Engineering Applications of Artificial Intelligence 133 (2024) 108261

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Research paper

Multi-scale spatial pyramid attention mechanism for image recognition: An effective approach

Yang Yu, Yi Zhang, Zeyu Cheng, Zhe Song, Chengkai Tang [*]

School of Electronics and Information, Northwestern Polytechnical University, Xi'an, 710072, Shaanxi, China

ARTICLE INFO

Keywords:
Attention mechanisms
Multi-scale spatial information
Structural information
Long-range channel dependencies
Multi-scale spatial pyramid attention

ABSTRACT

Attention mechanisms have gradually become necessary to enhance the representational power of convolutional neural networks (CNNs). Despite recent progress in attention mechanism research, some open problems still exist. Most existing methods ignore modeling multi-scale feature representations, structural information, and long-range channel dependencies, which are essential for delivering more discriminative attention maps. This study proposes a novel, low-overhead, high-performance attention mechanism with strong generalization ability for various networks and datasets. This mechanism is called Multi-Scale Spatial Pyramid Attention (MSPA) and can be used to solve the limitations of other attention methods. For the critical components of MSPA, we not only develop the Hierarchical-Phantom Convolution (HPC) module, which can extract multi-scale spatial information at a more granular level utilizing hierarchical residual-like connections, but also design the Spatial Pyramid Recalibration (SPR) module, which can integrate structural regularization and structural information in an adaptive combination mechanism, while employing the Softmax operation to build long-range channel dependencies. The proposed MSPA is a powerful tool that can be conveniently embedded into various CNNs as a plug-and-play component. Correspondingly, using MSPA to replace the 3 × 3 convolution in the bottleneck residual blocks of ResNets, we created a series of simple and efficient backbones named MSPANet, which naturally inherit the advantages of MSPA. Without bells and whistles, our method substantially outperforms other state-of-the-art counterparts in all evaluation metrics based on extensive experimental results from CIFAR-100 and ImageNet-1K image recognition. When applying MSPA to ResNet-50, our model achieves top-1 classification accuracy of 81.74% and 78.40% on the CIFAR-100 and ImageNet-1K benchmarks, exceeding the corresponding baselines by 3.95% and 2.27%, respectively. We also obtained promising performance improvements of 1.15% and 0.91% compared to the competitive EPSANet-50. In addition, empirical research results in autonomous driving engineering applications also demonstrate that our method can significantly improve the accuracy and real-time performance of image recognition with cheaper overhead. Our code is publicly available at https://github.com/ndsclark/MSPANet.

1. Introduction

In recent years, convolutional neural networks (CNNs) have become incredibly popular in computer vision research and applications, thanks partly to the success of models like AlexNet (Krizhevsky et al., 2012). The evolution of CNNs has made significant progress in various complex visual recognition tasks, such as image recognition (Yu et al., 2023c; Liu et al., 2022a), object detection (Wan et al., 2023; Wang and Wang, 2023), and semantic segmentation (Zhou et al., 2022; Zhang et al., 2022a). Various studies have devoted enormous time and energy to the three fundamental factors of depth, width, and cardinality to advance the development of CNNs in the past decade, leading to the remarkable growth of many excellent models that meet the requirements of visual tasks. However, studies have also claimed that when the network's depth, width, and cardinality increase a certain threshold, its performance rapidly reaches saturation or degrades. Training these models will waste more computing resources, and more learnable layers will introduce many parameters and floating-point operations (FLOPs), slowing down inference speed. To circumvent this problem, instead of striving to design complex network architectures, recent works have concentrated on another critical factor that has a favorable impact on network performance: attention mechanisms, which are beneficial in diverting the network's attention to essential regions in the image and ignoring irrelevant parts. In the visual recognition system, attention

[*] Corresponding author.
E-mail addresses: yuyang_lark@mail.nwpu.edu.cn (Y. Yu), zhangyi@nwpu.edu.cn (Y. Zhang), czy1102@mail.nwpu.edu.cn (Z. Cheng), hamlet@mail.nwpu.edu.cn (Z. Song), cktang@nwpu.edu.cn (C. Tang).

https://doi.org/10.1016/j.engappai.2024.108261
Received 12 October 2023; Received in revised form 1 March 2024; Accepted 9 March 2024
Available online 5 April 2024
0952-1976/© 2024 Elsevier Ltd. All rights reserved.

mechanisms can be regarded as an adaptive dynamic weight adjustment process based on essential features of the input image, which can inform a CNN model where to pay attention, what to pay attention to, and which to pay attention to, so that the network can achieve better performance with fewer layers.

Recently, attention mechanisms have garnered widespread attention in visual recognition tasks, benefiting from prioritizing task-relevant features while suppressing irrelevant ones. Among these mechanisms, Squeeze-and-Excitation Network (SENet) (Hu et al., 2020) is gaining widespread recognition as one of the most representative attention methods. By computing channel attention for each convolutional block with the help of 2-D global average pooling (GAP), SENet has brought notable performance gains to various CNNs at a considerably low computational cost. Following the SE philosophy, researchers either optimized the output of the squeeze module (*e.g.*, CA (Hou et al., 2021)), simplified the excitation module by reducing complexity (*e.g.*, ECA (Wang et al., 2020b)), or improved both (*e.g.*, SRM (Lee et al., 2019)). Although these methods obtain superior performance compared to their baseline counterparts that do not involve any attention, three challenging issues still need to be addressed. First, how to effectively extract and utilize informative features at multiple scales in feature maps to enrich multi-scale representations. Second, indiscriminately employing GAP to squeeze spatial contextual information can significantly lose structural information in attention learning. Third, channel attention can only encode local feature correlations but fails to establish long-range channel dependencies. In response to the problems above, some efforts have been devoted to multi-scale feature representations, such as HS-ResNet (Yuan et al., 2020), PyConv (Duta et al., 2020), CoConv (Duta et al., 2021), and Res2Net (Gao et al., 2019a). Other works proposed SPANet (Ma et al., 2021) and SPPNet (He et al., 2015b) from the perspective of capturing structural information and channel relationships. Concurrently, long-range channel dependencies were established, as shown in Wang et al. (2018), Fu et al. (2019), Cao et al. (2019). However, all the mentioned methods only alleviate the problem to a certain extent from a particular perspective, bringing higher model complexity and suffering from heavier computational burden. Given this, a pressing challenge is to develop a lightweight and efficient channel attention module capable of modeling multi-scale feature representations, structural information, and long-range channel dependencies in a simple yet powerful manner.

