ISSN 2616-5775 Vol. 8, Issue 9: 1-8, DOI: 10.25236/AJCIS.2025.080901

Enhancing Image Recognition with Adaptive Interaction and Cross Attention in Convolutional Neural Network

Liwen Kong[1,a], Jianqiang Mei[1,b,*], Fan Jia[2,c], Weixiang Du[3,d]

[1]*School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin, China*
[2]*Raysov Instrument Co. Ltd., Dandong, China*
[3]*Gansu Province Special Equipment Inspection & Testing Research Institute, Lanzhou, Gansu, China*
[a]*liwen_kong@163.com,* [b]*meijianqiang@tute.edu.cn,* [c]*info@raysov.com,* [d]*119273184@qq.com*
**Corresponding author*

Abstract: *Convolutional Neural Network (CNN)-based classifiers have been extensively employed in image recognition tasks. However, as CNN networks continue to deepen, existing deep architectures often result in a large number of parameters and substantial model sizes. Although deep features often contain rich semantic information, the continuous deepening of the network leads to a loss of detailed target information due to resolution reduction, ultimately decreasing image recognition accuracy. To address this issue, we propose a convolutional classifier that incorporates adaptive interaction and cross attention mechanisms. In this study, we design a VGG-like network where the adaptive interaction module enhances the feature transformation process of traditional convolution. This module expands the receptive field of convolutional kernels, adaptively constructs spatial and channel relationship indicators, and outputs more discriminative feature representations. Additionally, the cross attention module effectively captures global contextual information, enabling the network to learn spatial dependency relationships among features. Our proposed method is compared with both classical and state-of-the-art classification models, and experimental results on the CIFAR-10 dataset demonstrate that our method achieves the highest accuracy of 88.97%. This advancement will contribute to the improved capture of advanced semantic features in the domain of image recognition and target measurement.*

Keywords: *Image Recognition, Classifier, Convolutional Neural Network, Adaptive Interaction, Cross Attention*

1. Introduction

In the past decade, deep neural networks, and especially convolutional neural networks (CNNs), have been extensively utilized as the foundational architecture for training large-scale datasets in image classification tasks. These networks rely heavily on convolution operations to provide robust feature extraction and representation capabilities, significantly contributing to the success of various downstream applications. Among the ongoing efforts to enhance CNNs, improving the representations of advanced semantic features stands out as a critical research direction, particularly for boosting the performance of image classification systems.

Recent advancements in deep learning have spurred the evolution of image classification models, with convolutional networks like VGGs [1] and ResNets [2] demonstrating remarkable progress in capturing high-level abstractions. While these models have outperformed traditional networks like AlexNet [3] and ZFNet [4], their improvements are primarily achieved through stacking convolution and pooling layers. This approach, while enhancing the network's ability to represent complex features, is limited by the receptive field size, hindering the acquisition of contextual information necessary for handling natural scenes with multiple categories [5]. Additionally, excessive network layers lead to increased model size and parameter count, potentially resulting in the loss of detailed information. To address these challenges, recent methods have integrated non-local blocks [6] or distributed convolutions [7] into CNNs to improve their capacity to capture spatial dependency relationships and generate richer feature representations. Attention mechanisms, employing large kernel convolution operators to capture local contextual information, have also shown promise in this regard [9-12]. However, their ability to connect information using square-shaped kernels remains limited.

ISSN 2616-5775 Vol. 8, Issue 9: 1-8, DOI: 10.25236/AJCIS.2025.080901

In this paper, we introduce an adaptive interaction module aimed at enhancing the discriminative representation of features, thereby replacing the traditional convolution operation module. This module leverages multiple convolutional filters to independently transform each component, utilizing one output to interact with the outputs of the other filters. The close interaction between these filters enables the network to adaptively encode spatial contextual information, resulting in the generation of more discriminative feature representations. In real-world scenarios, where interference from unrelated areas is common [8], we further propose a cross attention module. This module constructs long and narrow kernels along the horizontal and vertical dimensions, optimizing the encoding of horizontal and vertical contextual information. This approach improves the network's ability to focus on the spatial location of features, thereby capturing spatial dependency relationships more efficiently.

The remainder of this paper is structured as follows: Section 2 introduces the proposed method, providing detailed explanations of the adaptive interaction module and cross attention module. Section 3 presents the entire process and results of the experiments, including both quantitative and qualitative evaluations. Finally, Section 4 offers a summary of the paper's contributions and findings.

