}$$

MPA is the average pixel accuracy across all classes, reflecting the model's average classification performance over all categories. The higher the MPA, the better the overall classification performance of the model. In practical applications, MPA is an important metric, especially in cases where class distribution is imbalanced, as it can provide a more comprehensive performance evaluation compared to Overall Accuracy.

4) Recall

Recall measures the model's ability to identify positive samples and represents the proportion of actual positive samples that are correctly recognized. The higher the recall, the greater the proportion of positive samples the model can correctly identify.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

5) Confusion Matrix

A confusion matrix is a table used to visualize the classification results, showing the relationship between true labels and predicted labels. Each element represents the number of samples predicted as a specific class, helping to analyze the model's performance on different categories and identify areas that need improvement. The rows of the confusion matrix represent the true labels, the columns represent the predicted labels, and the element values indicate the number of corresponding samples.

B. Experimental Data Preparation

The dataset for this study was carefully curated by collecting and processing high-quality images depicting power transmission towers and lines with various foreign objects. The collection process involved both internet sourcing and field photography to ensure diversity in environmental conditions, lighting situations, and viewing angles.

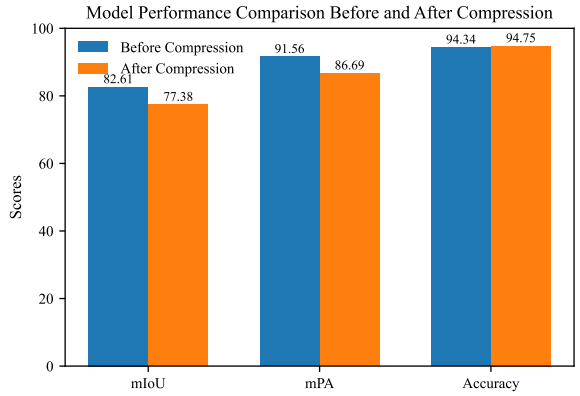

Fig. 5. Comparison of average metrics for compression models

This comprehensive dataset includes four primary types of foreign objects commonly found on power towers: bird nests, balloons, kites, and rubbish (such as plastic bags). Each category contains 100 images, creating a balanced dataset totaling 400 images. To ensure robust model training and evaluation, we implemented a stratified sampling approach, dividing the dataset into training and test sets at a 9:1 ratio while maintaining class distribution. The training set was further augmented using techniques such as rotation, scaling, and color adjustments to enhance model generalization capability.

All images are standardized to a resolution of 512 × 512, and the annotation process follows strict quality control protocols to ensure accuracy and consistency. Table 1 shows the specific data distribution of the training and testing sets.

C. Experimental Hardware Environment and Parameter Settings

The hardware environment configuration for this experiment includes a 12th generation Intel Core i5-12400F processor, an NVIDIA GeForce GTX 1660 SUPER graphics card, 16GB of RAM, and the Windows 10 Professional operating system. The input image resolution is standardized to 512×512, and the training process uses an SGD optimizer with a maximum learning rate set to 7e-3 and a minimum learning rate set to 0.01 times the maximum.

D. Average Performance of the Compressed Model

The Taylor pruning method was implemented for model compression in this study, with an optimal pruning ratio of 25% determined through extensive experimentation with different compression rates (10%, 25%, and 40%). After compression with the 25% pruning ratio, the average mIoU and mPA metrics decreased slightly from 82.61 to 77.38 and from 91.56 to 86.69, respectively. However, the overall accuracy notably improved from 94.34 to 94.75.

In addition, the compressed model has achieved significant improvements in computational efficiency, with inference time reduced by 3.7 times and memory usage reduced by 2.8 times compared to the uncompressed model. These results indicate that applying compression techniques can simultaneously reduce model complexity and improve system performance, making the solution more efficient and practical, especially when deployed on resource constrained devices in real-world power system monitoring applications.

E. Segmentation Performance of the Foreign Object Detection System

Figure 6 illustrates the comprehensive evaluation results of our system's segmentation performance across different metrics and object categories. The heatmap visualization clearly demonstrates that the system exhibits strong performance stability across various evaluation criteria, with particularly excellent results in certain categories.

When analyzing performance by category, we observe that the system achieves exceptional results in the Background and Balloon categories across all metrics (mIoU, Precision, Pixel Accuracy, and Recall). For the Background category, the system achieves near-perfect scores: 0.95 for mIoU, 0.97 for Precision, 0.98 for Pixel Accuracy, and 0.98 for Recall. Similarly, the Balloon category shows impressive performance with scores of 0.93, 0.95, 0.98, and 0.98 respectively. These results confirm the system's superior capability in segmenting visually distinct categories with clear shapes and well-defined boundaries.