Based on the above discussion, in this paper, we propose a novel, low-overhead yet high-performance attention module with excellent generalization ability for various CNNs and datasets, named Multi-Scale Spatial Pyramid Attention (MSPA). As illustrated in Fig. 1, the proposed MSPA consists of three major components. First, unlike the multi-scale feature extractors implemented by layer-wise operations in CNNs, we employ the proposed Hierarchical-Phantom Convolution (HPC) module to extract spatial information efficiently at different scales from the input feature maps at a more granular level. Specifically, the HPC module comprises three operators: Split, Conv, and Concat. The Split operator means equally splitting the original feature maps into multiple feature map subsets in the channel dimension. The Conv operator refers to different convolutional filter groups connected in a hierarchical residual-like style, operating on feature map subsets to process the input tensor at multiple scales. The Concat operator means concatenating feature map subsets with different scales in the channel dimension. Second, channel-wise attention weights for multi-scale feature maps are learned by leveraging the Spatial Pyramid Recalibration (SPR) module to build cross-dimension interaction. In SPR, we adaptively aggregate global and local feature responses using a spatial pyramid aggregation block, effectively combining structural regularization and structural information. We also learn channel relationships by leveraging two low-cost point-wise convolutional layers. Finally, the attention weights of the corresponding channels are recalibrated using the Softmax operation, establishing long-range dependencies between channels. Correspondingly, the MSPA module is adopted in the bottleneck residual blocks of ResNets (He et al., 2016a) to replace the 3 × 3 convolution while maintaining the integrity of other structures, resulting in a series of novel backbones, namely MSPANet. Following a similar scheme, our MSPA can be seemingly integrated into various CNNs, such as PreActResNet (He et al., 2016b) and ResNeXt (Xie et al., 2017), creating more variants with little effort. To comprehensively evaluate the performance of MSPA, we performed extensive experiments for image recognition on two benchmark datasets, *i.e.*, CIFAR-100 (Krizhevsky et al., 2009) and ImageNet-1K (Russakovsky et al., 2015), using various existing CNNs as backbones. Experimental results show that our MSPANet can consistently outperform other state-of-the-art (SOTA) counterparts with much cheaper computational overhead. For example, on the ImageNet-1K benchmark, for ResNet-50 with 25.557 M parameters and 4.122 GFLOPs, MSPANet-S-50 achieved significant performance gains of 2.05% and 1.23% in Top-1 and Top-5 accuracy, respectively, with 23.769 M parameters and 3.895 GFLOPs. Undoubtedly, similar performance improvements can also be observed on other benchmarks, reflecting that the effectiveness of MSPA is not confined to some specific backbones or datasets. The performance improvements are a testament to its potential for advancing the field of image recognition.

In a nutshell, the main contributions of this paper are summarized as follows:

- A novel, low-overhead, and high-performance MSPA module is proposed, which can effectively extract multi-scale spatial informative features at a more granular level, fully utilize structural regularization and structural information to achieve better feature representations, and efficiently build long-range channel dependencies.
- By replacing the 3 × 3 convolution with the MSPA module in the bottleneck residual blocks of ResNets, a series of novel backbones named MSPANet are proposed, which can not only obtain richer multi-scale feature representations but also adaptively recalibrate channel-wise attention weights.
- Through in-depth analysis and comprehensive ablation experiments, we verify the internal behavior and effectiveness of our method.
- As a flexible, modularized, and scalable attention module, MSPA can be incorporated into a variety of mainstream CNNs with no effort and significantly outperform other SOTA attention methods with much cheaper parameters and FLOPs on multiple benchmarks for image recognition. We also visualize the output of the model using GradCAM++ (Chattopadhay et al., 2018), providing intuitive insights into the effectiveness of our method.

2. Related work

This section briefly overviews the literature relevant to this paper's topic, including representative works on multi-scale feature representations and attention mechanisms.

2.1. Multi-scale feature representations

CNNs can naturally capture coarse-to-fine multi-scale features through a stack of layer-wise convolutional operators, indicating that the ability to extract multi-scale features is crucial for visual recognition tasks. Numerous works have emphasized that correctly incorporating multi-scale feature extraction operators into CNNs can further enhance their feature representations. The InceptionNet family (Szegedy et al., 2015, 2016, 2017) is a successful multi-scale representation architecture in which each path is meticulously configured with customized kernel filters, effectively extracting multi-scale features. Subsequently, PyConvResNet (Duta et al., 2020) exploits pyramidal convolution, where each level contains filters with different sizes and depths, capable of capturing details to varying scales while maintaining similar

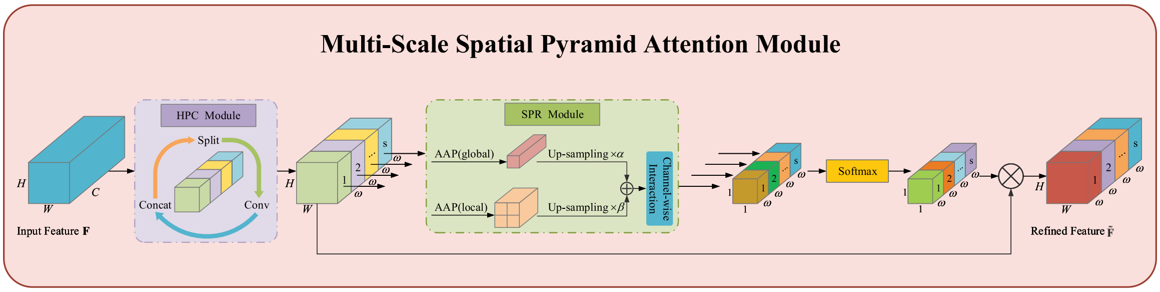


Fig. 1. The overall architecture of the proposed MSPA module. It contains three core components: the HPC module, the SPR module, and the Softmax operation. The HPC module is designed to extract multi-scale spatial information. The SPR module is responsible for learning channel attention weights to build cross-dimension interaction. The Softmax operation is used to recalibrate channel-wise attention weights to establish long-range channel dependencies. Here, the feature maps are shown as feature dimensions, where C, H, and W denote the number of channels, height, and width of a feature map, respectively. ⊕ denotes element-wise summation, and ⊗ denotes element-wise multiplication.

computational overhead as standard convolution. CoConvResNet (Duta et al., 2021) successfully inherits PyConvResNet by replacing the convolution operations in PyConv with dilated convolutions with different dilation ratios, effectively incorporating contextual information at multiple scales. HS-ResNet (Yuan et al., 2020) introduces a novel Hierarchical-Split Block, allowing the network to learn more robust multi-scale feature representations. This simple yet effective solution dramatically enhances the multi-scale representation capability of the network. Res2Net (Gao et al., 2019a) increases the range of receptive fields for each network layer at a more granular level to better capture local and global features. Concurrently, HRNet (Wang et al., 2020a) provides an effective strategy for multi-scale feature representations by concatenating feature maps of different resolutions. Based on the ViT (Dosovitskiy et al., 2020) pipeline, CrossViT (Chen et al., 2021) improves the multi-scale representation ability of the model by employing a dual-branch transformer to merge image patches of different sizes. Most recently, EMCA (Bakr et al., 2022) consolidates multi-scale feature aggregation while learning channel attention by reusing features from preceding attention modules. InceptionNeXt (Yu et al., 2023c) is based on ConvNeXt (Liu et al., 2022a), which decomposes large kernels of depthwise convolution into several groups of small kernels with Inception-style, strengthening feature representations. Additionally, some representative methods have been successively proposed in other critical vision applications, significantly enhancing the feature representation capability of networks, such as medical image classification (Öztürk et al., 2023), text information processing (Çoğalmiş and Bulut, 2022), medical image retrieval (Öztürk et al., 2021), and mineral type detection (Çalışkan, 2023). In short, driven by the importance of multi-scale feature extraction capability, distinct from the mentioned works, we aim to effectively incorporate multi-scale feature representations while learning channel attention.

