IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 January 2024, accepted 14 February 2024, date of publication 21 February 2024, date of current version 28 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368801

RESEARCH ARTICLE

PCSA: Enhancing CNN Performance With Pyramid Channel and Spatial Attention

YEHANG ZHANG
School of Computer Science, China University of Geosciences, Wuhan 430074, China
e-mail: ZhangYeHang2002@outlook.com

ABSTRACT Recent studies have demonstrated that the attention mechanism can effectively enhance the effectiveness of deep convolutional neural networks. In this paper, we propose a ''Pyramid channel and spatial attention'' (PCSA), which consists of reconstructing the features after pyramidal multiscale convolution by extracting spatial weights and channel weights. This dual weight extraction process helps to merge the multiscale information more accurately and enhances the model's focus on the complex locations of image objects. As a plug-and-play module, PCSA can be easily added to various backbone networks to enhance the modeling effect. We apply the PCSA module to two kinds of backbone networks, VGG and ResNet, and the improved models are named: VGG-PCSA and PCSANet, respectively. Experimental results show that on the CIFAR-10, CIFAR-100, and NaSC-TG2 datasets, our model has a significant performance improvement over the backbone networks while keeping the number of parameters low and performs better than most of the state-of-the-art channel attention methods. In addition, we visualize feature maps and class activation diagrams to explain the better performance of PCSA.

INDEX TERMS Image classification, attention mechanism, CNNs.

I. INTRODUCTION

Attention mechanism diverts attention to the most important areas of the image and ignores extraneous parts. It play a key role in computer vision by learning and focusing on the most relevant features in an image to improve performance in tasks such as image recognition, target detection, and image segmentation [1], [2], [3], [4], [5], [6], [7], [8].

SE [9] first proposed an attention mechanism for learning channel information. Attention mechanisms are mainly divided into two types, channel attention represented by SE mechanism and spatial attention represented by CBAM [10]. Based on SE, ECA-Net [11], which models inter-channel information, and Fca-Net [12], which is based on the frequency domain, are proposed. CBAM used average pooling and maximum pooling to introduce spatial attention, and achieved multiplexed attention that fuses channel and spatial information. Distinct from CBAM, BAM [13] and DA-Net [6] used parallel ideas to combine channel and spatial attention. However, they also suffer from not exploring multi-scale information and establishing long-distance dependencies. To solve the multi-scale feature extraction problem, the researchers proposed SK-Net [14] using different sizes of convolutional kernels and branching designs, which makes the feature maps of different receptive field branches have different importance. PyConv [15] proposed a pyramidal multi-scale grouped convolution module to extract multi-scale features. Coordinate Attention [16] constructed global dependencies by embedding spatial information in the channel feature maps. EPSA (Efficient Pyramid Squeeze Attention) [17] mechanism combines pyramidal convolution and channel attention, which has a more granular multi-scale representation capability and develops long-range channel dependency. To address these limitations, we propose a novel attention mechanism PCSA (Pyramid Channel and Spatial Attention). PCSA combines spatial weights with multi-scale channel weight extraction, integrating global location information into the feature map constructed from channel weights to establish a global dependency. Furthermore, as a plug-and-play module, PCSA needs to be applied to a backbone network to function. In this paper, two backbone networks, VGG and ResNet, are used.

The associate editor coordinating the review of this manuscript and approving it for publication was Yudong Zhang.

Firstly, VGG [18] is one of the milestones of convolutional neural networks and is suitable as a backbone network because of its simple and effective structure. However, it also has problems such as inadequate feature extraction and large number of model parameters. To solve these problems, researchers have developed many improved VGG models in terms of simplifying the model structure [19], [20] and adding attention to enhance the model effectiveness [21], [22], [23]. However, these models suffer from the drawbacks of not considering multi-scale information and network degradation after model deepening. To address these problems, we improve VGG using the PCSA mechanism and propose a new VGG-PCSA model that achieves the unification of feature extraction capability and number of parameters. Secondly, to further validate the generalization performance of the PCSA module, we propose a ResNet-style network PCSANet. It is more common to combine attention mechanisms with ResNet, and the network structure is simpler compared to VGG-PCSA. The main contributions of this work include.

- We propose a new attention mechanism, PCSA block, which can extract and combine multi-scale information more efficiently by fusing channel weights and spatial weights. As a plug-and-play module, the PCSA block can be applied to the backbone network for improving the performance of the model.
- We apply PCSA to two backbone networks, VGG and ResNet, and the improved models are named VGG-PCSA and PCSANet, respectively. They can learn richer multi-scale features with fewer parameters and adaptively adjust inter-channel weights precisely.
- We conducted extensive experiments on CIFAR-10, CIFAR-100, and NaSC-TG2 and showed that PCSA obtained better performance than other attention mechanisms.

II. RELATED WORK

This section first introduces related work on attention mechanisms, and then presents research on the application of attention mechanisms to VGG and ResNet backbone networks.

A. ATTENTION MECHANISM

SE (Squeeze-and-Excitation) [9] is a classical and commonly employed attention mechanism, which models channel attention to enhance performance. However, SE overlooks spatial attention and fails to consider inter-channel relationships. Spatial attention such as CBAM [10] and BAM [13] proposes spatial attention, aggregating both spatial and channel attention information more comprehensively and reliably. A^2Net proposes a dual-attention mechanism that collects key features of the space into a compact set and then adaptively distributes them to each location. Dual-attention [6] proposes a new dual parallel attention model based on dilated convolution. On the other hand, for the inter-channel relationship problem, ECA-Net [11] is proposed, which integrates adjacent channel information using 1×1 convolution to obtain more accurate channel attention. The subsequent Fca-Net [12] proves that the GAP of channel compression is a special case of feature decomposition in the frequency domain, and proposes a novel multi-spectral channel attention.

However, they still have two limitations: the inability to capture information at multiple scales and the inability to establish long-range channel dependencies. To tackle these challenges, researchers have proposed various methods for multi-scale information representation and cross-channel information interaction. Based on the first problem, the researchers proposed methods for multi-scale information representation and cross-channel information interaction. SK-Net [14] utilized parallel 3×3 and 5×5 convolutions to extract features in parallel. Inspired by SK-Net, PyConv [15] employed image pyramids to achieve multi-scale feature extraction. Differently from PyConv, which used different size convolutions, Res2Net [24] achieved multi-scale feature extraction by constructing hierarchical residual connections. To address the second problem, researchers proposed Coordinate Attention [16], which embedded location info into channel attention, capturing long-range information for understanding global dependencies. However, these approaches often suffer from the complexity of models and a high number of parameters.

