Efficient Channel Attention Optimized Multi-layer Fusion Edge Detection Network for Boundary Extraction

Tianyi Qin, Bingyu Qu*, Houlin Wang

China Communications Information & Technology Group CO.LTD., Beijing 101399, China

*qintianyi1@ccccltd.cn

Abstract. In modern Building Information Modeling (BIM), precise boundary extraction is crucial for terrain construction and model generation. Addressing the limitations of current edge detection methods in boundary extraction from remote sensing images, this paper proposes a novel edge detection model named Efficient Channel Attention Optimized Multi-layer Fusion Edge Detection Network (EMF-NET). To retain more feature information during the network's downsampling process and improve the accuracy of boundary extraction, we integrate the Efficient Channel Attention (ECA) mechanism with max-pooling layers, creating the ECA Poolblock. The ECA Poolblock enables the network to more accurately identify target boundaries and structures during edge detection tasks, enhancing the precision and robustness of boundary extraction. Additionally, EMF-NET adopts a multi-layer end-to-end network architecture based on the concept of multi-value fusion, significantly outperforming traditional single-layer encoder-decoder architecture edge detection networks. Experimental results demonstrate that the proposed edge detection network achieves an F1 score of 90.18% and an Intersection over Union (IOU) of 80.78% in remote sensing image boundary extraction tasks on GF-2 dataset, markedly superior to other state-of-the-art edge detection methods, showcasing excellent edge detection performance.

Keywords: Edge Detection, Efficient Channel Attention, Multi-layer Fusion, Boundary Extraction

1 Introduction

In modern Building Information Modeling (BIM), the accuracy of boundary extraction is crucial for precise terrain construction and model generation[1]. Traditional edge detection methods often struggle with the complex boundaries present in remote sensing images, resulting in suboptimal performance in BIM applications[2]. For instance, classic edge detection techniques like the Canny edge detector[3] and Sobel[4] operator are susceptible to noise interference when processing high-resolution remote sensing images, leading to inaccurate boundary extraction. Moreover, single-layer encoder-decoder architectures based on convolutional neural networks (CNNs)[5-7] have improved edge detection accuracy to some extent but still face limitations in extracting

Z. Ahmad et al. (eds.), *Proceedings of the 2024 5th International Conference on Urban Construction and Management Engineering (ICUCME 2024)*, Advances in Engineering Research 242,
https://doi.org/10.2991/978-94-6463-516-4_47

boundary details and complex structures. To address these issues, researchers have proposed various improvements. For example, the U-NET architecture, which integrates multi-scale features, performs well in edge detection tasks[8]. However, its single-layer structure restricts the depth and effectiveness of feature fusion. To further enhance the accuracy and robustness of boundary extraction, some studies have introduced attention mechanisms to improve detection performance by enhancing the expression of key features. Nevertheless, existing multi-layer fusion[9] and attention mechanism[10] methods still have room for improvement in terms of computational efficiency and feature retention.

To address the above problems, this paper proposes a novel edge detection model, EMF-NET. This method combines the ECA [11] mechanism with max-pooling layers to create the ECA Poolblock. The ECA Poolblock allows the network to more accurately identify target boundaries and structures in edge detection tasks, thereby enhancing the precision and robustness of boundary extraction. Our contributions are as follows:

1. ECA Poolblock. This paper introduces the ECA Poolblock, a novel module combining the Efficient Channel Attention (ECA) mechanism with max-pooling layers. This integration allows the network to retain more feature information during the downsampling process, resulting in more accurate identification of target boundaries and structures. Consequently, it enhances precision and robustness.
2. Multi-layer Fusion Architecture. This paper proposes a multi-layer fusion architecture based on the concept of multi-value fusion. This network architecture amalgamates multiple independent results to enhance the model's robustness and performance. It addresses the common issue where single-layer networks fail to effectively capture boundary features in remote sensing images due to complex textures and structures, thereby improving the level of detail in edge extraction results.
3. Experimental results on the the GF-2 and Denmark Marker satellite remote sensing datasets demonstrate that the proposed EMF-NET significantly outperforms commonly used edge detection methods across various metrics in boundary extraction tasks, exhibiting superior performance.

