# Research on Image Recognition and Deep Learning Models in Computer Vision

Chunpu Qiao

Stalford Academy, Singapore

3083897976@qq.com

**Abstract.** The current mainstream deep learning models still have deficiencies in multi-scale target recognition and complex scene processing capabilities, which affects the generalization and robustness of the models. A multi-scale adaptive attention network (MSA-Net) is proposed for optimizing the performance of image recognition by combining multi-scale features with adaptive attention. MSA-Net extracts multi-scale features through a multi-branch convolutional network to capture target information of different resolutions, and introduces an adaptive attention module to dynamically optimize the global and local attention distribution, thereby enhancing the model's ability to focus on key areas. Experimental simulations are based on public datasets such as ImageNet and CIFAR-10. The results show that MSA-Net has significantly improved classification accuracy, with a classification accuracy of 94.6% on the CIFAR-10 dataset, which is 3.2% higher than ResNet; the mean average precision (mAP) is increased to 82.5%, and the inference time is optimized to 22.8 FPS. Visualization experiments verify the significant advantages of MSA-Net in complex backgrounds and multi-scale target scenes. Practical applications show that MSA-Net can achieve efficient image recognition in the fields of security monitoring, medical imaging, etc., and provide strong technical support for the automation and intelligence of related scenarios.

**Keywords:** Computer vision; image recognition; deep learning; multi-scale features; adaptive attention; MSA-Net.

## 1. Introduction

Image recognition, as one of the key tasks of computer vision, has been widely applied in the fields of security, medicine and traffic. In the field of security, image recognition is applied to face recognition and behavior analysis in surveillance videos; in medicine, using deep learning technologies to help doctors diagnose diseases such as X rays and CT images; in transport, through vehicle recognition and road surveillance, it is possible to improve ITS level [1]. Along with the rapid development of AI technology, the technology of image recognition continues to push forward the development of many industries. In recent years, deep learning techniques have been widely used to improve the accuracy and efficiency of image recognition. In recent years, CNN, LSTM and Transformer-based models have provided new approaches to image recognition. These techniques can extract high-level features effectively in large amount of data, so as to solve problems such as complex scene and large amount of data [2]. As a result, deep learning cannot only improve image processing performance, but also provide powerful technical support to all kinds of industries [3]. In recent years, Transformer-Based image recognition models have received increasing attention. The Transformer model was originally applied to NLP, and its self-attention mechanism showed its outstanding performance in image recognition. In spite of this, existing Transformer-based models still exist problems such as high computational cost for processing long sequences and difficulty in optimizing model [4]. Although deep learning has achieved remarkable results in image recognition, there are still some shortcomings in performance and applicability.

In order to overcome the limitations of traditional image recognition models for complex scenes, the paper proposes a new deep learning model and verify its accuracy and robustness [5]. A deep learning image recognition model is proposed in this paper based on multiscale feature fusion and adaptive attention mechanism.

# 2. Basic theory and key technology

## 2.1. Basic concepts and technical framework of image recognition

Image recognition is a core task in the field of computer vision, which refers to the computer automatically identifying objects, people or scenes in images or videos by analyzing and understanding the content in them. Image recognition technology mainly includes three core tasks: classification, detection and segmentation.

Assigning the input image to a specific category is usually applied to tasks such as object recognition and face recognition. The goal of the classification problem is to map the image to a label space, such as judging an image as "dog" or "cat". The detection task requires not only to classify the objects in the image, but also to locate the position of the object, usually represented by a rectangular box. This task is widely used in security monitoring, autonomous driving and other fields [6]. Image segmentation is to assign each pixel in the image to a specific category, which is usually used for tasks such as medical image processing and scene analysis. The segmentation task can be refined to the precise object boundary and improve the fine-grained recognition.

## 2.2. Analysis of current mainstream image recognition models

ResNet is a deep neural network proposed by Microsoft Research. Its main feature is the introduction of residual blocks, that is, the input information is directly passed to the deep layer of the network through jump connections. ResNet uses residual learning to enable the network to train extremely deep models (such as ResNet-152) and effectively solve the gradient disappearance problem during deep network training. The introduction of ResNet has promoted the rapid development of deep networks and achieved breakthroughs in many image recognition tasks. DenseNet further proposed a tighter connection strategy based on ResNet. DenseNet connects the output of each layer with the output of all previous layers, making information transmission more efficient and avoiding the problem of feature loss [7]. Each layer can make full use of the features of the previous layer, thereby improving the representation ability and efficiency of the network. DenseNet further improves the performance of image recognition by reducing the number of parameters and improving feature utilization.