2.2. Attention mechanisms

Attention mechanisms, owing to their ability to strengthen the allocation of salient feature representations while suppressing insignificant ones, have been plugged into CNNs to improve the information perception ability of models. In previous literature, there are two main ways to incorporate attention mechanisms into CNNs. One is as an independent additional attention module to capture the correlation between features, either across the channel-wise dimension as in Hu et al. (2020), Wang et al. (2020b), Lee et al. (2019), Qin et al. (2021), Yang et al. (2020), Gao et al. (2019b), across the spatial dimension as in Wang et al. (2018), Fu et al. (2019), Cao et al. (2019), Dosovitskiy et al. (2020), Ramachandran et al. (2019), Hu et al. (2018), or a combination of both as in Hou et al. (2021), Woo et al. (2018), Park et al. (2018), Yu et al. (2023a), Wang et al. (2017), Liu et al. (2020), Yang et al. (2021), Zhang and Yang (2021), Misra et al. (2021), Li et al. (2022), Yu et al. (2023b). Specifically, one of the representative examples is SENet (Hu et al., 2020), which is an effective method for successfully implementing channel attention while providing an end-to-end training paradigm for channel attention learning. Inspired by SENet, some methods further advance this idea. SPANet (Ma et al., 2021) improves the output of the squeeze module by introducing spatial pyramid pooling. ECA (Wang et al., 2020b) employs a simple 1-D convolutional layer to capture inter-channel correlations, significantly reducing the excitation module's complexity. SRM (Lee et al., 2019) adaptively recalibrates feature maps by exploiting the style cues of features, improving both the squeeze and excitation modules. Beyond channel attention, spatial attention, an adaptive spatial region selection mechanism, also plays a significant role in inferring attention. NLNet (Wang et al., 2018) generates spatial attention by calculating the correlation between each spatial point in feature maps and captures the long-range dependencies via non-local (NL) operations. Following the philosophy of NLNet, GCNet (Cao et al., 2019) combines the simplified NL block with the SE block to develop a novel NL network, combining contextual representations with channel weighting more effectively. Instead of modeling channel or spatial attention independently, SimAM (Yang et al., 2021) infers 3-D attention weights for feature maps directly based on some well-established neuroscience theories. On the other hand, SA (Zhang and Yang, 2021) exploits shuffle units to effectively combine complementary channel and spatial attention mechanisms. TA (Misra et al., 2021) emphasizes the significance of capturing cross-dimension interaction when inferring attention to generate more informative feature representations. HAM (Li et al., 2022) further improves the channel and spatial attention modules based on CBAM (Woo et al., 2018), achieving promising performance. Another line of research is to use only the attention module to replace specific components of CNNs while keeping other parts intact. SKNet (Li et al., 2019) introduces a new dynamic selection mechanism in which each neuron adaptively adjusts its receptive field size. Subsequently, EPSANet (Zhang et al., 2022b) explores a simple yet effective pyramid squeeze attention module to replace the standard 3 × 3 convolutional layer. Following the hybrid idea of ConvMixer (Trockman and Kolter, 2023) and the advantages of EPSANet, a lightweight EMANet (Yang et al., 2022) is introduced.

In contrast to the prior efforts, we are in pursuit of a low-overhead and high-performance channel attention mechanism, which can extract multi-scale spatial features at a more granular level in a more efficient way, make full use of structural regularization and structural information, and efficiently establish long-range channel dependencies, thus generating more expressive feature representations. Notably, our method performs favorably against other counterparts with fewer parameters and lower computational complexity.

3. Methodology

In this section, we first provide a detailed introduction to the proposed MSPA module and its core components, including the HPC module responsible for extracting multi-scale spatial information, the SPR module for modeling inter-channel relationships, and the Softmax operation for building long-range channel dependencies. Then, we showcase how to apply MSPA to the bottleneck residual blocks of ResNets and propose a series of new backbone networks named MSPANet. Additionally, we present the architecture details of MSPANet-S-50 and MSPANet-B-50 for CIFAR-100 and ImageNet-1K classification, respectively.

3.1. Multi-scale spatial pyramid attention

As discussed in Section 1, this work investigates how to develop a cheap, efficient, and scalable channel attention mechanism. Motivated by the intuition in Gao et al. (2019a), Ma et al. (2021), and Zhang et al. (2022b), we propose an MSPA module that can effectively extract multi-scale spatial information, fully integrate structural regularization and structural information, and efficiently establish long-range channel dependencies. The overall architecture of the MSPA module is shown in Fig. 1. Structurally, the MSPA module is implemented in the following four steps.

Specifically, let an intermediate feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ be the input of the MSPA module, where C, H, and W represent the number of channels, spatial height, and width, respectively. First, $\mathbf{F}$ is passed into our designed HPC module to process the input feature map efficiently at multiple scales. The resulting enhanced multi-scale feature map is expressed as $\hat{\mathbf{F}} = \left[\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2, \dots, \hat{\mathbf{F}}_s\right] \in \mathbb{R}^{C\times H\times W}$. Formally, this process is summarized as follows:

$$\hat{\mathbf{F}} = HPC\left(\mathbf{F}\right) \tag{1}$$

where the $HPC\left(\cdot\right)$ function means the proposed HPC module. Section 3.2 provides the detailed calculation process of the HPC module.

Second, the enhanced feature map $\hat{\mathbf{F}}$ with different scales is fed into the SPR module for efficiently learning channel attention, from which the channel-wise attention weights $\mathbf{V} = \left[\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_s\right] \in \mathbb{R}^{C\times 1\times 1}$ can be successfully inferred. In this process, for the i-th enhanced feature map subset $\hat{\mathbf{F}}_i$ (where $i \in \{1, 2, \dots, s\}$), the corresponding channel-wise attention weights $\mathbf{V}_i$ is represented as follows:

$$\mathbf{V}_i = SPR\left(\hat{\mathbf{F}}_i\right) \tag{2}$$

where the $SPR\left(\cdot\right)$ function refers to the proposed SPR module, and its more detailed introduction can be found in Section 3.3. In order to achieve the interaction of attention information without destroying the original attention weights, the whole channel-wise attention weights are obtained through the following equation:

$$\mathbf{V} = Concat\left(\left[\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_s\right]\right) \tag{3}$$

where the $Concat\left(\cdot\right)$ function represents the concatenation operation along the channel dimension.

Third, soft attention across channels allows for the adaptive selection of different spatial scales while facilitating the interaction between local and global channel attention. Therefore, the recalibrated channel-wise attention weights $\mathcal{A} \in \mathbb{R}^{C\times 1\times 1}$ is generated by applying the Softmax function to the channel-wise attention weights $\mathbf{V}$. Specifically, for $\mathbf{V}_i$, the corresponding $\mathcal{A}_i$ is formulated as follows:

$$\mathcal{A}_i = Softmax\left(\mathbf{V}_i\right) = \frac{exp\left(\mathbf{V}_i\right)}{\sum_{i=1}^{s} exp\left(\mathbf{V}_i\right)} \tag{4}$$

By doing this, long-range channel dependencies between different feature map subsets are established. Next, the whole recalibrated channel-wise attention weights can be generated in a concatenation way, as illustrated in Eq. (5).

$$\mathcal{A} = Concat\left(\left[\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_s\right]\right) \tag{5}$$

Finally, for the i-th enhanced feature map subset $\hat{\mathbf{F}}_i$, we multiply the corresponding $\mathcal{A}_i$ by $\hat{\mathbf{F}}_i$ to obtain the refined output feature map $\tilde{\mathbf{F}}_i$. Mathematically, this process is written as follows:

$$\tilde{\mathbf{F}}_i = \mathcal{A}_i \otimes \hat{\mathbf{F}}_i \tag{6}$$

where $\otimes$ refers to the element-wise multiplication operation. Similarly, the whole refined output feature map is computed as follows:

$$\tilde{\mathbf{F}} = Concat\left(\left[\tilde{\mathbf{F}}_1, \tilde{\mathbf{F}}_2, \dots, \tilde{\mathbf{F}}_s\right]\right) \tag{7}$$

Based on the above description, our MSPA module can effectively integrate multi-scale spatial information and cross-channel attention into one building block, which is beneficial for capturing more discriminative features and improving multi-scale representation capability. Subsequently, we comprehensively verified the superiority of our method, and detailed experimental results and analysis can be found in Section 4.