2 Approach

2.1 EMF-NET

Traditional edge detection methods often suffer from a loss of detail and critical semantic information during the downsampling process due to the reduction in feature map resolution, ultimately affecting the accuracy of edge detection results. Additionally, remote sensing images inherently possess complex ground cover and blurred boundaries, leading to common problems such as boundary ambiguity, misclassification, and omissions in the output. To address these issues, this paper proposes the EMF-NET, which integrates the novel ECA Poolblock and a multi-layer fusion architecture based on the concept of multi-value fusion. The ECA Poolblock combines the Efficient Channel Attention (ECA) mechanism with pooling layers, enabling the automatic learning of

channel and spatial information in feature maps and optimizing the representation capability of feature maps. This enhances the clarity, accuracy, and robustness of the results. By introducing the ECA Poolblock at each downsampling stage in the encoder layer, the network can focus more on important features while preserving more detail during downsampling.

The multi-layer fusion architecture employs the multi-value fusion concept, applying data augmentation techniques such as rotation and mirroring to input images to generate multiple varied inputs. These inputs are then processed by segmentation networks with identical structures. The multi-value fusion approach allows for the integration of multiple predicted images to obtain the final result. The multi-layer fusion architecture designed in this paper endows EMF-NET with the modularity to replace the backbone network module based on specific requirements. Common end-to-end edge detection networks, such as U-Net, Convolutional Encoder-Decoder (CED), Richer Convolutional Features (RCF), and Holistically-Nested Edge Detection (HED), can be integrated into EMF-NET. This flexibility allows the construction of models suitable for various tasks and datasets, enhancing EMF-NET's performance and robustness in handling different types of noise and variations in remote sensing images. Additionally, this provides EMF-NET with strong scalability and flexibility in fields such as medical image segmentation, natural scene segmentation, and remote sensing image segmentation. To better address the complexities and variations in ground cover in remote sensing images and enhance the model's robustness, we chose U-Net as the backbone network. The structure of the EMF-NET network with U-Net as the backbone is illustrated in Figure 1.

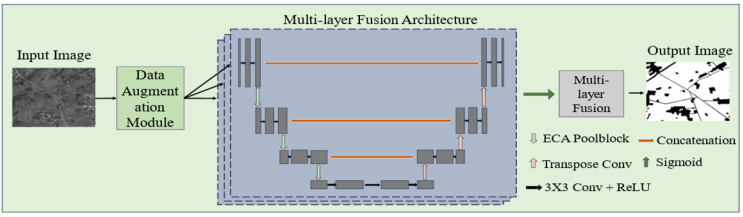

Fig. 1. EMF-NET structure

As illustrated in Figure 1, EMF-NET begins by taking remote sensing images as input into the data augmentation module. Through techniques such as mirroring and flipping, various different inputs are generated. These inputs are then fed into the multi-layer fusion architecture. In the U-NET network, optimized with the introduction of the ECA Poolblock to enhance the downsampling process, multiple independent results are fused based on the multi-value fusion concept. This process yields the final refined boundary results.

2.2 ECA PoolBlock

During the downsampling process in U-NET, the resolution of feature maps is reduced, leading to the loss of detail and important semantic information. This results in

decreased segmentation accuracy and blurred boundaries. The Efficient Channel Attention network proposed by Hang Zhang et al., is an attention mechanism designed to enhance neural network performance. Its core idea is to capture inter-channel dependencies using one-dimensional convolution, thereby improving the model's performance and generalization ability. Compared to traditional attention mechanisms, ECA avoids complex dimensionality reduction and expansion processes, preserving the integrity of the original channel features. This helps the model better learn and utilize the correlations between channels, achieving both efficiency and lightweight characteristics. The ECA mechanism adaptively computes the kernel size k of the one-dimensional convolution based on the number of channels. The formula for calculating the kernel size k is as follows:

$$k = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right| \quad (1)$$

This paper combines the Efficient Channel Attention (ECA) mechanism with max-pooling layers to construct the ECA Poolblock. The ECA Poolblock can adaptively compute and adjust the kernel size k of the one-dimensional convolution based on the number of channels in the input feature map. Consequently, it dynamically captures channel dependencies across different ranges. This powerful adaptive mechanism significantly enhances the model's ability to extract boundary feature information.