In order to efficiently extract multi-scale features while establishing channel dependencies, EPSA [17]mechanism was proposed, which extracted the weights of multi-scale convolution groups through the SE mechanism and efficiently implements the global and local feature dependencies. However, EPSA only extracts channel weights and ignores the attention to spatial weights, and the resulting weights cannot accurately combine multi-scale feature maps. Based on this, this paper proposes a new attention mechanism, PCSA, which uses a improved pyramidal multi-scale convolution with simplified parameters to obtain multi-scale information, and fuses channel weights and spatial weights in parallel to obtain a more informative feature map.

B. ATTENTION MECHANISMS IN BACKBONE NETWORKS

Attention mechanisms are usually applied in different backbone networks to enhance the effectiveness in different computer vision tasks. Commonly used backbone networks in image classification tasks are ResNet and VGG, etc. SENet is composed by inserting the SE module into the residual structure of ResNet. Fca-Net, ECANet are similar to SENet, which are further improved based on ResNet. Unlike the above mentioned channel attention extraction of only the features extracted from the backbone network, the EPSA module also uses multi-scale convolution. EPSANet uses the PSA module instead of the convolutional layers of ResNet and recombines the features between each multiscale convolution using the channel attention weights. The PCSANet proposed in this paper is inspired by EPSANet, based on ResNet34, ResNet50 and incorporating new PCSA modules.

In addition to ResNet, we added PCSA to the VGG backbone network. VGG, as a classical convolutional neural network, has a small convolutional kernel, a large receptive field and a simple network structure. These advantages make VGG an important backbone network. Researchers have proposed a combination of multiple attention mechanisms and VGG to solve image classification tasks. Paper [23] improves VGG by adding SE mechanism. Distinct from channel attention, the paper [21] uses the addition of spatial attention on hopping connections to enhance the extraction of spatial information by VGG. In addition to focusing on feature information, Paper [25] proposes a discrete wavelet transform-based Wavelet- Attention mechanism and used it for the improvement of VGG. However, there are two problems with these improvements: firstly, the introduced attention mechanism lacks multi-scale consideration, and secondly, as the model becomes more complex, model degradation and excessive parameter amount of VGG. As a result, this paper proposes the new VGG-PCSA applied to the backbone network VGG.

III. CNN WITH PCSA MECHANISM

This section first presents a review of the channel weights and spatial weights extraction methods used by PCSA, then presents the detailed design of the novel attention mechanism PCSA, and finally introduces the newly proposed VGG-PCSA and PCSANet.

A. REVISITING CHANNEL AND SPATIAL ATTENTION

1) CHANNEL ATTENTION

In PCSA, CA (Channel Attention) [10] is used for the extraction of channel weights. The CA mechanism is shown in Fig. 1. It performs average pooling and global maximum pooling on the feature map, generating two distinct spatial context description vectors for each channel. These vectors represent the average pooling feature and the maximum pooling feature. A MLP (Multilayer Perceptron) is applied to model the channel relationship. The resulting description vectors are then combined through element-wise leveling and summation to obtain the channel attention vector. The CA mechanism can be defined as:

$$CA = Sig(Conv_{1\times1}(AvgPool(F)) + Conv_{1\times1}(MaxPool(F))) \quad (1)$$

$$Conv_{1\times1}\,(X) = W_1(\sigma(W_0(X))) \quad (2)$$

where the $W_0 \in \mathbb{R}^{c\times\frac{c}{r}}$ and $W_1 \in \mathbb{R}^{\frac{c}{r}\times c}$ is 1×1 convolution, σ is the ReLU activation function, X is the input feature, which can be the result of $Maxpool(F)$ or $AvgPool(F)$. We use 1×1 convolutional layers instead of fully connected layers to more effectively combine linear information between channels and achieve information interaction between channels.

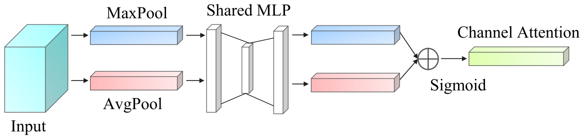


FIGURE 1. CA structure.

2) SPATIAL ATTENTION

In PCSA, SA (Spatial Attention) [10] mechanism is used for obtaining spatial information weights at the spatial pixel level. The SA structure is shown in Fig. 2. Different from channel attention, spatial attention is more concerned with information location. In computing spatial attention, firstly, average pooling and maximum pooling are performed in each channel, and the obtained feature map stitching is performed by channel, and then output after convolution layer and Sigmoid activation function. The SA mechanism can be defined as:

$$SA = Sig(Conv_{3\times3}([AvgPool(F_i); MaxPool(F_i)])) \quad (3)$$

where $Conv_{3\times3}$ is 3×3 convolution, Sig is Sigmoid activation function, AvgPool is global average pooling and MaxPool is global max pooling.

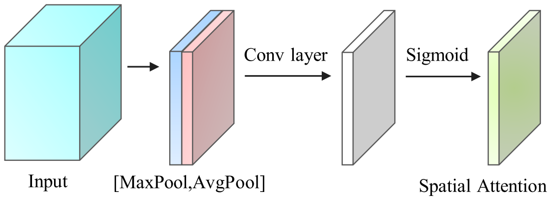


FIGURE 2. SA structure.

B. PCSA MECHANISM

Inspired by PSA, the motivation of this work is to establish a more precise and efficient attention mechanism. The PCSA (Pyramid Channel and Spatial Attention) can be divided into four parts. First, we use improved Pyramid Convolution Module to obtain grouped multiscale convolution results. Second, channel and spatial features are extracted in parallel for multi-scale features. Third, the channel and spatial features are rescaled with Softmax function to obtain the channel and spatial weights, respectively. Fourth, the features are reconstructed according to the 0-1 weights, and the adjusted features are linearly fused and output by 1×1 convolution adaptive combination.