3.2. Hierarchical-phantom convolution module

As discussed above, this section mainly focuses on the following question: What strategies are used to design a simple yet efficient multi-scale feature extraction scheme compatible with channel attention learning? To answer this question, rather than following the idea behind most existing methods, which enhance multi-scale representations in a layer-wise convolutional manner, we attempt to boost the multi-scale representation capability at a more granular level, designing an HPC module. Fig. 2 depicts the proposed HPC module, from which we can intuitively observe that it is implemented by three operators—Split, Conv, and Concat. In the following, we elaborate on this calculation process.

Following the above notations, given $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ as an input feature map, we first evenly split $\mathbf{F}$ into s feature map subsets along the channel dimension via the Split operator, denoted by $\mathbf{F}_i \in \mathbb{R}^{\omega\times H\times W}$ (where $i \in \{1, 2, \dots, s\}$), and each feature subset has the same spatial shape as $\mathbf{F}$ but with ω channels, without loss of generality $C = s \times \omega$. Second, each $\mathbf{F}_i$ has a group of corresponding Conv operators, *i.e.*, 3 × 3 standard convolution + batch normalization, denoted by $\mathcal{T}_i\left(\cdot\right)$, and the enhanced output feature subset of $\mathcal{T}_i\left(\cdot\right)$ is denoted by $\hat{\mathbf{F}}_i \in \mathbb{R}^{\omega\times H\times W}$. The most innovative idea is that different groups of Conv operators are connected in a hierarchical residual-like style to increase the number of scales the output features can represent. Structurally, the first group of Conv operators extracts features from $\mathbf{F}_1$, from which $\hat{\mathbf{F}}_1$ is produced. Next, the feature map subset $\mathbf{F}_i$ (where $1 < i \leq s$) is added with the output of $\mathcal{T}_{i-1}\left(\cdot\right)$, and then fed into $\mathcal{T}_i\left(\cdot\right)$, from which $\hat{\mathbf{F}}_i$ is obtained. This process repeats several times until all input feature splits are processed. Mathematically, this calculation process is expressed as follows:

$$\hat{\mathbf{F}}_i = \begin{cases} \mathcal{T}_i\left(\mathbf{F}_i\right), & i = 1 \\ \mathcal{T}_i\left(\mathbf{F}_i \oplus \hat{\mathbf{F}}_{i-1}\right), & 1 < i \leq s \end{cases} \tag{8}$$

where $\oplus$ stands for the element-wise summation operation. Finally, the whole enhanced multi-scale feature map $\hat{\mathbf{F}} \in \mathbb{R}^{C\times H\times W}$ can be obtained in a concatenation way, as shown in the following function:

$$\hat{\mathbf{F}} = Concat\left(\left[\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2, \dots, \hat{\mathbf{F}}_s\right]\right) \tag{9}$$

It is worth emphasizing that each set of Conv operators $\mathcal{T}_i\left(\cdot\right)$ can extract feature information from all feature map subsets $\left\{\mathbf{F}_j, j \leq i\right\}$ in the HPC module. Every time $\mathbf{F}_j$ passes through a 3 × 3 convolution operation, the output result has a larger receptive field. As a result, the output of the HPC module encompasses different combinations and different numbers of receptive field scales due to the combinatorial explosion effect. Outputs with smaller receptive fields can capture more details in the scene, which is crucial for recognizing key parts of objects, while outputs with larger receptive fields can focus on larger objects.

In brief, in the HPC module, the input feature maps are processed in a more efficient multi-scale manner, facilitating the extraction of local and global feature information. Meanwhile, we utilize s and ω as control parameters for the number of feature splits and channels to limit our method's parameters and computational overhead. The larger s corresponds to stronger multi-scale feature representation ability, while the larger ω corresponds to richer feature maps. Moreover, we empirically validate the effect of different s and ω on performance. More experimental results are found in Section 4.2.2.

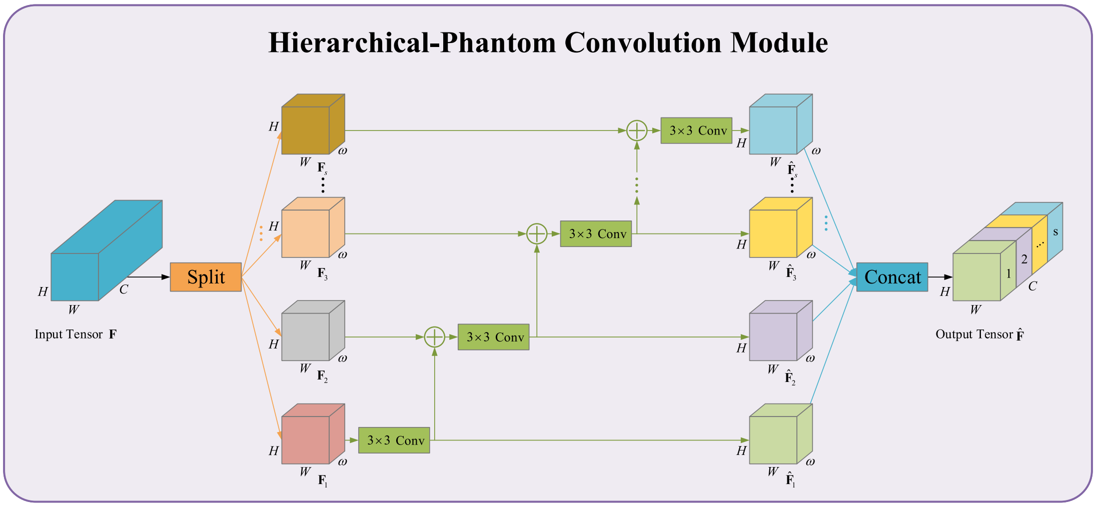


Fig. 2. A detailed illustration of the proposed HPC module, where Split means equally splitting in the channel dimension, Conv represents a 3 × 3 standard convolutional layer followed by batch normalization, and Concat refers to concatenating features in the channel dimension.

3.3. Spatial pyramid recalibration module

From the above description, it is difficult for the proposed HPC module to model channel relationships. Explicitly building channel interdependencies is advantageous for enhancing the model's sensitivity to informative channels, thus making a more substantial contribution to the final decision process. Therefore, extracting information from the multi-scale enhanced feature maps outputted by the HPC module to learn channel relationships is particularly important. In general, channel attention learning can be roughly divided into two steps: feature aggregation, which is responsible for global information embedding, and feature transformation, which captures inter-channel correlations. Numerous attention-based works (Hu et al., 2020; Wang et al., 2020b; Yang et al., 2021; Zhang and Yang, 2021) have shown that GAP is a simple and effective method for aggregating global feature responses. Although GAP can achieve global information embedding, aggregating 3-D feature maps into 1-D channel descriptors inevitably results in a loss of structural information in the original feature maps. At the same time, GAP behaves similarly to a structural regularizer (Ma et al., 2021). Unthinkingly applying GAP to every feature map will lead to overemphasizing the effect of structural regularization while ignoring the detailed feature representation and structural information, especially when a feature map is large. To alleviate this issue, we creatively incorporate structural information in channel attention learning to develop an SPR module. Fig. 3 illustrates the SPR module architecture. As depicted in Fig. 3, the SPR module can be decomposed into two main components: one is the Spatial Pyramid Aggregation (SPA) block, which aggregates global and local contextual features in an adaptive fusion manner. The second is the Channel-wise Interaction (CI) block, which exploits a combination of two point-wise convolutional layers and a Sigmoid activation function to extract inter-channel relationships. Both components are designed to be lightweight. Next, we describe the design of each component in the SPR module in detail.