2.3 Multi-layer Fusion Architecture

In image processing tasks, due to the complexity and diversity of input images, single-layer models often struggle to capture all relevant information. Therefore, fusing multiple prediction results can improve the model's performance and robustness. The basic principle of the multi-value fusion concept is to integrate the results of multiple independent predictions to obtain a more accurate and reliable final prediction. Its advantage lies in its ability to leverage the differences between multiple models or prediction results, thereby overcoming the limitations of a single model or prediction. This enhances the accuracy of the final result and the generalization capability of the model, making it more suitable for different data distributions and scenarios. This paper proposes a multi-layer fusion architecture based on the concept of multi-value fusion, aiming to improve the model's ability to extract boundary feature information from remote sensing images. The workflow of this architecture is illustrated in Figure 2.

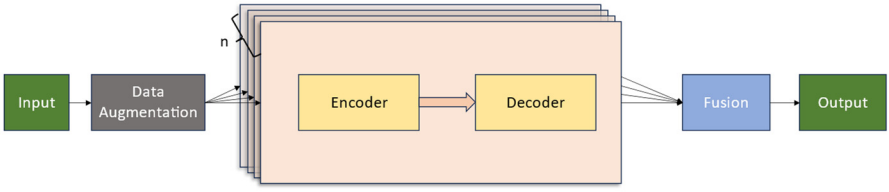

Fig. 2. Architecture of Multi-layer Fusion

For the input image, the convolution operation sequentially samples the image, and the weights of the feature map correspond to the pixels. Thus, inputting the same image in different orientations results in different sampling outcomes. In this paper, we apply data augmentation methods such as rotation and mirroring to the same input image, generating multiple different images based on the original input. These augmented images are then input into identical backbone networks, producing multiple distinct prediction results. Using the weighted fusion strategy from the multi-value fusion concept, we fuse these independent prediction results to obtain the final output of the multi-layer fusion architecture. The number of stacked layers can range from 1 to n. When the number of layers is 1, the proposed stacked architecture degenerates into the original single-layer architecture, demonstrating the generalization capability of the multi-layer stacked network architecture. In this paper, the original input image is subjected to 90°, 180°, and 270° rotations, followed by mirroring along the X-axis and Y-axis. Consequently, a total of 6 augmented images are generated after data augmentation. The fusion based on the multi-value fusion concept is described by equation (2):

$$P_{\text{final}} = \frac{w_1P_1+w_2P_2+w_3P_3+w_4P_4+w_5P_5+w_6P_6}{w_1+w_2+w_3+w_4+w_5+w_6} \tag{2}$$

In equation (2), Pi represents the predicted segmentation result image from the i-th segmentation network in the stacked network architecture, and w_i denotes the weight of this segmentation result image. By applying data augmentation techniques such as rotation and mirroring to the input image, we generate multiple varied images, which enhance the model's coverage of image information, improve its understanding of complex scenes, and reduce the occurrence of information loss. Using the multi-value fusion concept, the model integrates multiple predicted segmentation images to obtain the final segmentation result. This fusion method compensates for the limitations of individual prediction results, further enhancing the model's understanding of the image. It enables the model to better adapt to different scenes and data distributions, thereby improving the accuracy and stability of the final boundary recognition results.

The main feature of the multi-layer fusion architecture in this paper is the replaceability of its backbone network module. This feature allows for the selection of different backbone network modules according to the specific requirements of the task and data characteristics, thereby constructing edge detection models suitable for various scenarios and enhancing the model's performance and robustness. For example, when the target task requires the model to better handle issues arising from complex ground cover and variability, and the dataset size for the target task is relatively small, selecting U-Net as the backbone network module and integrating it into the network can fully leverage U-Net's characteristics for small dataset tasks. These characteristics include skip connections and the use of an encoder-decoder structure for feature extraction and restoration at different levels, thereby improving the accuracy and robustness of the results. Conversely, when the target task demands high accuracy in edge detection and is sensitive to details, U-Net can be replaced with the HED module. HED utilizes deep convolutional neural networks for multi-scale edge detection and improves detection accuracy by fusing the detection results at various scales.

3 Experiments

3.1 Data Set and Evaluation Metrics

To evaluate the proposed EMF-NET, experiments were conducted on the GF-2 and Denmark Marker satellite remote sensing datasets. The Denmark Marker dataset[12], released in 2016 by Denmark's European Union Land Parcel Identification System (LPIS), contains nearly 600,000 parcels, each with a unique identification number. The GF-2 dataset is from a multispectral Earth observation satellite developed by the China Academy of Space Technology (CAST), primarily tasked with providing high-resolution remote sensing image data for China's land resource surveys and urban planning. The datasets were divided into training, validation, and test sets in a 6:2:2 ratio.