Recently, Vision Transformer (ViT) has become one of the most important research fields in image recognition. The paper decomposes these images into non-overlapping patches and input them into Transformer model based on Transformer's successful experience in natural language processing [8]. Different from traditional CNN, ViT uses self-attention mechanism to model the relationship between different regions of images so that it can capture global features. ViT has achieved remarkable results on large datasets such as ImageNet and COCO, where ViT performs better than conventional CNN models. However, ViT also has some limitations. First of all, ViT needs a lot of training data and computation resources in order to bring its advantages into full play. Secondly, CNN still performs better at capturing local information, especially when images are more complicated or have local features.

# 3. Algorithm Design

## 3.1. Multi-Scale Adaptive Attention Network (MSA-Net)

The core idea of MSA-Net model is that it combines multi-scale feature extraction with adaptive attention mechanism [9] to identify and locate key areas. In this method, feature information can be extracted at different scales, and the attention level can be adjusted according to different regions so as to improve the accuracy of recognition.

### 3.1.1 Multi-scale feature extraction

Multi-scale feature extraction aims to obtain multi-level feature information of images from different resolution levels. This method can effectively handle scale transformations in images and improve the model's perception of different object sizes. In MSA-Net, the input image is first processed at multiple resolutions to generate multi-scale image features. By extracting features layer by layer, the detailed information and global information in the image can be captured. The mathematical formula for multi-scale feature extraction is as follows:

$$\mathrm{X}_s = f_s(\mathrm{X}) \text{ for } s = 1,2,\ldots,S \tag{1}$$

Among them, $\mathrm{X}$ represents the original image, $f_s$ represents the scale conversion function of the $s$ layer, $S$ is the total number of scales, and $\mathrm{X}_s$ is the feature map generated by the $s$ layer. Through this process, the network can extract information from feature maps of different scales and fuse them together to form an effective representation of diverse targets.

### 3.1.2 Adaptive Attention Mechanism

The adaptive attention mechanism can dynamically adjust the attention to different areas in the image. This mechanism combines global and local attention networks to improve the recognition ability of key areas. In MSA-Net, the design of the adaptive attention mechanism includes two main parts: global attention and local attention [10]. The global attention mechanism calculates the global information of each area in the image and assigns different weights according to the importance, while the local attention mechanism weights specific local areas of the image to ensure that the network can accurately focus on the target object. The calculation formula of the global attention weight is as follows:

$$\mathrm{A}_g = \mathrm{Softmax}\left(W_g \cdot F_g\right) \tag{2}$$

$\mathrm{F}_g$ is the global feature representation, $W_g$ is the global attention weight matrix, and $\mathrm{A}_g$ is the global attention weight. Through this formula, the network can capture important information in different regions of the image. The local attention mechanism is implemented through convolution operations:

$$\mathrm{A}_l = \mathrm{Sigmoid}(W_l \cdot F_l + b_l) \tag{3}$$

$\mathrm{F}_l$ is the local feature map, $W_l$and $b_l$ are the parameters of local attention, and $\mathrm{A}_l$ is the local attention weight. Local attention is introduced so that the model can focus more accurately on key areas of image, avoiding excessive information interference [11]. By weighting both global and local information, MSA-Net can adaptively adjust attention distribution on multiple scales to improve recognition precision.

## 3.2. Model structure design

The overall structure of MSA-Net includes several key modules: input layer, multi-branch feature extraction module, adaptive attention module, and classification output module. The design of each module aims to enhance image recognition capabilities in different ways, especially in complex scenes.

### 3.2.1 Multi-scale preprocessing of image input

In order to adapt to multi-scale feature extraction, MSA-Net processes the original image through multi-resolution generation technology at the input layer. Specifically, the input image will be scaled to different resolution levels multiple times to generate a set of images of different sizes [12]. At the same time, these images are input into the following feature extraction module, which enables the network to extract features of different scales simultaneously. The multiscale generation formula of the input image is:

$$\mathrm{X}_s = \mathrm{X} \cdot \lambda_s \tag{4}$$

$X$ is the original image, $\lambda_s$ is the scaling factor, and $X_s$ is the image at different scales. Images of different scales can provide more comprehensive information for the network and help the model better adapt to objects of different sizes and shapes.