First, we describe the SPA block in detail. Based on the above discussion, instead of exploiting GAP directly to aggregate feature responses, we utilize the average pooling of two sizes in SPA to aggregate global and local contexts, aiming to achieve structural regularization and explore structural information simultaneously. Specifically, our SPA adaptively pools an input feature map into two scale channel descriptors. The first is to use the traditional GAP (*i.e.*, 1 × 1 average pooling) with strong structural regularization to obtain the global channel descriptor. The second is the local channel descriptor generated using local average pooling (LAP) (*i.e.*, 2 × 2 average pooling) to capture richer feature representation and structural information. These channel descriptors are then upsampled to the spatial shape of the local channel descriptor and fused in a weighted summation manner. Subsequently, we rescale the output into a 1-D vector as the final channel descriptor. Therefore, based on the above notations, taking the enhanced feature map subset $\hat{\mathbf{F}}_i \in \mathbb{R}^{\omega \times H \times W}$ as input, the corresponding channel descriptor $\mathbf{Z}_i \in \mathbb{R}^{4\omega \times 1 \times 1}$ produced by the SPA module is expressed as follows:

$$\mathbf{Z}_i = SPA\left(\hat{\mathbf{F}}_i\right) = \psi_{re}\left(\alpha \otimes T_{up}\left(P\left(\hat{\mathbf{F}}_i, 1\right)\right) \oplus \beta \otimes T_{up}\left(P\left(\hat{\mathbf{F}}_i, 2\right)\right)\right) \quad (10)$$

where the $SPA(\cdot)$ function refers to the SPA block, $P(\cdot,\cdot)$ denotes the adaptive average pooling layer, $T_{up}(\cdot)$ represents the upsampling function, and $\psi_{re}(\cdot)$ serves as resizing a tensor to a vector. Moreover, α and β are two learnable floating-point parameters optimized by stochastic gradient descent (SGD). In short, we effectively integrate structural regularization and structural information in SPA by adaptively combining channel descriptors of two different scales, significantly improving feature representations. Subsequently, we verify the impact of channel descriptors of different scales on performance, thereby confirming the effectiveness of our scheme. We present the details in Section 4.2.3.

Next, we introduce the computation flow of the CI block. It is worth noting that the channel descriptor $\mathbf{Z}_i$ generated by the SPA block cannot be directly used to learn inter-channel relationships. To address this issue, following the pipeline of the excitation module in Hu et al. (2020), we utilize a sequence consisting of two point-wise convolutional layers and a Sigmoid function to encode $\mathbf{Z}_i$, resulting in corresponding channel-wise attention weights $\mathbf{V}_i$. Mathematically, this process is formulated as follows:

$$\mathbf{V}_i = CI\left(\mathbf{Z}_i\right) = \sigma\left(f_2^{1\times1}\left(ReLU\left(f_1^{1\times1}\left(\mathbf{Z}_i\right)\right)\right)\right) \quad (11)$$

where the $CI(\cdot)$ function is our CI block, $ReLU(\cdot)$ refers to the Rectified Linear Unit, and $\sigma(\cdot)$ is the Sigmoid activation function responsible for mapping the output to a range of $(0, 1)$. $f_1^{1\times1}(\cdot)$ and $f_2^{1\times1}(\cdot)$ denote 1 × 1 convolution operation with parameter matrices $(4\omega, \omega/r)$ and $(\omega/r, \omega)$, respectively, where r represents the reduction ratio, which is mainly used to control computational overhead through dimensionality reduction. We empirically set r to 16 throughout all the experiments.

3.4. Instantiation and network configuration

We mainly take the bottleneck residual block in ResNets as an example to introduce in detail how to apply the proposed MSPA module to demonstrate better our method's superiority over other SOTA counterparts. Fig. 4 visually depicts the exact location of our MSPA when it is incorporated into the bottleneck residual block. As shown

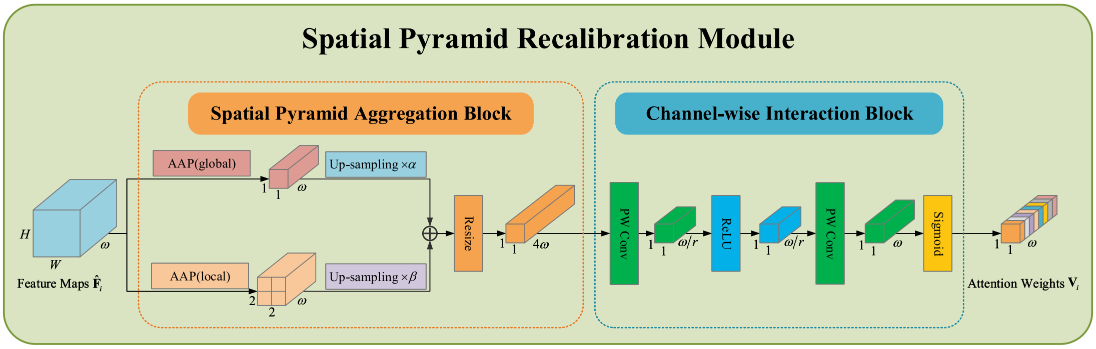
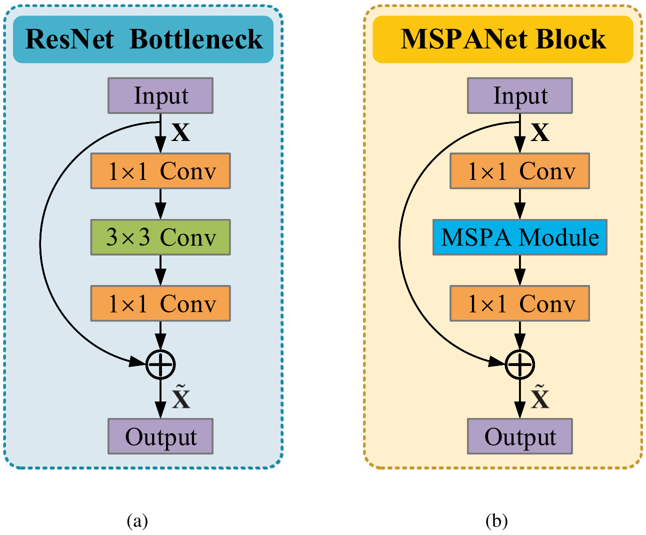

Spatial Pyramid Recalibration Module
Spatial Pyramid Aggregation Block
Channel-wise Interaction Block
H
W
ω
Feature Maps $\hat{\mathbf{F}}_i$
AAP(global)
1
1
ω
Up-sampling ×α
AAP(local)
2
2
ω
Up-sampling ×β
Resize
1
1
4ω
PW Conv
ω/r
ReLU
ω/r
PW Conv
ω
Sigmoid
1
1
ω
Attention Weights $\mathbf{V}_i$

Fig. 3. Diagram of the proposed SPR module. It comprises two essential components, *i.e.*, spatial pyramid aggregation block and channel-wise interaction block. The spatial pyramid aggregation block utilizes a pyramid-like 2-layer adaptive average pooling of different sizes to combine structural regularization and structural information in the attention path. The channel-wise interaction block learns attention maps from the output of the spatial pyramid aggregation structure. Here, AAP () represents the adaptive average pooling, Up-sampling stands for upsampling using nearest neighbor interpolation, and PW Conv refers to a point-wise convolution.