For the Kite category, the system demonstrates good but slightly lower performance, with mIoU at 0.81, Precision at 0.91, Pixel Accuracy at 0.87, and Recall at 0.87. This moderate decrease can be attributed to the more complex shapes and occasionally transparent nature of kites, which present greater challenges for precise segmentation.

However, the system shows relatively lower performance in the more complex Nest and Rubbish categories. The Nest category achieves an mIoU of 0.71, Precision of 0.88, Pixel Accuracy of 0.78, and Recall of 0.78. The Rubbish category performs the lowest, with an mIoU of 0.63, Precision of 0.72, Pixel Accuracy of 0.83, and Recall of 0.83. This performance gap between simple and complex categories can be attributed to several key factors:

1)Target complexity:

Categories like nests and rubbish often have irregular shapes and complex structures, making them more prone to confusion with the background or other categories, thereby increasing the segmentation difficulty.

2)Data imbalance:

In the training data, complex categories (e.g., nests and rubbish) may have relatively fewer samples, leading to suboptimal model performance for these categories.

3)Detail extraction ability:

Segmenting these complex categories requires higher detail extraction capabilities, and the current model may not fully capture the characteristic

Despite the performance decline in complex categories, the system still demonstrates a high level of segmentation precision and recall overall, indicating excellent performance in handling highly salient target categories. This highlights the system's strong application potential, particularly in real-world scenarios like detecting foreign objects on transmission towers, where it can quickly identify and segment prominent target objects (such as backgrounds and balloons), thereby improving the overall task's efficiency and reliability.

Future Optimization Directions:

1)Data augmentation and sample balancing:

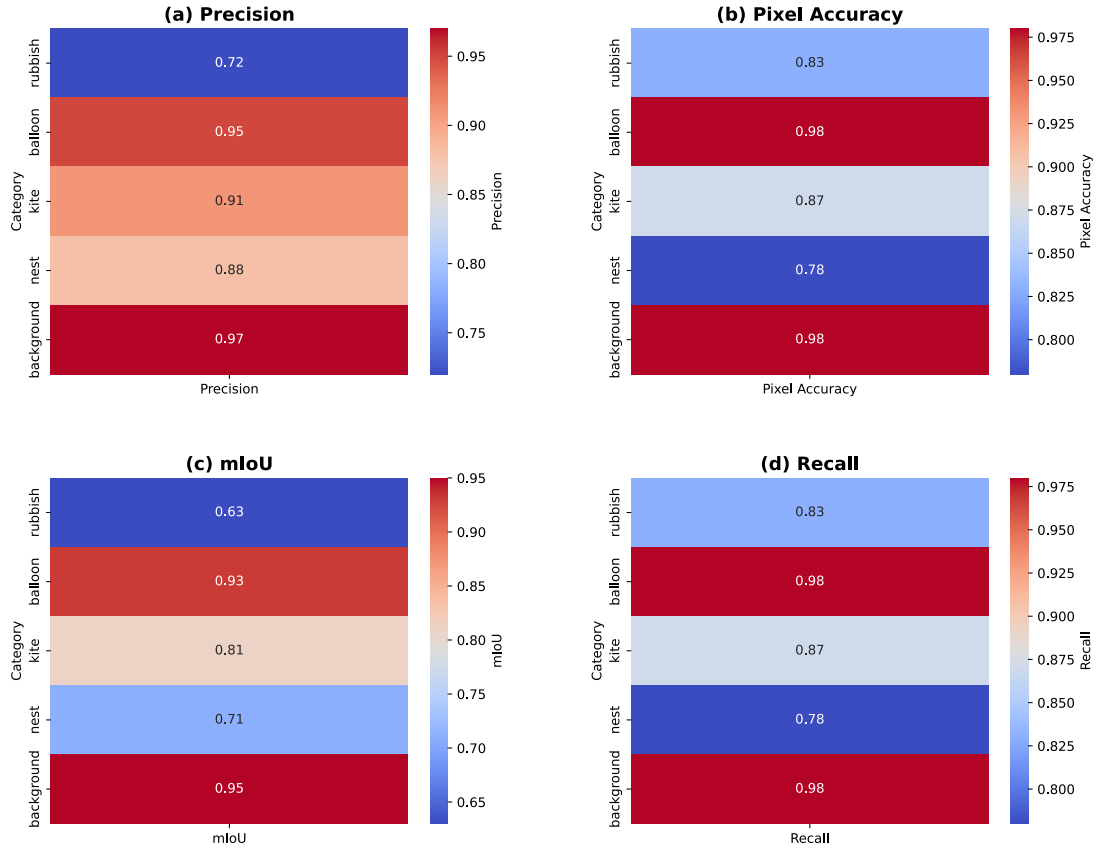

Fig. 6. Evaluation Indicators

Using data augmentation methods such as rotation, scaling, and color adjustments to generate more samples for complex categories, thereby enhancing the model's adaptability to these categories.