To evaluate the performance of the proposed EMF-NET, we utilized two commonly used evaluation metrics: Intersection over Union (IOU) and F1 score. IOU is one of the most widely used metrics in edge detection tasks, quantifying the ratio of the intersection between the predicted instances and the ground truth instances. The calculation formula is shown in equation (3):

$$IOU = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i+FN_i} \tag{3}$$

In equation (3), TPi represents the number of pixels correctly predicted as class i and overlapping with the ground truth, FP_i represents the number of pixels incorrectly predicted as class i without overlapping with the ground truth, and FN_i represents the number of pixels that belong to class iii but were not correctly predicted. N is the total number of classes. The IOU value ranges from 0 to 1, with higher IOU values indicating better model performance.

The F1 score is a metric that considers both precision and recall. It is commonly applied in binary classification scenarios between the target and background in instance segmentation tasks. The formula for calculating the F1 score is shown in equation (4):

$$F1 = \frac{2\times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

In equation (4), *Precision* is the ratio of the number of pixels correctly predicted as the target to the total number of pixels predicted as the target. *Recall* is the ratio of the number of pixels correctly predicted as the target to the total number of actual target pixels.

3.2 Implementation Settings

All models were trained using the Adam optimizer with a learning rate set to 0.001 and a batch size of 8. The training was conducted on an NVIDIA GeForce GTX 4090 GPU with 24GB of memory, using PyTorch as the deep learning framework.

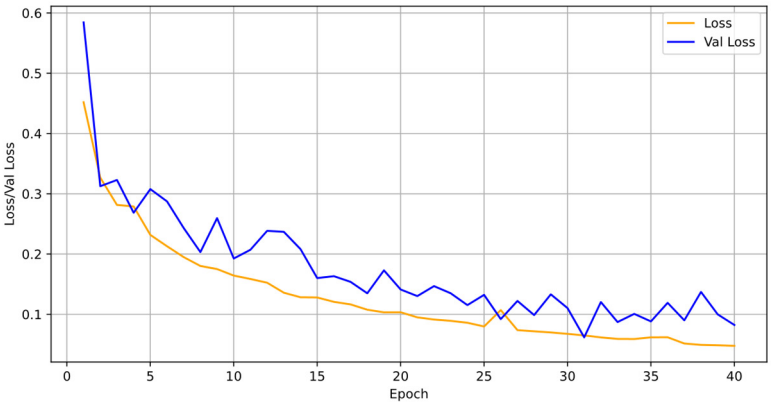

Fig. 3. Loss and Val Loss over Epochs

Figure 3 shows the relationship between the loss function of the model and the number of epochs in this study. It can be observed that as the number of training epochs increases, the model's loss gradually decreases. When the training epochs approach 40, the loss on the training set converges to approximately 0.05, and the loss on the validation set stabilizes. This indicates that the model is able to extract useful features from the data and learn effectively, demonstrating good convergence. Additionally, it shows that the model has relatively low computational resource requirements.

3.3 Performance Comparison

To investigate the segmentation performance of the proposed EMF-NET, we conducted a performance comparison experiment. The experiment compared the performance of EMF-NET with other classic models: U-NET, Canny, Sobel, ED-NSNP, and EDTER on the two remote sensing datasets. The experimental results are shown in Table 1.

Table 1. Comparison results of all methods on two datasets.

DataSet	GF-2		Denmark Marker	
Model	F1	IOU	F1	IOU
U-NET	81.71	72.64	78.31	69.34
Canny	83.69	71.68	77.5	65.82
Sobel	83.01	70.96	78.16	66.81
ED-NSNP[13]	86.42	76.42	81.61	71.16
EDTER[14]	88.94	79.14	82.01	73.35
EMF-NET	**90.18**	**80.78**	**83.91**	**74.9**