We design new pyramidal multiscale convolution to extract multiscale features. PCM using pyramid-shaped multiscale dilated group convolution and point-wise group convolution. Compared to ordinary convolution, this design can effectively reduce the number of parameters. To extract multi-scale features, the inputs are divided into 4 groups for group convolution with receptive fields of [3, 5, 7, 9] and group sizes of [1, 4, 8, 16], respectively. The number of channels of input X is C, and the channel dimension of the multi-scale convolution group F_i is $\frac{C}{4}$. The output feature maps from these convolutions are combined to form a multi-scale representation. The dilated group convolution (DGC) operation can be defined as:

$$DF_i = Conv(k_i \times k_i, d_i, G_i)(X_i)\, i = 1, 2, 3, 4 \tag{4}$$

where $G_i \in \{1, 4, 8, 16\}$, $k_i \in \{3, 3, 7, 5\}$, $d_i \in \{0, 1, 0, 1\}$, $F \in \mathbb{R}^{C\times W\times H}$ is the multi-scale feature map after stitching by channel, DF is the grouped convolution result. Since the use of grouped convolution may result in some missing information, point-wise convolution (PWC) is then introduced to correct the information. In order to be consistent with the grouped convolution, the input to the PWC is also divided into 4 parts. The point-wise convolution operation can be defined as:

$$PF_i = Conv_{1\times 1}(X_i)\, i = 1, 2, 3, 4 \tag{5}$$

where PF_i is the point-wise convolution result. After that we directly sum DF and PF to cancel out the grouping information loss.

$$F_i = DF_i + PF_i\, i = 1, 2, 3, 4 \tag{6}$$

Next, channel attention (CA) extraction and spatial attention (SA) extraction are performed on F, respectively.

$$CA_i = ChannelWeight(F_i), i = 0, 1, 2, 3 \tag{7}$$

$$SA_i = SpatialWeight(F_i), i = 0, 1, 2, 3 \tag{8}$$

where $CA_i \in R^{C_i\times 1\times 1}$ is the channel weight vector and $SA_i \in R^{1\times W\times H}$ is the spatial weight matrix. After extracting the features, soft attention is used to adaptively select the weights of the different channels and spatial pixels.

$$CW_i = Softmax(CA_i) = \frac{exp(CA_i)}{\sum_{i=0}^{4} exp(CA_i)} \tag{9}$$

$$SW_i = Softmax(SA_i) = \frac{exp(SA_i)}{\sum_{i=0}^{4} exp(SA_i)} \tag{10}$$

We multiply the two weights with their extracted feature counterparts and add the results linearly to achieve the fusion of local and global features.

$$Z_i = CW_i \odot F_i + SW_i \odot F_i \tag{11}$$

where the $\odot$ is the corresponding multiplication of elements. The multi-scale feature maps with reassigned weights are stitched by channel to obtain the final output with fused channel and spatial information.

$$Z = Cat\ ([Z_1, Z_2, Z_3, Z_4]) \tag{12}$$

By linearly fusing the features of the channel weight combination with those of the spatial weight combination, we obtain a new and more accurate multi-scale feature combination. The fused feature map constructs the inter-relationship between channels by 1×1 convolution. The use of 1×1 convolution kernel can increase the nonlinear characteristics and make the network deeper by adding the activation function while keeping the feature map scale unchanged. In summary, PCSA integrates multi-scale information more precisely and enhances the model's focus on global information through the embedding of spatial information.

C. APPLICATION OF PCSA IN VGG: VGG-PCSA

For the problems of inadequate VGG feature extraction, easy overfitting, network degradation and excessive number of parameters, we propose the VGG-PCSA improvement model. VGG-PCSA uses VGG16 as the backbone network and utilizes BN (batch normalization), residual structure and PCSA mechanism. The model structure is shown in Fig. 4.

First, we insert the BN layer between the convolutional layer and the activation function to regulate the input of each layer. Then, we add the newly proposed PCSA mechanism, which enable VGG16 to better extract the channel and spatial information capability of multi-scale features. By calibrating the channel weights and spatial weights of multi-scale features, the model obtains a feature map with more reasonable information interaction between channels and richer features contained, which improves the ability of the model to extract features. However, the addition of PCSA makes the network deeper and network degradation occurs. To alleviate this problem, we use a residual structure that adapts to VGG channel changes and downsampling mechanisms.

Specifically, we use a 3 × 3 convolution layer with a step size of 2 to achieve downsampling and channel variation on the strip edges. This design not only preserves the extraction of texture and edge features by maximum pooling while downsampling, but also preserves a certain amount of local information due to the introduction of the convolutional layer on the residual connection to avoid the information loss that may result from pooling. Finally, the model uses an adaptive average pooling layer to replace the fully connected layer with an excessive amount of parameters.

D. APPLICATION OF PCSA IN RESNET: PCSANET

The novel PCSA module integrates multiscale information and constructs inter-channel links through multiscale feature extraction, generating spatial and channel weights. The inclusion by residual structure and 1×1 convolution makes PCSA construct long-distance spatial dependency based on the extraction of multi-scale attention. The new PCSANet is obtained by stacking the PCSA modules according to the ResNet style. The replacement PCSANet module is composed as shown in Fig. 3. The PCSANet synthesizes the layers of ResNet18 and the Block structure of ResNet50, which allows it to maintain an excellent feature extraction capability with a smaller number of parameters. The structure of the PCSANet is shown in Table 1.