2)Improving model architecture:

Incorporating more efficient feature extraction modules, such as multi-scale fusion modules or attention mechanisms, to enhance the model's ability to capture details and complex targets.

3)Increasing training samples:

Adding high-quality labeled data for complex categories like nests and rubbish can significantly improve the model's segmentation performance.

4)Dynamic adjustment of loss weights:

Assigning dynamic loss weights to different categories during training to ensure that complex categories receive more attention.

Overall, this system has demonstrated outstanding segmentation capabilities, particularly for simple target categories. Its high precision and recall rates validate the system's superiority. However, for applications in more complex scenarios, further optimization and improvement are needed to ensure stable segmentation performance across diverse categories and environments, ultimately better serving the needs of practical application scenarios.

F. Analysis of Recognition Results

Figure 7 presents a comparative analysis of foreign object recognition results between two approaches: the standalone DeepLabv3+ model (second column) and our proposed integrated approach combining DeepLabv3+ with a large language model (third column). This side-by-side comparison across four representative foreign object categories (Balloon, Kite, Nest, and Rubbish) reveals substantial differences in recognition capability, result presentation, and practical utility.

When using the standalone DeepLabv3+ model, the system performs image segmentation and recognition independently. The model successfully detects foreign objects on transmission towers and visually marks them with color coding (e.g., highlighting bird nests in red). This approach effectively labels the foreign object category directly on the image and provides spatial information about its location and shape. For example, in the Nest category example, the model accurately identifies the position and approximate contours of the bird nest on the power tower.

However, this visual-only approach has significant limitations. The results are restricted to visual representations without contextual information or analysis of potential impacts. For maintenance personnel, this presents several practical challenges:

1) Limited operational guidance: The mere identification of an object without context about its potential risks makes prioritization difficult.

2) Absence of risk assessment : No information about potential hazards or urgency of intervention is provided.

3) Lack of actionable insights: Maintenance crews receive no guidance on appropriate response protocols.