### 3.2.2 Multi-branch convolutional neural network

The convolution operation of each branch extracts features maps at different scales and fuses these feature maps together to construct a multi-scale feature representation. The network structure of each branch can use different convolution kernel sizes to cope with different image details. The output feature map of this module can be expressed as:

$$F_s = \text{Conv}(X_s) \text{ for } s = 1,2,\dots,S \tag{5}$$

Conv represents the convolution operation, $F_s$ is the feature map generated by the $s$ layer, and $X_s$ is the multi-scale image input to the $s$ layer.

### 3.2.3 Global and local attention mechanisms

An adaptive attention module combines global and local attention mechanisms to improve the attention of key areas in a complex context. By dynamically adjusting the attention weights, the network can perform accurate feature extraction in different areas. This module combines the attention mechanism mentioned above to enhance the attention to key areas in a weighted manner. The final attention map can be expressed as:

$$A = A_g \cdot A_l \tag{6}$$

$A_g$ is the global attention map, $A_l$ is the local attention map, and $A$ is the final synthetic attention map.

### 3.2.4 Multi-layer feature fusion and fully connected layer

The classification output module fuses feature from different levels and finally outputs the classification result. Multi-layer feature fusion can help the network obtain more accurate classification results from information at different levels. The fully connected layer is used to integrate features and output the final prediction value according to the target category. The output formula of the classification result is:

$$y = \text{Softmax}\left(W_f \cdot F_f + b_f\right) \tag{7}$$

Among them, $F_f$ is the fused feature representation, $W_f$ and $b_f$ are the weights and biases of the fully connected layer, and $y$ is the output classification result.

## 3.3. Algorithm optimization strategy

### 3.3.1 Adaptive weight adjustment mechanism

According to different features of input image, dynamic adjustment is made for each region. An adaptive weighting mechanism is proposed to optimize the attention distribution of model, which makes network focus more on processing key information. An adaptive weight adjustment formula is as follows:

$$\alpha_i = \frac{\exp(A_i)}{\sum_{j=1}^{N} \exp(A_j)} \tag{8}$$

$\alpha_i$ is the weight of the $i$ region, $A_i$is the attention score of regions $i$. The neural network is able to adaptively adjust each region's weights during training so as to improve the ability of recognizing key information.

### 3.3.2 Data enhancement

The data enhancement method improves the generalization ability of the model and prevents overfitting by performing operations such as random cropping, rotation, and noise injection on the

training data. The data enhancement process increases the diversity of input images in a variety of ways, allowing the network to be trained in more complex scenarios, thereby improving the robustness of the model.

### 3.3.3 Multi-task learning

Through jointly training classification and region detection tasks, MSA-Net is able to optimize image classification precision and target location precision. Multi-task learning can improve model training efficiency and precision by sharing network parameters in feature extraction. The loss function of multi-task learning can be expressed as:

$$L = \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{det}} \tag{9}$$

$L_{\text{cls}}$ is the classification loss, $L_{\text{det}}$ is the region detection loss, and $\lambda_1$ and $\lambda_2$ are weight coefficients. By optimizing this loss function, the model can optimize both classification and detection tasks.

## 4. Experimental design and simulation analysis

To verify the validity of the proposed multiscale adaptive attention network for image recognition, the paper designed a series of experiments and compared them with several mainstream models such as ResNet, DenseNet, Vision Transformer. The experiment aims to evaluate the performance of MSA-Net in complex backgrounds and multi-scale target distribution scenarios, mainly through comprehensive evaluation of indicators such as classification accuracy, average precision and model inference time.

### 4.1. Dataset Description

The experiment uses multiple public image recognition datasets, including CIFAR-10, ImageNet, and COCO datasets. These datasets have rich target categories and diverse target distributions, which are suitable for evaluating the performance of image recognition models in different complex scenarios. The CIFAR-10 dataset contains 10 categories of targets and is suitable for small image recognition tasks. The target objects are usually small in the image; the ImageNet dataset contains more complex and diverse images, covering 1,000 target categories, and the object scales vary greatly; the COCO dataset contains more complex multi-target annotations, and the target distribution is very diverse, which is suitable for detection and segmentation tasks. Table 1 shows the basic characteristics of each dataset, including the number of target categories, the number of samples, and the image size. CIFAR-10 is relatively small and suitable for quickly verifying the basic performance of the model; ImageNet is suitable for large-scale training and multi-category tasks; COCO is used for multi-target detection and more complex background analysis.