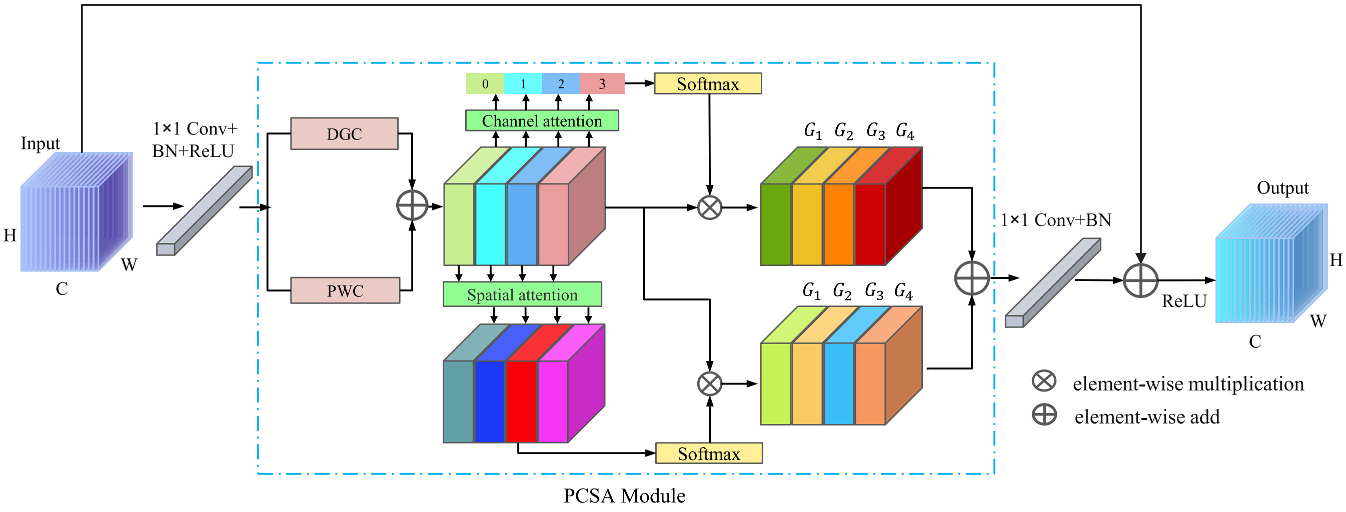


FIGURE 3. Replacing 3×3 convolution in ResNet with PCSA module.

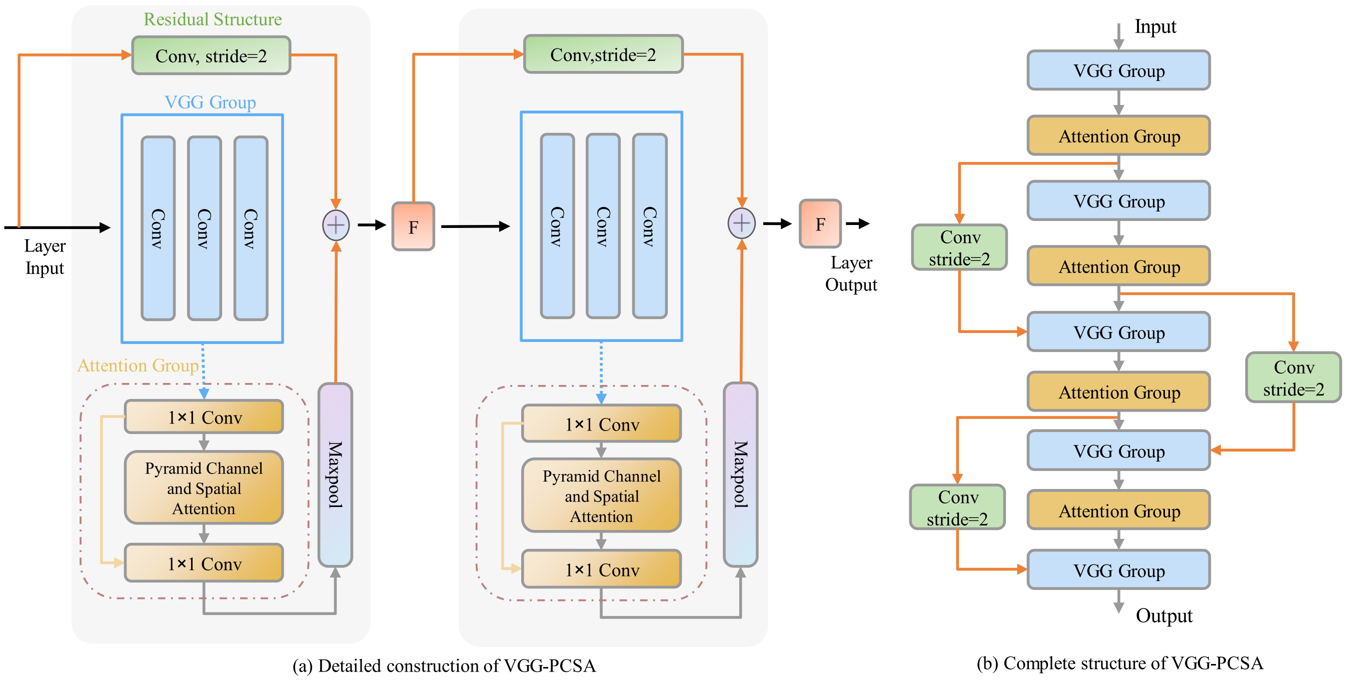


FIGURE 4. VGG-PCSA structure.

IV. EXPERIMENTS

In this section, we conduct extensive experiments on three data sets to analyze the performance of the proposed VGG-PCSA and PCSANet. The image classification benchmarks include CIFAR-10, CIFAR-100 and NaSC-TG2. Top-1 accuracy is used as an evaluation metric for classification. In addition, we repeated the experiment several times to prevent the effect of fluctuation and took the mean value as the experimental result.

A. DATA SETS

In this paper, we assess the generalization capability of our model by conducting experiments on three benchmark datasets, including CIFAR-10, CIFAR-100 and NaSC-TG2.

1) CIFAR-10 DATA SET

CIFAR-10 [26] is a small dataset for identifying pervasive objects organized by Alex Krizhevsky and Ilya Sutskever.

TABLE 1. PCSANet structure.

Output	ResNet-50	PCSANet
112×112	7×7, 64, stride 2	
56×56	3×3 maxpool, stride 2	
56×56	$\begin{bmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, & 64 \\ PCSA, & 64 \\ 1 \times 1, & 256 \end{bmatrix} \times 2$
28×28	$\begin{bmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, & 128 \\ PCSA, & 128 \\ 1 \times 1, & 512 \end{bmatrix} \times 2$
14×14	$\begin{bmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \\ 1 \times 1, & 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, & 256 \\ PCSA, & 256 \\ 1 \times 1, & 1024 \end{bmatrix} \times 2$
7×7	$\begin{bmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \\ 1 \times 1, & 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, & 512 \\ PCSA, & 512 \\ 1 \times 1, & 2048 \end{bmatrix} \times 2$
1×1	7×7, global average pool, 10/100 fc	

The dataset contains 60,000 32×32 color images divided into 10 categories with 6,000 images in each category.

2) CIFAR-100 DATA SET

The dataset has 60,000 color images with the same size of 32×32. There are in total 100 classes and each contains 600 images. Compared with CIFAR-10, CIFAR-100 divides the 100 classes into 20 superclasses, i.e., each image has a "fine" label and a "coarse" label.