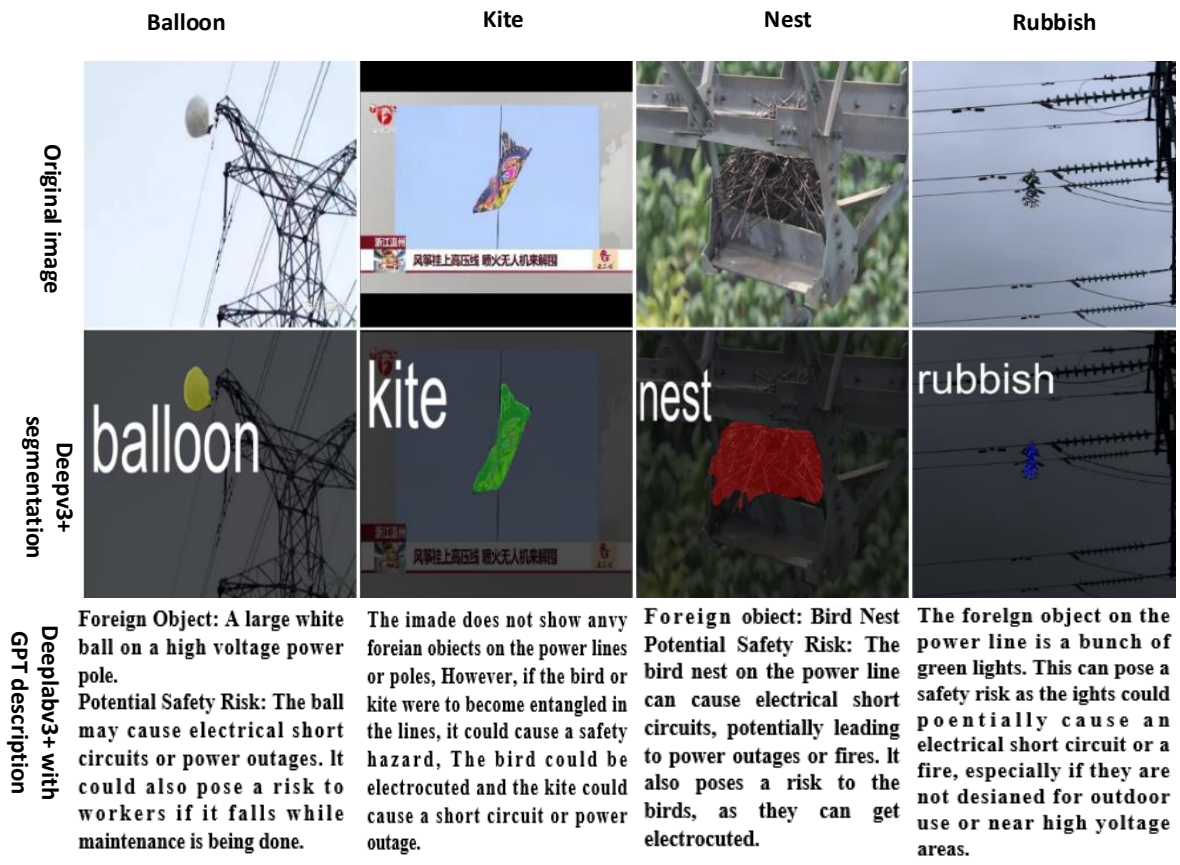

Fig. 7. Evaluation Indicators

From the above comparison, it is evident that the Deeplabv3+ model combined with GPT not only improves the system's ability to recognize target objects but also enhances the interpretability and practical value of the results. In scenarios such as transmission tower foreign object detection, this approach enables a more comprehensive analysis of the potential impacts and safety risks of foreign objects, providing clearer guidance for operational personnel and improving maintenance efficiency while reducing the risk of potential accidents.

In the future, this combined approach can be further optimized. For instance, introducing more efficient feature extraction modules to improve the baseline segmentation capability of Deeplabv3+ and enhancing GPT's domain knowledge to generate more accurate and in-depth textual descriptions can further improve the comprehensiveness and applicability of the system.

IV. Conclusion

This paper presents a novel foreign object detection system for power transmission towers that integrates DeepLabv3+ with large language models, demonstrating significant technological advancement and practical value. Our experimental results show that this integrated approach not only improves overall detection accuracy from 94.34% to 94.75% compared to the standalone DeepLabv3+ model but also generates intuitive textual descriptions that transform complex detection results into actionable operational information. The system effectively addresses three critical challenges in power system maintenance: accurate object detection, contextual understanding, and actionable insight generation.

The key contributions of this work include:

1) A novel multi-model fusion architecture that leverages the complementary strengths of semantic segmentation and large language models.

2) An effective model compression approach using Taylor pruning that maintains high accuracy while significantly reducing computational requirements.

3) Comprehensive evaluation across multiple metrics demonstrating superior performance particularly for complex object categories.

4) A practical implementation framework that bridges the gap between anomaly detection and operational decision-making.

The text generation capability powered by LLMs extends the system's utility beyond mere detection to support various operational needs. The detailed contextual descriptions enable the creation of standardized incident reports and maintenance logs, documenting specific anomalies with corresponding handling suggestions. This provides a reliable foundation for subsequent review, analysis, and event tracking, while enhancing the information management capabilities of power utilities.

In real-world deployment scenarios, our system significantly reduces the cognitive load on remote monitoring personnel by providing clear visual annotations accompanied by concise, interpretable textual information. This enables operators to quickly focus on critical issues within complex information streams, improving decision-making efficiency and accuracy. For example, when detecting a bird's nest,

operators can immediately access key risk information and efficiently schedule appropriate maintenance actions.

Future work will focus on further optimizing the model for complex real-world scenarios including adverse weather conditions and challenging lighting environments. We also plan to explore alternative architectures to improve segmentation performance for the more challenging object categories like nests and debris. Additionally, we aim to integrate this system with existing power grid monitoring platforms to create a comprehensive solution for intelligent power system maintenance.

REFERENCES

[1] Fen Qin, Hongfang Yao, Caiguo Ma, Junting Zhang, Tao Ni, and Jintong Xu, "Optimization Method for Manual Inspection of Transmission Lines Based on Dijkstra Algorithm," *Zhejiang Electric Power*, vol. 40, no. 6, pp. 49-53, 2021.

[2] Pengcheng Wang, "Design and Implementation of Helicopter and UAV Power Line Inspection System Based on Web," Tianjin University, 2020.