(a)

(b)

Fig. 4. Comparison between the original bottleneck residual block (a) and the basic building block of the proposed MSPANet (b).

in Fig. 4, the basic building block of MSPANet can be obtained by replacing the 3 × 3 convolutional layer with the MSPA module in the bottleneck residual block of ResNets while keeping other configurations unchanged. By stacking such basic building blocks as the ResNets style, a series of novel and efficient backbones called MSPANet are developed. Remarkably, the proposed MSPANet networks inherit the advantages of the MSPA module, making them powerful in multi-scale feature representations and capable of adaptively recalibrating cross-dimension channel weights. Following similar schemes, our MSPA can also be easily integrated into various well-established CNNs, *e.g.*, PreActResNet (He et al., 2016b) and ResNeXt (Xie et al., 2017), to build more variants. For concrete examples of MSPANet architectures, two variants, MSPANet-S (Small $28\omega \times 3s$) and MSPANet-B (Base $30\omega \times 3s$), are proposed by controlling the ω and s parameters of MSPA in the basic building blocks to evaluate the effectiveness of our design scheme comprehensively. Table 1 presents the detailed configurations of MSPANet-S-50 and MSPANet-B-50 for CIFAR-100 and ImageNet-1K datasets.

4. Experiments

In this section, we conduct an extensive and compelling series of experiments to evaluate the performance and practical benefits of the proposed MSPA module for image recognition across a range of datasets and model architectures. We first elaborate on the characteristics of the datasets and the implementation details of all experiments. Second, we perform comprehensive ablation experiments to thoroughly investigate the contribution of each component in the proposed MSPA and further verify the correctness and effectiveness of our design scheme. Third, our MSPA module can be seamlessly implanted into various mainstream network architectures and performs favorably against other SOTA competitors on different benchmarks, demonstrating our approach's lightweight, effectiveness, and generalization. Moreover, we provide Grad-CAM++ (Chattopadhay et al., 2018) visualization results for several sample images from the ImageNet-1K validation set, intuitively showcasing the ability of our method to capture more discriminative feature-rich representations. At last, we perform empirical studies on an engineering-related campus scene dataset to rigorously evaluate the practical value of our approach in specific engineering applications.

4.1. Datasets and implementation details

We perform all experiments on the CIFAR-100 (Krizhevsky et al., 2009) and ImageNet-1K (Russakovsky et al., 2015) benchmarks, utilizing various well-established CNNs as backbone models to comprehensively evaluate the performance of the proposed MSPA module. These datasets are publicly available. The CIFAR-100 dataset comprises 60K color images of 32 × 32 pixels, of which 50K images are for training and 10K for testing. This dataset has 100 classes, each containing 600 images, including 500 training images and 100 testing images. Moreover, the dataset is divided into 5 training batches and 1 test batch, each with 10K images. The test batch includes 100 randomly selected images from each class, whereas the training batches contain the remaining images in random order. For all experiments performed on CIFAR-100, we follow the training practices as (He et al., 2016a,b; Xie et al., 2017) for data augmentation and optimization to facilitate practical comparative analysis between models. Concretely, for data augmentation, each image is zero-padded with 4 pixels on each side, then randomly cropped to the original size (*i.e.*, 32 × 32 pixels). Afterward, half of the resulting images are horizontally flipped at random. The input images are then normalized using channels' means and standard deviations. We employ a synchronous SGD optimizer with a weight decay of 5e-4, momentum of 0.9, and mini-batch size of 128 for optimization. The initial learning rate is set to 0.1 and is divided by 5 at the 60th, 120th, and 160th epochs. All models are trained from scratch for 200 epochs on two RTX A6000 GPUs. During evaluation, all models only accept the original images. Differently, ImageNet-1K is a large-scale labeled dataset organized according to the WordNet hierarchy and serves as a benchmark dataset for image recognition. This dataset contains 1.28 million training images and 50K validation images belonging to 1000 categories. The training set consists of a

Table 1

Architecture details of the original ResNet-50 (left), the proposed MSPANet-S-50 (middle), and MSPANet-B-50 (right) for ImageNet-1K and CIFAR-100. The configuration shown in the first row of the starting stage is for ImageNet-1K, while the second row is for CIFAR-100. Shapes and operations with specific parameter settings of a building block are listed inside the brackets. Moreover, the number of stacked building blocks in a stage is presented outside. A batch normalization layer and ReLU activation function follow each convolutional layer. Down-sampling is performed by the first building block of each stage with a stride of 2, except for the first stage. The $k{-}d\ fc$ operation in the ending stage refers to a fully-connected layer with k classes.

Stage	ResNet-50 (64ω)	MSPANet-S-50 ($28\omega \times 3s$)	MSPANet-B-50 ($30\omega \times 3s$)	Output	
				ImageNet-1K	CIFAR-100
starting	(7 × 7, 64, stride 2), (3 × 3, max pool, stride 2)			56 × 56	–
	3 × 3, 64, stride 1			–	32 × 32
1	$\begin{bmatrix}\text{conv}, 1\times1, 64\\ \text{conv}, 3\times3, 64\\ \text{conv}, 1\times1, 256\end{bmatrix}\times 3$	$\begin{bmatrix}\text{conv}, 1\times1, 84\\ \text{MSPA}, 3\times3, 84\\ \text{conv}, 1\times1, 256\end{bmatrix}\times 3$	$\begin{bmatrix}\text{conv}, 1\times1, 90\\ \text{MSPA}, 3\times3, 90\\ \text{conv}, 1\times1, 256\end{bmatrix}\times 3$	56 × 56	32 × 32
2	$\begin{bmatrix}\text{conv}, 1\times1, 128\\ \text{conv}, 3\times3, 128\\ \text{conv}, 1\times1, 512\end{bmatrix}\times 4$	$\begin{bmatrix}\text{conv}, 1\times1, 168\\ \text{MSPA}, 3\times3, 168\\ \text{conv}, 1\times1, 512\end{bmatrix}\times 4$	$\begin{bmatrix}\text{conv}, 1\times1, 180\\ \text{MSPA}, 3\times3, 180\\ \text{conv}, 1\times1, 512\end{bmatrix}\times 4$	28 × 28	16 × 16
3	$\begin{bmatrix}\text{conv}, 1\times1, 256\\ \text{conv}, 3\times3, 256\\ \text{conv}, 1\times1, 1024\end{bmatrix}\times 6$	$\begin{bmatrix}\text{conv}, 1\times1, 336\\ \text{MSPA}, 3\times3, 336\\ \text{conv}, 1\times1, 1024\end{bmatrix}\times 6$	$\begin{bmatrix}\text{conv}, 1\times1, 360\\ \text{MSPA}, 3\times3, 360\\ \text{conv}, 1\times1, 1024\end{bmatrix}\times 6$	14 × 14	8 × 8
4	$\begin{bmatrix}\text{conv}, 1\times1, 512\\ \text{conv}, 3\times3, 512\\ \text{conv}, 1\times1, 2048\end{bmatrix}\times 3$	$\begin{bmatrix}\text{conv}, 1\times1, 672\\ \text{MSPA}, 3\times3, 672\\ \text{conv}, 1\times1, 2048\end{bmatrix}\times 3$	$\begin{bmatrix}\text{conv}, 1\times1, 720\\ \text{MSPA}, 3\times3, 720\\ \text{conv}, 1\times1, 2048\end{bmatrix}\times 3$	7 × 7	4 × 4
ending	global average pool, $k{-}d\ fc$, softmax			1 × 1	1 × 1