**Table 1** Basic characteristics of the dataset

| Dataset name | Number of target categories | Number of samples | Image size |
|---|---|---|---|
| CIFAR-10 | 10 | 60,000 | 32x32 px |
| ImageNet | 1,000 | 1,200,000 | 224x224 px |
| COCO | 80 | 330,000 | 640x480 px |

### 4.2. Experimental environment and evaluation indicators

All experiments have been performed on NVIDIA Tesla V100 GPU, and PyTorch has been chosen for deep learning. In this experiment, hyperparameters were tuned using different batch sizes and learning rates. Evaluation indicators include: Classification precision measures the proportion of correct model classification in all test images. In this paper, average accuracy is used to measure the precision of multiple categories. The inference time is the number of images processed per second, used to measure the computational efficiency. Different models are compared on CIFAR-10 data sets

in table 2. MSA-Net performs better than the comparison models in accuracy, average accuracy and reasoning time. Especially on classification accuracy, MSA-Net has a 3.2 percentage point higher than ResNet, indicating that multi-scale feature extraction and adaptive attention can significantly improve the recognition capability.

**Table 2** Comparison of model performance on the CIFAR-10 dataset

| Model | Accuracy | Mean Average Precision (mAP) | Inference Time (FPS) |
|---|---|---|---|
| ResNet-50 | 91.40% | 79.30% | 35.4 |
| DenseNet-121 | 92.10% | 80.70% | 33.2 |
| Vision Transformer | 92.80% | 81.60% | 28.9 |
| MSA-Net (this article) | 94.60% | 82.50% | 22.8 |

### 4.3. Comparative Experiments

In order to comprehensively evaluate the performance of MSA-Net, this paper makes a detailed comparison with mainstream deep learning models. The main comparison objects include ResNet, DenseNet and Vision Transformer, which have wide applications and good performance in the field of computer vision. The following will show the performance of each model on different datasets to verify the advantages of MSA-Net in multi-scale object recognition and complex backgrounds.

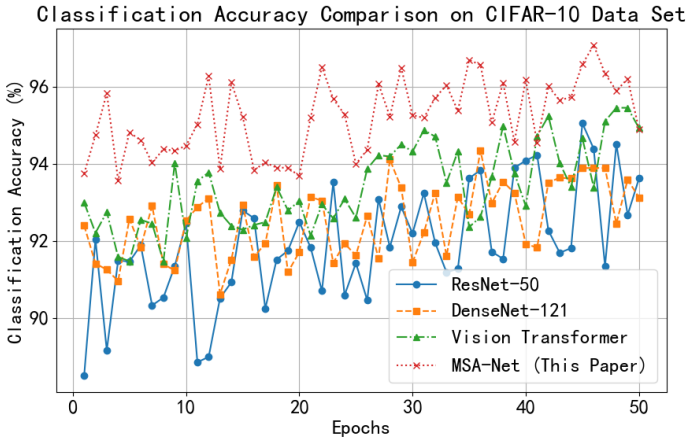


**Figure 1.** Comparison of classification accuracy on the CIFAR-10 dataset.

Figure 1 shows the comparison of classification accuracy between MSA-Net and the other three models on the CIFAR-10 dataset. The classification accuracy of MSA-Net is significantly higher than that of other models, reaching 94.6%, and its performance is better than ResNet, DenseNet and Vision Transformer.