2. Methodology

In this section, we first introduce the overall architecture of the proposed method. Subsequently, we provide detailed descriptions of the adaptive interaction module and the cross attention module, respectively.

2.1 Overall Architecture

As illustrated in Figure 1, the proposed network adopts a streamlined architecture inspired by VGG. Initially, the input image undergoes a standard 3×3 convolution to extract fundamental low-level features. Subsequently, these features are processed through an adaptive interaction module (AIM), which dynamically encodes spatial contextual information, expands the receptive field of the convolution, and produces more distinctive feature representations. This module notably enhances the network's capacity to capture intricate spatial relationships. Furthermore, a cross attention module (CAM) is incorporated to efficiently encode and integrate both horizontal and vertical information, thereby further boosting the network's ability to comprehend complex spatial dependencies among features. The combined use of these modules significantly improves the network's performance.

To maintain the integrity of the original features, a residual connection is implemented, followed by a standard 2×2 maxpooling layer to reduce the dimensionality of the feature map. The features then pass through another 3×3 convolution layer to extract more abstract, higher-level image features. These features are iteratively refined through the AIM to enhance their representations. To enhance the network's robustness and mitigate overfitting, a dropout layer with a rate of 0.4 is strategically incorporated. Finally, the refined features are fed into a fully connected layer to produce the final prediction result. This design ensures that the network's output is not only accurate but also reliable, with added measures to improve generalization and performance.

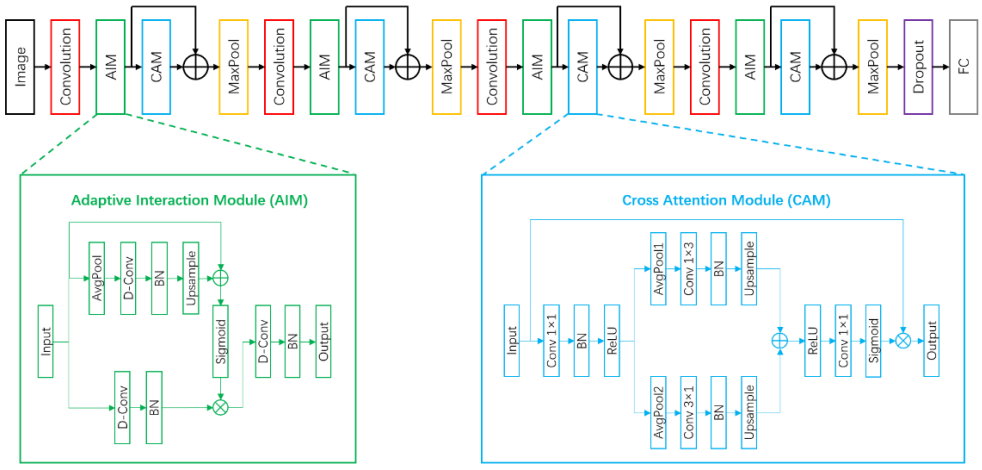


Figure 1: Overall architecture of the proposed method.

ISSN 2616-5775 Vol. 8, Issue 9: 1-8, DOI: 10.25236/AJCIS.2025.080901

2.2 Adaptive Interaction Module

Image classification tasks involving natural scenes typically necessitate that the network possess the capability to represent advanced abstractions. However, the receptive field of traditional 2D convolution operations is primarily governed by the predefined kernel size, leading to a lack of sufficiently large receptive fields to capture comprehensive high-level semantic features. This limitation often impedes the flow of contextual information within the network. To address the aforementioned challenges, we introduce an adaptive interaction module. As illustrated in the green part of Figure 1, the process of this module is mainly divided into three parts, each part contains a dilated convolution filter (D-Conv).