3) NASC-TG2 DATA SET

NaSC-TG2 [27] is the Tiangong-2 natural scene dataset. The dataset includes 10 types of scenes: beach, circular farmland, clouds, desert, woodland, mountains, rectangular farmland, built-up areas, rivers, and snowy mountains, and there are 2000 color images per class. We use the division of paper [27] and [28], with 20% for the test set and 80% for the training set. The dataset preview is shown in the Fig. 5.

B. EXPERIMENTAL SETTINGS

Parameter settings of VGG-PCSA and PCSANet are consistent in the paper. We use cross-entropy loss function with label smoothing and set the label smoothing factor to 0.1. SGD optimizer with a weight decay factor of 0.01 is adopted. The learning rate is initialized to 0.05 and is reduced by half every 30 epochs. We set the batch size to 128 and the maximum epoch is 300. We use a cross-entropy loss function with label smoothing as the loss of the model with a label smoothing parameter of 0.1. For the CIFAR dataset, the images are scaled to 224 × 224 and then input to the model. For NaSC-TG2, we retained the original size of the dataset of 256 × 256 as input to the model. All experiments were run on a server with the following hardware and software environment. The operating system is Ubuntu 20.04 with NVIDIA GeForce GTX 3080 GPU and Intel(R) Xeon(R) Platinum 8255C CPU @2.50GHz. The methods were implemented on PyTorch.

C. INTEGRATION METHOD

In order to explore the optimal insertion location of PCSA blocks in VGG, we designed several integration methods, as shown in Fig. 6. The integration of PCSA blocks in VGG is a very simple process. 1) Standard, in which the PCSA is placed inside the residual structure, after the VGG layer. 2) PRE-PCSA, in which the PCSA block is placed inside the residual structure, before the VGG layer. 3) POST-PCSA, in which the PCSA block is placed outside the residual structure, after the VGG layer. The results in Table 2 show that the standardized integration method achieved the best results in all three datasets, with an improvement of 1.59%, 1.41%, and 0.94% over the least effective POST-PCSA method, respectively. It also shows that placing the PCSA module in the VGG block better utilizes the ability of PCSA to extract the channel and spatial weights of the multi-scale features and recombine them to significantly improve the performance of the model.

TABLE 2. Classification accuracy changes with the combination of weight extraction modules.

Methods	Top1-Accuracy(%)		
	CIFAR-10	CIFAR-100	NaSC-TG2
Standard	96.50	81.74	98.18
PRE-PCSA	96.13	81.41	98.06
POST-PCSA	94.91	80.33	97.24

D. ABLATION EXPERIMENTS FOR PCSA

We designed the ablation experiments shown in Table 3 to verify its effectiveness on each dataset by replacing the modules for extracting channel weights and spatial weights. The experiments in Table 3 show that the CA mechanism is more effective than the SE mechanism in channel weight extraction due to the consideration of spatial information. The SA mechanism, as a module for extracting spatial weights in parallel, allows PCSA to embed global spatial information in the generated feature maps, which improves the model performance.

E. COMPARISONS WITH OTHER METHODS

As depicted in Table 4, the accuracy of our VGG-PCSA model and PCSANet outperforms all previous networks in all cases. We first compared the effect of some basic backbone networks such as GhostNet and MobileNet. It can be seen that the use of PCSA makes the VGG and ResNet models gain significant improvement, in which VGG-PCSA improves 13.21%, 9.27%, 4.01% over VGG, and PCSANet improves 3.52%,6.04%, 9.42% over ResNet. Then we used SENet, ECANet, EPSANet, SCConv with ResNet as the backbone to compare with PCSANet, and the results showed that PCSANet achieved better results, and its accuracy on

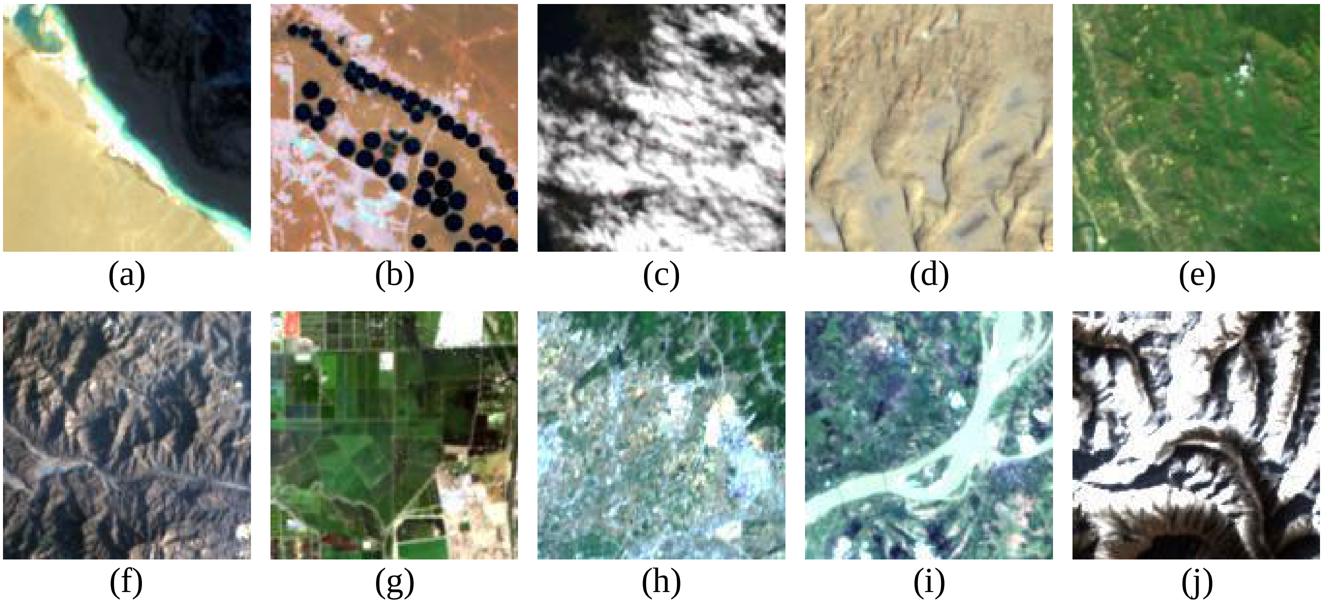
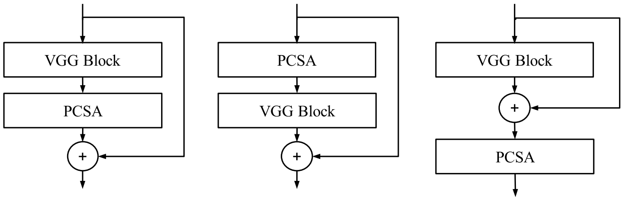


FIGURE 5. NaSC-TG2 dataset. (a) beach; (b) circularfarmland; (c) cloud; (d) desert; (e) forest; (f) mountain; (g) rectangularfarmland; (h) residential; (i) river; (j) snowberg.