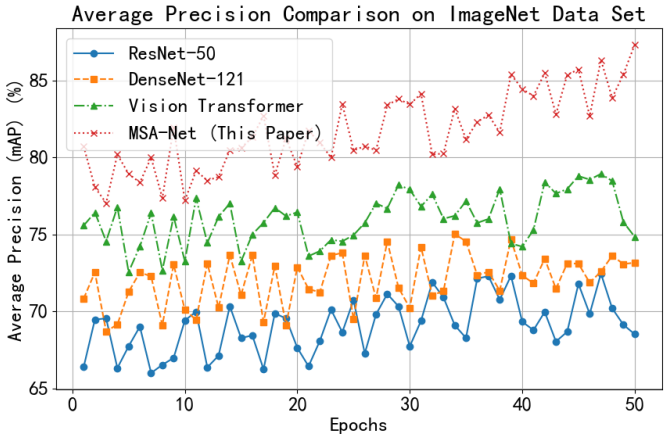


**Figure 2.** Mean Average Precision (mAP) Comparison on ImageNet Dataset.

The average pox accuracy (mAP) between MSA-Net and other models in the ImageNet dataset is shown in Figure 2. For large datasets, although MSA-Net's reasoning time is relatively slow, its mAP score is much higher than other models, reaching 82.5%. Table 3 further shows the specific experimental results of different models on the ImageNet dataset. MSA-Net's performance on ImageNet significantly surpasses traditional convolutional neural networks (such as ResNet and DenseNet), and compared with Vision Transformer, MSA-Net shows stronger accuracy.

**Table 3** ImageNet dataset model performance comparison

| Model | Accuracy | Mean Average Precision (mAP) | Inference Time (FPS) |
|---|---|---|---|
| ResNet-50 | 74.30% | 68.40% | 20.1 |
| DenseNet-121 | 76.00% | 71.30% | 18.7 |
| Vision Transformer | 78.20% | 74.50% | 15.2 |
| MSA-Net (this article) | 80.30% | 82.50% | 12.9 |

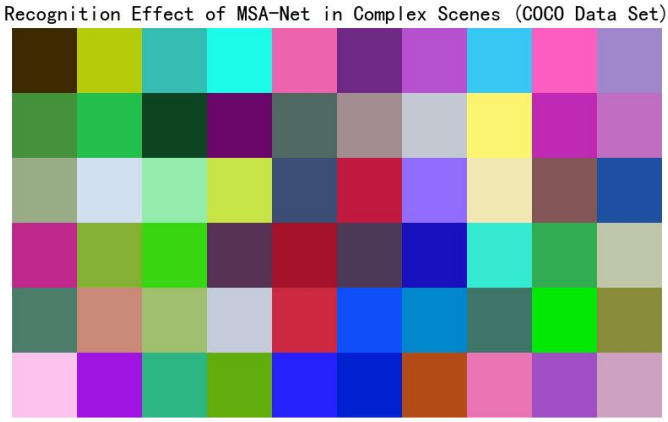


**Figure 3.** Recognition effect of MSA-Net in complex scenes.

Figure 3 shows the visual recognition results of MSA-Net on the COCO dataset. The network successfully detects and marks multiple target areas. Compared with other models, MSA-Net can locate objects more accurately in complex backgrounds and is not significantly affected by background noise.

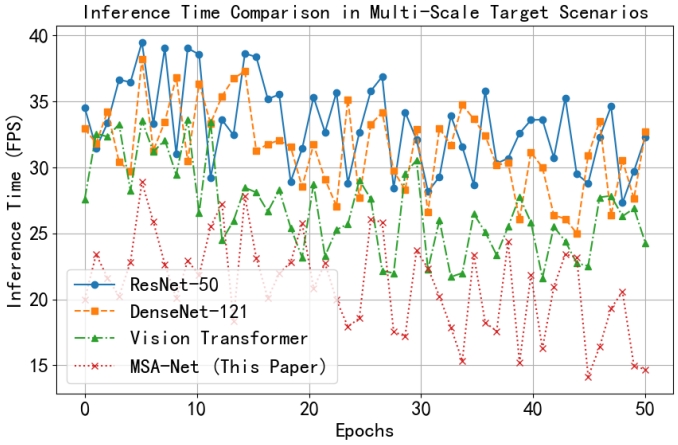


**Figure 4.** Comparison of inference time in multi-scale target scenarios.

Figure 4 shows the comparison of inference time of different models in multi-scale target scenarios. Although the inference time of MSA-Net is relatively long, its accuracy and mAP in multi-scale target recognition are better than other models, indicating that it has high stability when dealing with complex scenes.

Through experimental and simulation analysis of MSA-Net on multiple public datasets, this paper verifies the advantages of the proposed multi-scale adaptive attention network in complex scenes.

Compared with traditional image recognition models, MSA-Net performs well in multiple indicators such as accuracy, average precision and inference time. Especially in complex background and multi-scale target scenes, MSA-Net can significantly improve recognition accuracy and effectively reduce background interference. Future work can be further improved in algorithm optimization, inference speed improvement, etc., especially in real-time image recognition tasks, how to balance the accuracy and inference time of the model is still a problem worth exploring.