Considering the input X with dimensions C×H×W, where C represents the channel, H and W represent the height and width, respectively. From top to bottom, we designate the three inputs as X_1, X_2, and X_3, each with dimensions C×H×W. Notably, X_1 shares the same feature maps as the original input X. In the path of X_2, we initially apply an average pooling operation to downsample the input features, resulting in a pooled shape of C×(H/r)×(W/r). Here, the kernel size and stride are both set to r×r (with r=4 in this study). Following a dilated convolution filter, the number of output channels remains unchanged. Subsequently, we upsample the feature map size via bilinear interpolation to restore its original spatial dimensions, yielding an output shape of C×H×W, which generates feature weights X_2' to interact X_1. The interaction operation can be expressed as:

$$X_2' = \mathcal{F}(\mathrm{Conv_d}(\mathrm{AvgPool_r}(X_2))) \tag{1}$$

Where d represents the dilation rate of convolution, which is set to 2 in this study. $\mathcal{F}$ represents the bilinear interpolation operator used for calculating feature map size's upsampling. In the path of X_3, we only perform a dilated convolution to generate X_3' with the shape of C×H×W. This process can be simply expressed as:

$$X_3' = \mathrm{Conv_d}(X_3) \tag{2}$$

Unlike the processing of X_2, our goal is to extract contextual information from various spatial scales. Subsequently, we add X_2' and X_1, and the resulting feature map is element-wise multiplied with X_3'. This operation not only facilitates adaptive interaction of contextual information surrounding each spatial position but also captures the interdependencies among channels. Lastly, the output of this module is obtained by using the result of the final dilated convolution. The aforementioned procedure can be formally expressed as follows:

$$\mathrm{Output} = \mathrm{Conv_d}(X_3' \otimes \sigma(X_2' \oplus X_1)) \tag{3}$$

In which $\oplus$ denotes element-wise addition, $\otimes$ denotes element-wise multiplication, and σ denotes the Sigmoid activation function.

To validate the performance of the adaptive interaction module, we employ VGG-16 as a case study. While preserving the overall architecture, we adjust the network parameters of VGG-16 to accommodate training images with a resolution of 32×32. We substitute all 3×3 convolution layers in VGG-16 with the proposed adaptive interaction module (AIM) and visualize the network's receptive field using Grad-CAM (Gradient-weighted Class Activation Mapping). For this visualization, we select the intermediate feature map from the final convolution operation module. As illustrated in Figure 2, the adaptive interaction module effectively enlarges the receptive field of traditional convolutions, enabling accurate localization of target features. It is apparent that VGG-16 equipped with AIM captures more extensive contextual information.

Figure 2: The visualization of intermediate feature maps produced by VGG-16 is presented. The first row displays the original input image, whereas the second and third rows exhibit the visualization outcomes of intermediate feature maps generated by VGG-16 using the standard convolution module and the proposed adaptive interaction module, respectively.

ISSN 2616-5775 Vol. 8, Issue 9: 1-8, DOI: 10.25236/AJCIS.2025.080901

In contrast to dynamic group convolution [13], our approach organizes convolutional filters in a heterogeneous manner, with each filter serving a distinct purpose. This arrangement facilitates the fusion of contextual information from multiple spatial scales. Compared to traditional convolution, our proposed module significantly enlarges the receptive field of the layer, thereby enabling the generation of more discriminative feature representations.

2.3 Cross Attention Module

When identifying targets with irregular shapes and sizes in real scenes, the feature information of targets is often inevitably interfered by unrelated areas, which can affect the network's ability to correctly recognize images.

Therefore, we propose a cross attention module. As shown in the blue part of Figure 1, the module can be roughly divided into three steps from left to right. Firstly, the channel of input is compressed, and the feature map size is pooled horizontally and vertically, respectively. Then, they are upsampled to the original spatial size and fused by element-wise addition. Finally, the channel is extended to the original size, and the original input is reweighted by the fused result, which serves as output of the module.

Given the shape of the original input X as C×H×W, the first 1×1 convolution operation compresses the input channel in the ratio of α (which is set to 4 in this study). The output tensor is set as X' with the shape of (C/α)×H×W, which can be expressed as:

$$X' = \mathrm{ReLU}(\mathrm{Conv}_{1\times1}(X)) \tag{4}$$

Next, we perform avepooling operations on X' vertically and horizontally along two sub-paths. As shown in Figure 1, $\mathrm{AvgPool}_1$ represents vertical pooling, while $\mathrm{AvgPool}_2$ represents horizontal pooling. After pooling, the spatial ranges convert to (1, W) and (H, 1), respectively. Unlike 2D pooling of the square window, these pooling involve summing and averaging the values on columns and rows, respectively. Thus, the vertical pooling calculation and horizontal pooling calculation can be represented as:

$$\overline{P}_j = \frac{1}{H}\sum_{1\le i\le H} p_{(i,j)} \tag{5}$$

$$\overline{P}_i = \frac{1}{W}\sum_{1\le j\le W} p_{(i,j)} \tag{6}$$

Where $p_{(i,j)}$ represents the value of the i-th row and j-th column, $\overline{P}_i$ and $\overline{P}_j$ respectively represent the value of a certain column and a certain row after avepooling. These pooling can help collect remote contextual information from different spatial dimensions.