As observed from the experimental results in Table 1, the baseline model U-NET achieved F1 and IOU scores of 81.71% and 72.64%, respectively, on the GF-2 remote sensing dataset. Edge Detection Transformer (EDTER) is an edge detection model

based on the Transformer architecture. Compared to traditional convolutional neural network (CNN) models, EDTER leverages the self-attention mechanism to capture both global and local information, thereby improving the accuracy and robustness of edge detection. Compared to the U-NET baseline model, its F1 and IOU values increased by 7.23% and 6.5%. The EMF-NET designed in this paper consists of two core modules: the ECA Poolblock and a multi-layer fusion architecture. The ECA Poolblock uses one-dimensional convolution to capture inter-channel dependencies, avoiding complex dimensionality reduction and expansion processes, thereby preserving the integrity of the original channel features. This helps the model better learn and utilize the correlations between channels, achieving both efficiency and lightweight characteristics. The multi-layer fusion architecture improves the model's generalization ability by integrating multiple independent prediction results and leveraging the differences between them, thereby compensating for the limitations of single-layer structures and making the results more precise. EMF-NET improved the F1 and IOU scores by 8.47% and 8.14%, respectively, compared to the U-NET baseline model. Additionally, it showed significant improvements over the advanced edge detection model ED-NSNP on the GF-2 remote sensing dataset, demonstrating the excellent performance of EMF-NET in edge detection tasks. On the Denmark Marker dataset, EMF-NET achieved an F1 score of 83.91% and an IOU of 74.9%, significantly outperforming several other edge detection methods. This demonstrates the model's generalizability and excellent performance across different types of datasets. Figure 4 shows examples from the performance comparison experiments of the three models—U-Net, ED-NSNP, and CCIS-NET—on the GF-2 dataset.

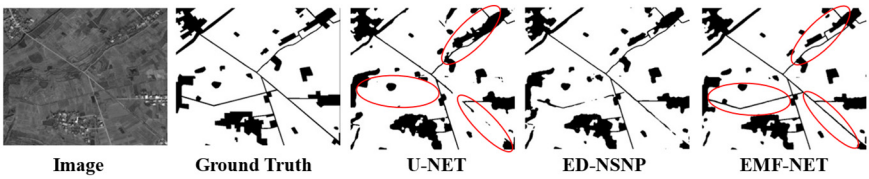

Fig. 4. Visualization results for case study

As shown in Figure 4, two images from the GF-2 remote sensing dataset were used to demonstrate the instance segmentation effects. It was observed that the results from the U-NET baseline model contained significant noise and misclassification/omission issues at the boundaries. The more advanced ED-NSNP model reduced the amount of noise and, to some extent, preserved the boundary parts that were ignored in the U-NET results. However, it still failed to completely resolve the noise and misclassification/omission problems. The rightmost images show the results of the proposed EMF-NET model. Through the multi-layer fusion architecture and the ECA Poolblock, which enhance the extraction and retention of information in the input feature maps, the number of noise points in the segmentation results is significantly reduced. The EMF-NET model effectively addresses the misclassification/omission issues that are challenging for other edge detection methods, preserving the boundary parts that were ignored in

both U-NET and ED-NSNP. This demonstrates the superior performance of the proposed EMF-NET model in edge detection tasks.

4 Model Analysis

4.1 Ablation Studies

To investigate the performance improvement effect of the ECA Poolblock, which is based on the ECA attention mechanism, ablation experiments were conducted using the U-NET network as the foundation. The experiments compared the performance of four segmentation models on the GF-2 remote sensing dataset: U-NET, U-NET combined with the SE attention mechanism[15] (U-NET_SE), U-NET combined with the STN attention mechanism[16](U-NET _STN), and U-NET combined with the proposed ECA Poolblock (U-NET _ECA). The experimental results are shown in Table 2.

Table 2. Ablation results of the attention mechanisms

Model	F1(%)	IOU(%)
U-NET	81.71	72.64
U-NET_SE	82.98	73.85
U-NET_STN	84.11	75.67
U-NET_ECA	85.12	76.17

Comparing the performance of the models combined with different attention mechanisms in Table 2, it is observed that both SE and STN improved the performance of U-NET to some extent. When combined with the proposed ECA Poolblock, the U-NET model's performance metrics improved by 3.41% and 3.53%, respectively, showing superior optimization effects compared to the SE and STN attention mechanisms. This demonstrates the effectiveness of the ECA-based pooling module constructed in this paper. To investigate the performance improvement effect of the multi-layer fusion architecture based on the multi-value fusion concept proposed in this paper, we compared the performance of two models on the GF-2 remote sensing dataset: U-NET_non using a single-layer network architecture and U-NET_mul using the multi-layer fusion architecture. The results of the ablation experiments are shown in Table 3.