(a) Standard (b) Pre-PCSA (c) Post-PCSA

FIGURE 6. Class activation map drawn on CIFAR-10 using the last layer of the network. Darker colors indicate higher model attention.

TABLE 3. VGG-PCSA classification accuracy changes with the combination of weight extraction modules.

Methods	Attention			Top1-Accuracy(%)		
	SE	CA	SA	CIFAR-10	CIFAR-100	NaSC-TG2
VGG-PCSA	✓			96.06	79.82	97.57
	✓		✓	96.19	80.13	97.98
		✓		96.24	80.56	98.11
		✓	✓	96.50	80.74	98.18
PCSANet	✓			96.27	80.32	97.68
	✓		✓	96.37	80.58	97.71
		✓		96.28	80.16	97.64
		✓	✓	96.43	80.92	97.79

CIFAR-10 reached the highest in the table, 96.50%. In addition, we replaced the PCSA module in VGG-PCSA with the attention mechanism mentioned above, keeping the other structures unchanged, and these models were named VGG-SE, VGG-CBAM, VGG-ECA, and VGG-EPSA. The results showed that VGG-PCSA was the most effective, and obtained on CIFAR-100, NaSC-TG2, respectively, an 80.74%, 98.18% accuracy. We plotted the training process curves of PCSANet and VGG-PCSA, and as can be seen from Fig. 7, compared to Baseline's VGG and ResNet, our model is faster to train, more accurate, and less likely to enter overfitting in the late stage of training.

In terms of visualization of the model, we used the Grad-CAM algorithm to plot the class activation images of the model under different use of attention mechanisms, and the results are shown in Fig. 8. PCSA enables the model to pay more precise attention to objects than other attention mechanisms. In addition, as shown in Fig. 9, we visualize the feature maps of the first stage in PCSANet, and the results show that PCSA can reduce feature redundancy and obtain information at more scales.

F. PARAMETER ANALYSIS

To find the most suitable convolutional kernel size for SA to focus on spatial information, we conducted experiments on NaSC-TG2 as in Table 5. Experimental results show that smaller convolutional kernels have a smaller field of perception and can better perceive local features, thus extracting a more discriminative feature representation.

G. ANALYSIS OF MODEL COMPLEXITY

In this section, we will analyze the complexity of different models by two metrics, which are parameters and FLOPs. The input image size of VGG-PSA and VGG-PCSA in Table 6 is 64 × 64, and that of the rest of the models is 224 × 224. VGG-EPSA is the substitution of PCSA in VGG-PCSA with EPSA. It is evident that the parameter count of VGG-PCSA amounts

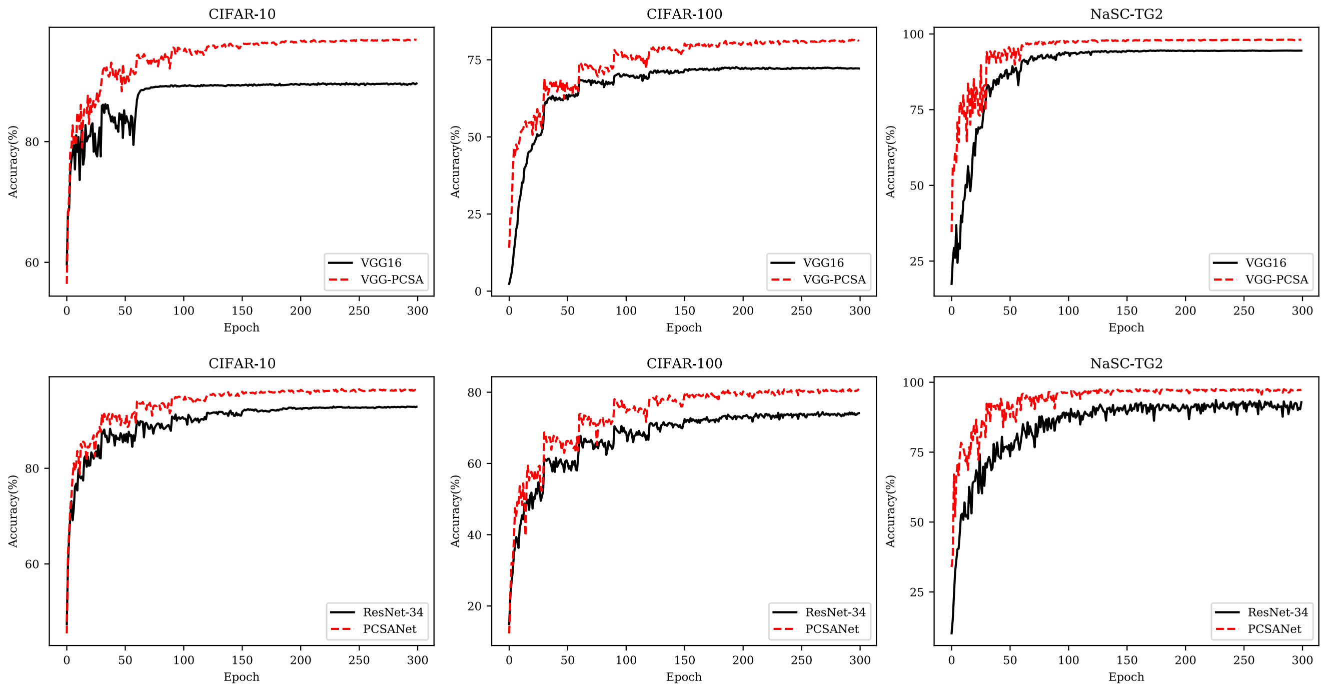


FIGURE 7. Training process curves for PCSANet and VGG-PCSA.