Then, we use a 1D convolution with the kernel size of 3 to encode the vertically and horizontally pooled regions separately, modulating the feature information of each spatial position and its adjacent areas. Due to the long and narrow pooled region, 1D convolution can easily establish long-range dependency relationships between discrete feature positions. The operations of encoding for the vertically pooled region and horizontally pooled region are shown as:

$$X'_v = \mathrm{Conv}_{1\times3}(\mathrm{AvgPool}_1(X')) \tag{7}$$

$$X'_h = \mathrm{Conv}_{3\times1}(\mathrm{AvgPool}_2(X')) \tag{8}$$

Where X_v' has the shape of (C/α)×1×W, and X_h' has the shape of (C/α)×H×1. We upsample the spatial sizes of X_v' and X_h' to H×W using interpolation algorithms, which allows the value of each column or row to be expanded vertically or horizontally, respectively. In order to obtain more global contextual information, the upsampled results are element-wise added and then extended in channel dimension, forming cross attention features. This operation can be expressed as:

$$X'_c = \mathrm{Conv}_{1\times1}(\mathrm{ReLU}(\mathcal{F}(X'_v)\oplus\mathcal{F}(X'_h))) \tag{9}$$

In which X_c' has the shape of C×H×W. Finally, X_c' is activated by the Sigmoid function and then reweights the original input X by element-wise multiplication, serving as the output of the module. This operation can be shown as:

$$\mathrm{Output} = X\otimes\sigma(X'_c) \tag{10}$$

All the aforementioned processes are illustrated in Figure 3. The purple box, situated at the intersection of the red and blue stripes, serves to integrate the spatial contextual information from both vertical and horizontal dimensions. It establishes dependency relationships with the spatial positions of

ISSN 2616-5775 Vol. 8, Issue 9: 1-8, DOI: 10.25236/AJCIS.2025.080901

the original input, thereby enhancing the network's ability to comprehend the underlying structure of the data.

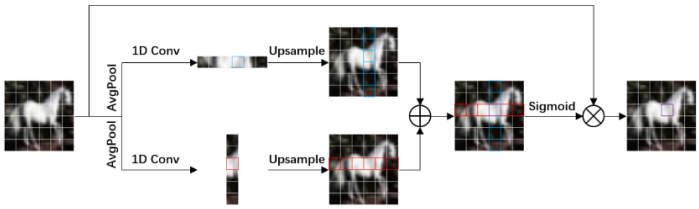


Figure 3: The illustration of the cross attention module.

In comparison to the attention module proposed in [14], our cross attention module necessitates significantly less computational overhead for spatial transformations. Consequently, our module is more lightweight and capable of capturing spatial dependency relationships more efficiently. This enhanced efficiency allows the network to better concentrate on the features of targets, thereby improving its overall performance.

3. Experiments

In this section, we present the experimental setup and results to evaluate the performance of the proposed method. Firstly, we provide details regarding the dataset used in our experiments. Following this, we outline the experimental settings and implementation specifics. Subsequently, we introduce the performance metrics employed for evaluation. Finally, we present and analyze the experimental results.

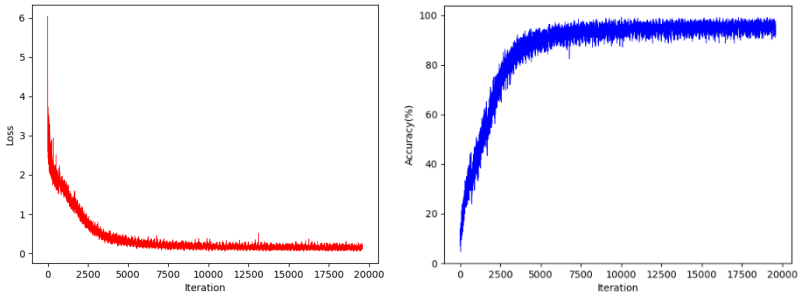
3.1 Dataset

In this study, we employ the CIFAR-10 dataset, a well-established benchmark specifically designed for image classification tasks. The dataset consists of 60,000 color images, each with a resolution of 32×32 pixels, evenly distributed across 10 categories, with 6,000 images per category. The dataset is partitioned into a training set and a testing set in an 8:2 ratio, comprising 50,000 and 10,000 images, respectively. It should be noted that the dataset remains unaltered and unprocessed during our experiments, ensuring a fair evaluation.