[3] Quan Xu, "Identification and Measurement of Power Towers Based on UAV Inspection," China University of Mining and Technology, 2023.

[4] Yongpeng Ling, "Design of Transmission Line Online Video Monitoring System," *Integrated Circuit Applications*, vol. 39, no. 8, pp. 132-133, 2022.

[5] Yixin Sun, "Preliminary Study on UAV and Manual Joint Inspection Mode for Overhead Transmission Lines," *Science & Technology Information*, vol. 16, no. 31, pp. 64-65, 2018.

[6] Cheng Jiang, Jiangang Zhang, Tingjian Li, Fu Zhang, and Feng Gao, "Research and Application of Data Processing in Helicopter Inspection System," *Microcomputer Applications*, vol. 36, no. 10, pp. 152-155, 2020.

[7] Huawei Meng, Gao Liu, Cong Wang, Sheng Guo, and Liyi Luo, "Automatic Inspection System for Transmission Lines Based on UAV Images," *Microcomputer Applications*, vol. 39, no. 8, pp. 193-196+200, 2023.

[8] Yu Jin, Sanlong Ma, and Jingkai Feng, "Online Monitoring Method for Icing on High-Voltage Transmission Lines Based on Machine Vision Technology," *Wireless Internet Technology*, vol. 20, no. 15, pp. 97-99, 2023.

[9] Zongqi M., "Transmission Line Inspection Image Recognition Technology Based on YOLOv2 Network," *Proceedings of the 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, IEEE, pp. 1-6, 2018.

[10] Gulzar M. A., Kumar K., and Javed M. A., "High-Voltage Transmission Line Inspection Robot," *Proceedings of the 2018 International Conference on Engineering and Emerging Technologies (ICEET)*, IEEE, pp. 1-5, 2018.

[11] Md. Faiyaz A., J. C. M., and Alok S., "Inspection and Identification of Transmission Line Insulator Breakdown Based on Deep Learning Using Aerial Images," *Electric Power Systems Research*, vol. 211, pp. 1-5, 2022.

[12] Dongmin Li, Jing Li, Dachuan Liang, et al., "Multi-Target Saliency Detection Method Based on Multi-Scale Prior Deep Features," *Acta Automatica Sinica*, vol. 45, no. 11, pp. 2058-2070, 2019.

[13] Lan Zhang, Lei Hu, and Lina Yu, "The Feasibility of Solving the Satisfiability Problem Using Various Machine Learning Approaches," Engineering Letters, vol. 32, no. 5, pp. 1004-1011, 2024.

[14] Wei Jin, Xiaoman Meng, and Yichao Wu, "Review of the Application of Deep Learning in Image Classification," *Modern Information Technology*, vol. 6, no. 16, pp. 29-31+35, 2022.

[15] Huijun Zou, Liangbao Jiao, Zhijian Zhang, et al., "Improved YOLO Network for Small Target Foreign Object Detection in Transmission Lines," *Journal of Nanjing Institute of Technology (Natural Science Edition)*, vol. 20, no. 3, pp. 7-14, 2022.

[16] Jianfeng Yang, Zhong Qin, Xiaolong Pang, et al., "Monitoring and Recognition Method for Foreign Object Intrusion in Transmission Lines Based on Deep Learning Network," *Power System Protection and Control*, vol. 49, no. 4, pp. 37-44, 2021.

[17] Wupan Li and Yuqi Liang, "Research on Improved Image Segmentation Method Based on DeepLabV3+ Model," *Modern Information Technology*, vol. 8, no. 19, pp. 39-43, 2024.

[18] Ruixuan Leng, "Application of Transmission Line Foreign Object Detection Algorithm Based on YOLOv8," Northeast Agricultural University, 2023.

[19] Hukai Li and Jie Wu, "LSOD-YOLOv8s: A Lightweight Small Object Detection Model Based on YOLOv8 for UAV Aerial Images," Engineering Letters, vol. 32, no. 11, pp. 2073-2082, 2024.

[20] ZhangFang Hu, FangYu Li, and JiXiang Shen, "A Semantic SLAM Integrated with Enhanced YOLOv7 Target Detection Algorithm," Engineering Letters, vol. 32, no. 10, pp. 1909-1920, 2024.

[21] Zhenbing Zhao and Yaping Cui, "Research Progress on Visual Detection Methods for Key Components of Transmission Lines Based on Deep Learning," *Electric Power Science and Engineering*, vol. 34, no. 3, pp. 1-6, 2018.