Table 3. Ablation results of the multi-layer fusion

Model	F1(%)	IOU(%)
U-NET_non	81.71	72.64
U-NET_mul	86.17	75.86

As observed from the results in Table 3, the multi-layer stacked architecture designed in this paper significantly improves the model's generalization ability by integrating multiple independent prediction results. U-NET_mul can fully leverage the differences between various prediction results, compensating for the limitations of obtaining segmentation results using only a single-layer structure. This leads to more accurate segmentation results, with an improvement of 3.22% in IOU and 4.46% in F1 compared

to U-NET_non. These results demonstrate the effectiveness of the multi-layer fusion architecture in optimizing model performance.

5 Conclusion

This paper proposes an Efficient Channel Attention Optimized Multi-layer Fusion Edge Detection Network (EMF-NET). By introducing the ECA attention mechanism during the downsampling process, it addresses the issue of boundary information loss caused by changes in feature map resolution in U-NET. Additionally, to tackle the problem of boundary blurring due to the inability to fully capture fine edge features in high-resolution remote sensing images, a multi-layer fusion architecture based on the multi-value fusion concept was constructed. This significantly enhances U-NET 's performance in high-resolution remote sensing image edge extraction tasks. Finally, experimental results on the GF-2 dataset demonstrate that EMF-NET outperforms advanced algorithms in boundary extraction.

Reference

1. Li S, Li S, Hu J, et al.: Intelligent Modeling of Edge Components of Prefabricated Shear Wall Structures Based on BIM. Buildings 13(5), 1252-1269 (2023).
2. Khudhair A, Li H, Ren G, et al.: Towards future BIM technology innovations: A bibliometric analysis of the literature. Applied Sciences 11(3), 1232-1252 (2021).
3. Wang L, Gu X, Liu Z, et al.: Automatic detection of asphalt pavement thickness: A method combining GPR images and improved Canny algorithm. Measurement 196, 111248-111259 (2022).
4. AS R A, Gopalan S.: Comparative analysis of eight direction sobel edge detection algorithm for brain tumor MRI images. Procedia Computer Science 201, 487-494 (2022).
5. Elharrouss O, Hmamouche Y, Idrissi A K, et al. Refined edge detection with cascaded and high-resolution convolutional network. Pattern Recognition 138, 109361-109370 (2023).
6. Xian R, Xiong X, Peng H, et al.: Feature fusion method based on spiking neural convolutional network for edge detection. Pattern Recognition 147: 110112-110120 (2024).
7. Zhou C, Huang Y, Pu M, et al.: The treasure beneath multiple annotations: An uncertainty-aware edge detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 15507-15517. Institute of Electrical and Electronics Engineers (IEEE) (2023).
8. Banerjee S, Lyu J, Huang Z, et al.: Ultrasound spine image segmentation using multi-scale feature fusion skip-inception U-NET (SIU-NET). Biocybernetics and Biomedical Engineering 42(1), 341-361 (2022).
9. Hou J, Zhou H, Hu J, et al.: A Multi-Scale Convolution and Multi-Layer Fusion Network for Remote Sensing Forest Tree Species Recognition. Remote Sensing 15(19), 4732 (2023).
10. Guo F, Liu J, Lv C, et al.: A novel transformer-based network with attention mechanism for automatic pavement crack detection. Construction and Building Materials 391, 131852-131861 (2023).
11. Ni H, Shi Z, Karungaru S, et al.: Classification of typical pests and diseases of Rice based on the ECA attention mechanism. Agriculture 13(5), 1066-1080 (2023).

12. Xu L, Yang P, Yu J, et al.: Extraction of cropland field parcels with high resolution remote sensing using multi-task learning. European Journal of Remote Sensing 56(1), 2181874-2181897 (2023).
13. Xian R, Lugu R, Peng H, et al.: Edge detection method based on nonlinear spiking neural systems. International journal of neural systems 33(01), 2250060 (2023).
14. Pu M, Huang Y, Liu Y, et al.: Edter: Edge detection with transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 1402-1412. Institute of Electrical and Electronics Engineers (IEEE) (2022).
15. Wei Z, Chang M, Zhong Y.: Fruit Freshness Detection Based on YOLOv8 and SE attention Mechanism. Academic Journal of Science and Technology 6(1), 195-197 (2023).
16. Cole R C, Espinoza A I, Singh A, et al.: Novelty-induced frontal–STN networks in Parkinson's disease. Cerebral Cortex 33(2), 469-485 (2023).