variable number of images for each category, ranging from 732 to 1300, while there are precisely 50 images per category in the validation set. For all experiments on ImageNet-1K, we adopt the same data augmentation and hyper-parameter settings as (He et al., 2016a; Zhang et al., 2022b) for a fair and consistent comparison with other competing methods. Correspondingly, the input images are randomly cropped to 224 × 224 patches, followed by horizontally flipping images with a probability of 0.5. The practical mean channel subtraction is adopted to normalize the input images for training and testing. During training, the network parameters are optimized using synchronous SGD with weight decay of 1e-4, momentum of 0.9, and mini-batch size of 256. The initial learning rate is set to 0.1 and decreased by a factor of 10 every 30 epochs. All networks are trained from scratch within 100 epochs on four RTX A6000 GPUs. When testing, the shorter side of the input image is first resized to 256, and a single center crop of 224 × 224 patches is completed for evaluation. Without bells and whistles, we use the common weight initialization strategy in He et al. (2015a) and train all networks on the training set utilizing naive softmax cross-entropy without label-smoothing regularization. Moreover, to perform better apple-to-apple comparisons, we re-implement all competitors in the PyTorch framework (Paszke et al., 2019) and report our reproduced results throughout the experiments.

4.2. Ablation studies

In this section, we evaluate the performance of each component of the proposed MSPA module and provide an in-depth understanding of its internal properties on the CIFAR-100 dataset through comprehensive ablation studies. Our ablation experiments are divided into three parts. First, we discuss the importance of the HPC module for extracting multi-scale spatial information, the SPR module for learning inter-channel relationships, and the Softmax operation for modeling long-range channel dependencies, and also verify the effectiveness of our design choice. Second, we explore the impact of configuring different scale and channel dimensions in the HPC module on the feature representation power of MSPA. Finally, we perform an in-depth analysis of how the behavior of MSPA changes with different scales of information embedding in the SPR module. The experimental results and analysis are introduced as follows.

4.2.1. Importance of HPC module, SPR module, and softmax operation

In this part, we employ ResNet-50 as the base model and carry out several ablation experiments on CIFAR-100 to evaluate the practical benefits of each core component in the MSPA module, including the HPC module, SPR module, and Softmax operation, further showing the validity of our design solution. The corresponding experimental results are presented in Table 2. For convenience, the ‘HPC’ means replacing the 3 × 3 convolutional layers with the HPC module in the bottleneck residual blocks of the ResNet-50 while keeping other configurations unchanged. The ‘SPR’ refers to the scale dimension (s) in the HPC module being set to 1, which is equivalent to the effect of removing the HPC module. The ‘HPC + SPR (no Softmax)’ is equipped with both HPC and SPR modules while removing the Softmax operation. The ‘HPC + SPR (Softmax)’ corresponds to our proposed method. We can derive the following findings from the experimental results reported in Table 2. First, both ‘HPC’ and ‘SPR’ show significant improvements in classification accuracy over their respective base models, and compared with the competitive channel attention SE/ECA, they achieve performance gains of 0.91%/0.25% and 1.23%/0.57% in Top-1 Acc, respectively. These results objectively indicate that the HPC and SPR modules are beneficial to improving recognition accuracy. Intuitively, the reason for the above phenomenon is that the HPC module effectively enhances the multi-scale representation capability of the network, while the SPR module efficiently models inter-channel correlations by aggregating more informative channel descriptors. Second, the ‘HPC + SPR (no Softmax)’ further improves classification accuracy, proving that the HPC and SPR modules are compatible and their efficacy is complementary, *i.e.*, embedding multi-scale feature representations in channel attention learning is advantageous. Finally, it is worth emphasizing that our approach attains the best result highlighted in Table 2. In particular, compared with the ‘HPC + SPR (no Softmax)’, our method provides a 0.26% improvement in Top-1 Acc with almost the same overhead (parameters and FLOPs), which reflects that establishing long-range channel dependencies is helpful for the interaction between local and global channel attention, making the network more efficient in attention inference. More importantly, our approach always obtains remarkable performance improvements over the original network and its high-performing ECA counterpart with minimal additional overhead. As a brief conclusion, these empirical results strongly support our design concept and validate the effectiveness of our design solution.

4.2.2. Effect of scale dimension (s) and channel dimension (ω)

To delve into the influence of different scales and channel dimensions in the HPC module on MSPA performance, we utilize ResNet-50

Table 2
The practical benefits are from the HPC module, SPR module, and Softmax operation on the CIFAR-100 dataset when using ResNet-50 as the baseline. Params and FLOPs denote the number of parameters and floating-point operations, respectively, and Top-1 Acc refers to top-1 classification accuracy. All accuracy results are the mean ± standard deviation of 5 runs.

Model	Description	Params (M)	FLOPs (G)	Top-1 Acc (%)
ResNet-50	N/A	23.705	1.308	77.79 ± 0.16
	SE	26.236	1.312	79.76 ± 0.17
	ECA	23.705	1.310	80.42 ± 0.36
MSPANet-B-50	HPC	23.667	1.338	80.67 ± 0.17
	SPR	24.181	1.309	80.99 ± 0.14
	HPC + SPR (no Softmax)	23.769	1.340	81.48 ± 0.08
	HPC + SPR (Softmax)	23.769	1.340	**81.74 ± 0.07**

Table 3
Performance comparisons of MSPANet-50 with different scales on the CIFAR-100 dataset. The control parameters s and ω represent the number of scales and the number of channels (*i.e.*, the width of filters), respectively. All accuracy results are the mean ± standard deviation of 5 runs.

Model	Setting	Params (M)	FLOPs (G)	Top-1 Acc (%)
ResNet-50	N/A	23.705	1.308	77.79 ± 0.16
MSPANet-50	$30\omega \times 2s$	16.884	0.935	80.83 ± 0.14
	$30\omega \times 3s$	23.769	1.340	**81.74 ± 0.07**
	$30\omega \times 4s$	30.653	1.744	81.34 ± 0.12
	$30\omega \times 5s$	37.538	2.149	80.98 ± 0.07

Table 4
Performance comparisons of MSPANet-50 with different channels on the CIFAR-100 dataset. All accuracy results are the mean ± standard deviation of 5 runs.