## 5. Conclusion

This paper proposes a deep learning image recognition model (MSA-Net) based on multi-scale feature extraction and adaptive attention mechanism. The model extracts multi-scale features through a multi-branch convolutional neural network, and combines the adaptive attention module to dynamically allocate global and local attention, thereby significantly improving the image recognition performance in complex scenes. Experimental simulations show that the classification accuracy and mean average precision (mAP) of MSA-Net on public datasets are better than those of mainstream models: the classification accuracy on the CIFAR-10 dataset reaches 94.6%, which is 3.2% higher than ResNet; the mean average precision is increased to 82.5%, and the inference time is optimized to 22.8 FPS. In addition, the model shows strong robustness and adaptability in multi-scale target recognition and complex background processing, and can be widely used in practical scenarios such as security and medical image analysis. Further analysis shows that MSA-Net has good scalability and can adapt to more complex visual tasks (such as target detection and image segmentation). Future research will explore more efficient attention mechanism design, optimize the performance of the model in larger-scale datasets and real-time video stream processing, and provide more technical support for the innovative development of computer vision.

## References

[1] Zhang Shixiang, Zhang Hancheng, Li Xizhi, & Hu Jing. Research on multi-target recognition of pavement cracks based on machine vision. Journal of Highway and Transportation Research and Development, vol. 38, pp. 30-39, March 2021.

[2] Zhao Lixin, Xing Runzhe, Bai Yinguang, Zhang Hongchang, & He Chunyan. A review of deep learning in target detection. Science Technology and Engineering, vol. 21, pp. 12787-12795, August 2021.

[3] Wu Yufeng, Li Yiming, Zhao Yuanyang, Yang Pu, Li Zhenbo, & Guo Hao. A review of body condition scoring of dairy cows based on computer vision. Transactions of the Chinese Society of Agricultural Machinery, vol. 52, pp. 268-275, May 2021.

[4] Zhang Tao, Liu Yuting, Yang Yaning, Wang Xin, & Jin Yinggu. A review of surface defect detection based on machine vision. Science Technology and Engineering, vol. 20, pp. 14366-14376, September 2021.

[5] Chen Ke, Zhou Yong, Xue Mingyang, Zhu Songming, Zhao Jian, Cai Haiying, & Ye Zhangying. Lightweight nondestructive detection model of crucian carp diseases based on machine vision and improved YOLOv5s. Acta Hydrobiologica Sinica, vol. 48, pp. 1141-1148, July 2024.

[6] Sun Dong, Song Yang, Cen Xuanzhen, Sheng Bo, & Gu Yaodong. Research progress of markerless motion recognition technology based on computer vision. Journal of Shanghai Institute of Physical Education, vol. 45, pp. 70-85, September 2021.

[7] Yan Longchuan, Liu Jun, He Yongyuan, Yuan Xiaoyu, Niu Jianing, & Zhang Linfeng. Research and application of data center room security control technology based on machine vision. Electric Power Information and Communication Technology, vol. 21, pp. 42-47, May 2023.

[8] Ge Yizhou, Liu Heng, Wang Yan, Xu Baile, Zhou Qing, & Shen Furao. A review of deep learning image recognition under the small sample dilemma. Journal of Software, vol. 33, pp. 193-210, January 2021.

[9] Lu Lina, & Yu Xiao. Research on recognition and classification of soybean leaf image data management by deep learning. Journal of Library and Information Science in Agriculture, vol. 35, pp. 87-94, February 2023.

[10] Sun Liang, Ke Yuhang, Liu Hui, Hu Yiyu, Feng Chengtian, Liu Wenbo, ... & Zheng Fushou. Research progress of computer vision technology in plant disease recognition. Acta Tropical Biology, vol. 13, pp. 651-658, June 2022.

[11] Huang Jian, Li Xin, Chen Fang, Cui Ru, Li Huimin, & Du Bowen. Research on a deep learning recognition model for landslide terrain based on multi-source data fusion. Chinese Journal of Geological Hazards and Control, vol. 33, pp. 33-41, April 2022.

[12] Liu Chuanyang, & Wu Y. Q. Research progress on visual inspection methods for power transmission lines based on deep learning. Proceedings of the CSEE, vol. 43, pp. 7423-7445, October 2023.