TABLE 4. Accuracy (%) of the different methods on CIFAR-10, CIFAR-100, NaSC-TG2 datasets.

Methods	Top1-Accuracy(%)		
	CIFAR-10	CIFAR-100	NaSC-TG2
AlexNet [29]	87.14	66.78	84.58
GoogleNet [30]	95.05	72.90	87.76
Inception-V3 [31]	91.99	79.05	86.75
DenseNet-40 [32]	94.76	75.58	90.67
GRSN [33]	90.64	63.18	-
MobileNetV3 [34]	94.60	77.70	91.39
PreActResNet18 [35]	96.03	78.49	-
ResNet34*	92.91	74.45	88.37
ResNet+SE*	96.17	78.98	97.41
ResNet+CBAM *	95.57	78.60	97.62
ResNet+ECA*	95.61	79.02	97.43
ResNet+SA [36]	95.20	76.93	-
ResNet+SPConv [37]	95.32	79.23	-
ResNet+TiedConv [38]	95.44	79.52	-
ResNet+SCConv [2]	95.92	79.89	-
EPSANet*	96.24	80.32	97.68
PCSANet(Ours)*	96.43	80.49	97.79
VGG16*	83.29	71.47	94.17
VGG+BN [39]*	94.60	73.17	95.69
VGG+WA [25]	93.89	73.66	-
VGG-SA [28]	-	-	96.68
VGG-SE*	95.73	80.31	97.02
VGG-CBAM*	95.97	81.36	96.40
VGG-ECA*	95.61	80.44	97.38
VGG-EPSA*	96.29	80.57	97.57
VGG-PCSA(Ours)*	**96.50**	**80.74**	**98.18**

The model marked with * is trained in this paper.

to 13.4% of that in VGG16. Additionally, the parameter count of VGG-PCSA is 3.3% lower compared to VGG-EPSA. Notably, PCSANet comprises 4.72M parameters, representing 22% of the parameter count in ResNet. The corresponding FLOPs for PCSANet stand at 0.88G, equating to 24% of the FLOPs seen in ResNet. Consequently, we can draw the conclusion that PCSA effectively enhances model performance while utilizing a reduced parameter count.

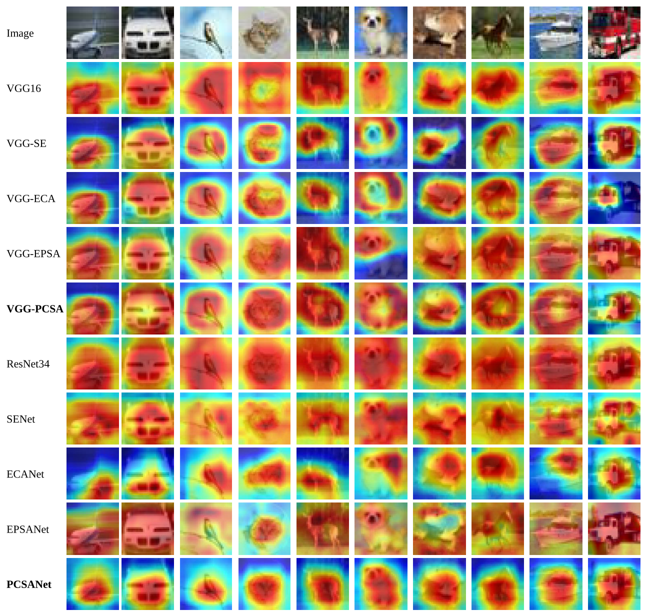


FIGURE 8. Class activation map drawn on CIFAR-10 using the last layer of the network. Darker colors indicate higher model attention. Our method has more bright regions in the heat map than other algorithms and pays more attention to the edge properties of the object, indicating that our method pays attention to more critical details and spatial information.

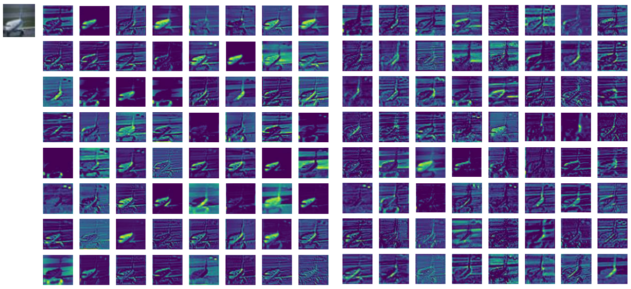


FIGURE 9. Left: Features from the first-stage of original ResNet34, Right: Features from the first-stage of PCSANet.

TABLE 5. Effect of using different convolution kernel sizes in SA on PCSA.

Methods	Kernel Size	Top1-Accuracy(%)
VGG-PCSA	3	98.18
VGG-PCSA	5	98.02
VGG-PCSA	7	97.89

TABLE 6. Analysis of model complexity.

Models	Parameter(M)	FLOPs(G)
VGG16	138.36	15.47
VGG-EPSA	19.28	2.07
VGG-PCSA	18.63	1.94
ResNet34	21.3	3.68
EPSANet	7.78	1.40
PCSANet	4.72	0.88

V. CONCLUSION

In this paper, we propose a new attention mechanism, PCSA, which uses improved pyramid multiscale convolution for feature extraction, extracted channels, and spatial attention weights to recombine the multiscale features to fully extract the spatial information of the image. We use dilated convolution to replace traditional convolution based on pyramidal group convolution to reduce computational cost, and point-wise convolution to mitigate the information loss caused by group convolution. In addition, PCSA extracts channel and spatial attention weights to recombine the features after multi-scale convolution. As a plug-and-play generalized module, PCSA can be directly used in model architecture. We apply PCSA to two backbone networks, ResNet and VGG, and redesign their structures, and the two new network models are PCSANet and VGG-PCSA, respectively.In order to validate the model effect, we conduct a large number of comparative experiments and ablation experiments on three datasets, namely, CIFAR-10, CIFAR-100, and NaSC-TG2. The results show that the network structure embedded with PCSA is more accurate than some art-of-state image classification methods. In subsequent studies, we will continue to apply PCSA in the backbone network for computer vision tasks such as target detection and image segmentation to test its effect.

REFERENCES

[1] Q. Wang, P. Qin, Y. Zhang, X. Wei, and M. Gao, "MLAN: Multi-level attention network," *IEEE Access*, vol. 10, pp. 105437–105446, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9905593/

[2] J. Li, Y. Wen, and L. He, "SCConv: Spatial and channel reconstruction convolution for feature redundancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6153–6162.