3.2 Experimental Settings and Implementation Details

The experiments were conducted on an NVIDIA GeForce RTX3050 GPU and an Intel Core i7-12700H CPU, using the PyTorch 1.12.0 framework. The hyper-parameters for training were carefully selected: the initial learning rate was set to 0.01, the optimizer chosen was Stochastic Gradient Descent (SGD) with a momentum of 0.9, and the weight decay was set to $5e^{-3}$. The training was carried out over 100 epochs with a batch size of 256, resulting in a total of 19,600 iterations. The loss function utilized was Cross Entropy.

Based on these settings, the proposed model was trained on the CIFAR-10 dataset. To evaluate the training process, loss and accuracy curves were plotted. Figure 4(a) depicts the loss curve, showcasing the reduction in loss over epochs. Figure 4(b), on the other hand, displays the accuracy curve, illustrating the improvement in model performance as training progresses. These curves provide insights into the training dynamics and help assess the model's learning efficiency.



(a) The training loss curve. *(b) The training accuracy curve.*

Figure 4: The training curves of the proposed method on CIFAR-10 dataset.

ISSN 2616-5775 Vol. 8, Issue 9: 1-8, DOI: 10.25236/AJCIS.2025.080901

3.3 Evaluation Results

For multi-class image classification tasks, the performance of the model is mainly evaluated by the following metrics:

■ Test accuracy represents the percentage of correctly predicted samples across total testing samples which can be calculated as:

$$\text{Test accuracy} = \frac{\text{Correctly predicted samples}}{\text{Total testing samples}} \times 100\%. \quad (11)$$

■ Params is used to measure the size of the model, which reflects the complexity of the model and has a significant impact on model's generalization and computation requirements.

■ Confusion matrix is usually used to evaluate the performance of classifiers. It provides visual representation of the classification performance and shows the relationship between true labels and predicted labels in the form of the matrix.

To validate the superiority of the proposed method in image classification, we conducted experiments using various classic models, including VGG-16, VGG-19, and ResNet-34, all adapted to handle images with a resolution of 32×32. We trained these models using the same dataset and settings as our proposed method, and evaluated their performance using the aforementioned metrics.

Table 1 presents a comparison of the test accuracy and params of different methods. Notably, CreINN [15], a recent method proposed in 2025 specifically for CIFAR-10 image classification, is also included for comparison. It is evident from the table that our proposed method achieves the highest test accuracy of 88.97% while having the smallest params. Figure 5 displays the confusion matrix of our proposed method, demonstrating its ability to accurately recognize a wide range of targets in natural images.

Table 1: The performance of each method.

Method	Test accuracy (%)	Params (M)
VGG-16	79.75	56.18
VGG-19	81.13	76.45
ResNet-34	79.84	81.18
CreINN	85.03	47.21
Proposed	88.97	43.39

The confusion matrix in Figure 5 provides a visual representation of the classification performance of our proposed method. Each row of the matrix represents the instances in an actual class, while each column represents the instances predicted to be in that class. The diagonal elements show the number of correctly classified instances, while off-diagonal elements indicate misclassifications. The high values along the diagonal and the relatively low values off-diagonal indicate that our method is effective in distinguishing between different classes, thereby confirming its superior performance in image classification tasks.

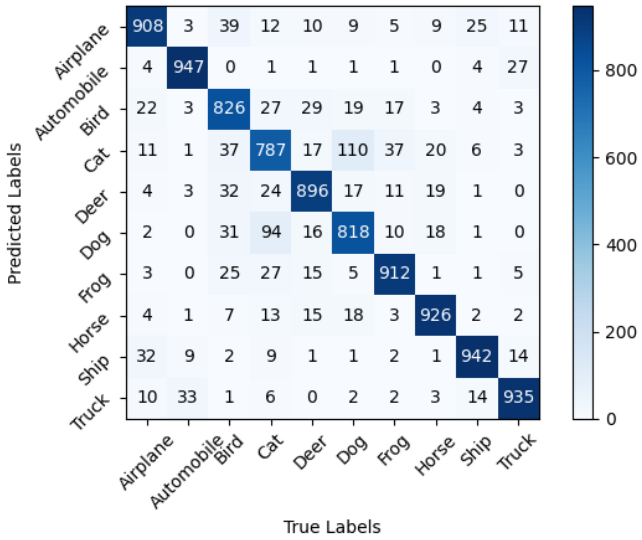


Figure 5: The confusion matrix of the proposed method.