[3] A. Kumar, P. Shivakumara, P. N. Chowdhury, U. Pal, and C.-L. Liu, "DPAM: A new deep parallel attention model for multiple license plate number recognition," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 1485–1491.

[4] L. Yang, R. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 11863–11874. [Online]. Available: https://proceedings.mlr.press/v139/yang21o.html

[5] Y. Liu, J. Zhou, L. Liu, Z. Zhan, Y. Hu, Y. Fu, and H. Duan, "FCP-Net: A feature-compression-pyramid network guided by game-theoretic interactions for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 6, pp. 1482–1496, Jun. 2022.

[6] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," 2018, *arXiv:1809.02983*.

[7] P. Wu, H. Huang, H. Qian, S. Su, B. Sun, and Z. Zuo, "SRCANet: Stacked residual coordinate attention network for infrared ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[8] Z. Wang, S. Zhang, W. Huang, J. Guo, and L. Zeng, "Sonar image target detection based on adaptive global feature enhancement network," *IEEE Sensors J.*, vol. 22, no. 2, pp. 1509–1530, Jan. 2022.

[9] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017, *arXiv:1709.01507*.

[10] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.

[11] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," 2019, *arXiv:1910.03151*.

[12] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 763–772.

[13] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*.

[14] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," 2019, *arXiv:1903.06586*.

[15] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Pyramidal convolution: Rethinking convolutional neural networks for visual recognition," 2020, *arXiv:2006.11538*. [Online]. Available: https://arxiv.org/abs/2006.11538

[16] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," 2021, *arXiv:2103.02907*.

[17] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "EPSANet: An efficient pyramid squeeze attention block on Convolutional neural network," in *Proc. 16th Asian Conf. Comput. Vis. (ACCV)* in Lecture Notes in Computer Science, vol. 13843. Macao, China: Springer, 2023, pp. 541–557.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[19] X. Jin, X. Du, and H. Sun, "VGG-S: Improved small sample image recognition model based on VGG16," in *Proc. 3rd Int. Conf. Artif. Intell. Adv. Manuf. (AIAM)*, Oct. 2021, pp. 229–232.

[20] J. A. Campos-Leal, A. Yee-Rendón, and I. F. Vega-López, "Simplifying VGG-16 for plant species identification," *IEEE Latin Amer. Trans.*, vol. 20, no. 11, pp. 2330–2338, Nov. 2022.

[21] S.-H. Wang, S. L. Fernandes, Z. Zhu, and Y.-D. Zhang, "AVNC: attention-based VGG-style network for COVID-19 diagnosis by CBAM," *IEEE Sensors J.*, vol. 22, no. 18, pp. 17431–17438, Sep. 2022.

[22] J. Huo, S. Qiao, Q. Qian, and J. Yang, "Flower classification based on improve VGG networks with attention," in *Proc. 3rd Int. Conf. Comput. Vis., Image Deep Learn. Int. Conf. Comput. Eng. Appl. (CVIDL ICCEA)*, Changchun, China, May 2022, pp. 596–600.

[23] Y. Zou, "Facial expression algorithm using attention mechanism," in *Proc. IEEE 4th Int. Conf. Civil Aviation Saf. Inf. Technol. (ICCASIT)*, Oct. 2022, pp. 1437–1440.

[24] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[25] Z. Xiangyu, "Wavelet-attention CNN for image classification," 2022, *arXiv:2201.09271*.

[26] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep., Jan. 2009, vol. 1.

[27] Z. Zhou, S. Li, W. Wu, W. Guo, X. Li, G. Xia, and Z. Zhao, "NaSC-TG2: Natural scene classification with Tiangong-2 remotely sensed imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3228–3242, 2021.

[28] Z. Liu, A. Dong, J. Yu, Y. Han, Y. Zhou, and K. Zhao, "Scene classification for remote sensing images with self-attention augmented CNN," *IET Image Process.*, vol. 16, no. 11, pp. 3085–3096, Sep. 2022, doi: 10.1049/ipr2.12540.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[30] N. Kulkarni, N. Singh, Y. Joshi, N. Hasabi, S. M. Meena, U. Kulkarni, and S. V. Gurlahosur, "Hybrid optimization for DNN model compression and inference acceleration," in *Proc. 2nd Int. Conf. Intell. Technol. (CONIT)*, Jun. 2022, pp. 1–8.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015, *arXiv:1512.00567*.

[32] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.

[33] X. Huang, "Improved model based on GoogLeNet and residual neural network ResNet," *Int. J. Cognit. Informat. Natural Intell.*, vol. 16, no. 1, pp. 1–19, Nov. 2022. [Online]. Available: http://www-igi–global-com-s.webvpn.cug.edu.cn:8118/article/improved-model-based-on-googlenet-and-residual-neural-network-resnet/313442

[34] D. Haase and M. Amthor, "Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved MobileNets," 2020, *arXiv:2003.13549*.

[35] Y. Zheng, R. Zhang, and Y. Mao, "Regularizing neural networks via adversarial model perturbation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 8152–8161. [Online]. Available: https://ieeexplore.ieee.org/document/9577273/

[36] S. Kundu and S. Sundaresan, "AttentionLite: Towards efficient self-attention models for vision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2225–2229.

[37] Q. Zhang, Z. Jiang, Q. Lu, J. Han, Z. Zeng, S.-h. Gao, and A. Men, "Split to be slim: An overlooked redundancy in vanilla convolution," 2020, *arXiv:2006.12085*.

[38] X. Wang and S. X. Yu, "Tied block convolution: Leaner and better CNNs with shared thinner filters," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, May 2021, pp. 10227–10235. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17226

[39] L. Tang, "Image classification based on improved VGG network," in *Proc. IEEE 6th Int. Conf. Signal Image Process. (ICSIP)*. Nanjing, China: Institute of Electrical and Electronics Engineers, Oct. 2021, pp. 316–320.

YEHANG ZHANG was born in Daqing, Heilongjiang, China, in 2002. He received the B.E. degree in computer science and technology from China University of Geosciences, Wuhan, in 2023. He is applying for a master's degree in computer science and artificial intelligence. His research interests include artificial intelligence, computer vision, and reinforcement learning.