ISSN 2616-5775 Vol. 8, Issue 9: 1-8, DOI: 10.25236/AJCIS.2025.080901

To assess the efficacy of each module introduced in this paper, we conducted a series of ablation experiments. Our baseline model consists of replacing the adaptive interaction module (AIM) with a traditional 3×3 convolution module, and removing both the cross attention module (CAM) and residual connections from the proposed architecture. The results of these ablation experiments are presented in Table 2. The baseline method, denoted as Method A, achieves a test accuracy of 77.83% with a params of 17.95M.

Table 2: The results of ablation experiments.

Method	AIM	CAM	Test accuracy (%)	Params (M)
A			77.83	17.95
B	√		84.46	41.88
C	√	√	88.97	43.39

In the next experiment, denoted as Method B, we replaced the traditional convolution module before each maxpooling layer with the AIM. This modification led to a notable 6.63% increase in test accuracy, albeit with a corresponding 23.93M increase in params. The incorporation of the AIM allows the network to capture a more extensive and richer spatial receptive field, thereby enhancing the overall feature representations. This demonstrates the significant contribution of the AIM to the network's performance.

Finally, we evaluated the proposed model, denoted as Method C, which includes the addition of the CAM and residual connections after each AIM, in comparison to Method B. The integration of the CAM resulted in a further 4.51% increase in test accuracy, with only a modest 1.51M increase in params. These results underscore the effectiveness of the CAM in capturing spatial dependency relationships and improving the network's attention to target features. Overall, the ablation experiments provide compelling evidence of the contributions of each module to the network's performance and robustness.

4. Conclusion

To address the challenge of insufficient classification accuracy in image recognition, this paper introduces a CNN-based method that leverages adaptive interaction and cross attention mechanisms. The adaptive interaction module effectively enlarges the receptive field of the network and enhances the discriminative representation of features, while the cross attention module efficiently establishes spatial dependency relationships among features and improves the network's ability to focus on targets. By addressing the limitations of other classification models in capturing spatial contextual information, the proposed method achieves the highest test accuracy of 88.97% on the CIFAR-10 dataset. This achievement underscores the method's capability in recognizing advanced semantic features of various targets, thereby contributing to the advancement of image recognition technology.

Acknowledgements

This work is supported by teaching Reform and Quality Construction Research Project of Tianjin University of Technology and Education (Grant No.JGY2022-03).

References

[1] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[2] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.
[4] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. Cham: Springer International Publishing, 2014: 818-833.
[5] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.
[6] Sun Z, Sun H, Zhang M, et al. A Non-Local Block With Adaptive Regularization Strategy[J]. IEEE Signal Processing Letters, 2024, 31: 331-335.
[7] Xiao Z, Ye K, Cui G. Differential self-feedback dilated convolution network with dual-tree channel attention mechanism for hyperspectral image classification[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 73: 1-17.

ISSN 2616-5775 Vol. 8, Issue 9: 1-8, DOI: 10.25236/AJCIS.2025.080901

[8] He J, Deng Z, Zhou L, et al. Adaptive pyramid context network for semantic segmentation[C] //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7519-7528.

[9] Guo M H, Lu C Z, Liu Z N, et al. Visual attention network[J]. Computational visual media, 2023, 9(4): 733-752.

[10] Peng C, Zhang X, Yu G, et al. Large kernel matters--improve semantic segmentation by global convolutional network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4353-4361.

[11] Azad R, Niggemeier L, Hüttemann M, et al. Beyond self-attention: Deformable large kernel attention for medical image segmentation[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2024: 1287-1297.

[12] Lau K W, Po L M, Rehman Y A U. Large separable kernel attention: Rethinking the large kernel attention design in cnn[J]. Expert Systems with Applications, 2024, 236: 121352.

[13] Su Z, Fang L, Kang W, et al. Dynamic group convolution for accelerating convolutional neural networks[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 138-155.

[14] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3146-3154.

[15] Wang K, Shariatmadar K, Manchingal S K, et al. Creinns: Credal-set interval neural networks for uncertainty estimation in classification tasks[J]. Neural Networks, 2025, 185: 107198.