Model	Setting	Params (M)	FLOPs (G)	Top-1 Acc (%)
ResNet-50	N/A	23.705	1.308	77.79 ± 0.16
MSPANet-50	$28\omega \times 3s$	21.917	1.234	81.47 ± 0.09
	$30\omega \times 3s$	23.769	1.340	81.74 ± 0.07
	$32\omega \times 3s$	25.694	1.449	81.86 ± 0.12
	$34\omega \times 3s$	27.680	1.562	81.95 ± 0.10
	$36\omega \times 3s$	29.739	1.678	**81.99 ± 0.06**

as the baseline and perform extensive ablation experiments on CIFAR-100, the corresponding experimental results of which are all listed in Tables 3 and 4. First, we provide an in-depth discussion of the performance changes induced by different scales. As shown in Table 3, we conduct this ablation experiment by fixing the number of channels to 30 (*i.e.*, ω = 30) and setting s from 2 to 5. From the experimental results, MSPANet-50 with different scales can consistently outstrip the baseline in classification accuracy. Interestingly, the MSPANet-50 with $30\omega \times 2s$ obtains a Top-1 accuracy of 80.83%, exceeding the vanilla ResNet-50 (77.79%) by 3.04%, while the parameters and FLOPs of the model are reduced by 28.77% and 28.52%. These results fully illustrate that strengthening multi-scale representations is crucial to improving the network's performance. Notably, the model overhead increases sharply with the increase of scale. However, its performance does not improve monotonically but reaches the optimal value when $s = 3$. These results validate the rationality of our parameter selection in the HPC module.

Next, we compare and analyze the influence of different channels on model performance. As illustrated in Table 4, we perform this ablation experiment by setting the scale to 3 and adopting five different channels, including $\{28, 30, 32, 34, 36\}$, to observe the performance changes better. From Table 4, the following insights become apparent. It is evident that MSPANet-50 with different channels consistently yields substantial improvements in Top-1 Acc over the baseline. In particular, compared with the vanilla ResNet-50, MSPANet-50 with $28\omega \times 3s$ reduces parameters and FLOPs by 7.54% and 5.66%, respectively, while achieving a 3.68% improvement in Top-1 accuracy. These results suggest that enhancing multi-scale feature representations can improve network performance. However, increasing the number of channels leads to a significant increase in model overhead, while the growth in classification accuracy tends to plateau and even reach saturation. For example, MSPANet-50 with $30\omega \times 3s$ provides a 0.27% improvement in Top-1 Acc over its counterpart with $28\omega{\times}3s$ while only increasing parameters and FLOPs by 1.852 M and 0.106 G, respectively. In comparison, MSPANet-50 with $36\omega{\times}3s$ achieves similar performance gains over its counterpart with $30\omega{\times}3s$, with an improvement of 0.25% in Top-1 accuracy but with an increase of 5.97 M and 0.338 G in parameters and FLOPs, respectively. These empirical findings indicate that the parameter configuration in the HPC module is conducive to achieving a good trade-off between performance and overhead, further verifying the correctness and effectiveness of the HPC module.

In addition, we depict the parameters and Top-1 accuracy curves of MSPANet-50 as s and ω vary in Fig. 5 to visually present the impact of different scales and channel dimensions on model performance, providing more in-depth insights to the research community.

4.2.3. Impact of different scales of information embedding

In this ablation study, we use ResNet-50 as the backbone model and primarily evaluate the impact of different scales of information embedding in the SPR module on the behavior of the proposed MSPA module. The experimental results are given in Table 5, presenting the following observations. First, the combination of different scales significantly impacts model performance, and all of them can surpass the backbone model by a clear margin, which once again reveals the importance of modeling inter-channel correlations. Further, compared to representative channel attention methods like SE and ECA, all models with varying scale combinations yield superior performance, indicating our design philosophy's advancement and effectiveness. Additionally, as the number of scales increases, the overhead of the corresponding model increases continuously, while its performance is not monotonically improved. Specifically, MSPANet-B-50 with adaptive global average pooling (*i.e.*, AAP(1)) obtains the worst result (80.89% Top-1 classification accuracy), while counterparts with other scale combinations can outperform it in classification accuracy by a non-trivial margin. These empirical results illustrate that capturing richer feature representations and structural information can significantly enhance the model's learning capability. Undoubtedly, our method (*i.e.*, the adaptive combination of AAP(1) and AAP(2)) consistently dominates the top performance (81.74% Top-1 classification accuracy), surpassing the remaining counterparts by 0.85% and 0.7% in terms of Top-1 Acc, respectively. These experimental results powerfully demonstrate that our method can fully incorporate structural regularization and structural information and is beneficial in providing a favorable balance between performance and overhead.

4.3. Classification performance on CIFAR-100

To comprehensively evaluate the proposed MSPA module's superior performance, we conduct a series of experiments through in-depth comparisons with several SOTA methods on the small-scale CIFAR-100 dataset employing three widely used CNNs as backbone models, including ResNet-50, ResNeXt-29, and PreActResNet-164. We always

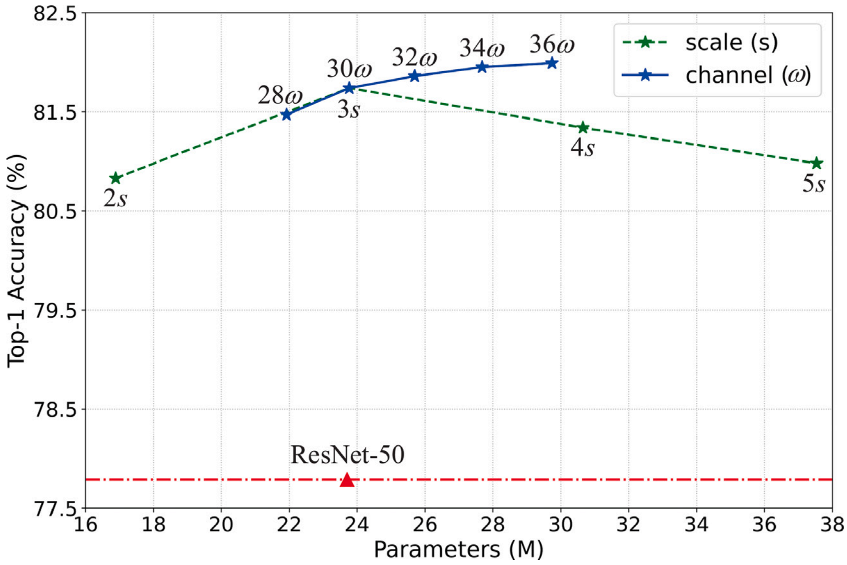


Fig. 5. Comparison of the performance of MSPANet-50 with the change of s and ω on CIFAR-100 classification. Here, we also plot the results of ResNet-50 as a benchmark. It is best viewed in color.

Table 5
Effect of configuring different scales of average pooling in MSPANet-B-50 on CIFAR-100 classification performance. Here, AAP () is the adaptive average pooling. All accuracy results are the mean ± standard deviation of 5 runs.

Model	Description	Params (M)	FLOPs (G)	Top-1 Acc (%)
ResNet-50	N/A	23.705	1.308	77.79 ± 0.16
	SE	26.236	1.312	79.76 ± 0.17
	ECA	23.705	1.310	80.42 ± 0.36
MSPANet-B-50	AAP (1)	23.702	1.339	80.89 ± 0.12
	AAP (1) & AAP (2)	23.769	1.340	**81.74** ± **0.07**
	AAP (1) & AAP (2) & AAP (4)	24.035	1.341	81.04 ± 0.09

follow the training policy specified in Section 4.1 and the network specifications presented in Table 1 for a fair and consistent comparison with competitors. The evaluation metrics we measure are the number of parameters and FLOPs for comparing model efficiency and the Top-1 accuracy for evaluating model performance. All experimental results are recorded in Table 6. As can be seen, no matter which backbone is selected, our proposed method can consistently improve the performance of the corresponding backbone model by a clear margin at cheaper or similar model complexity (*i.e.*, parameters and FLOPs). These results demonstrate that our method is low-overhead and high-performance and has strong robustness and generalization ability for various network architectures. It is worth emphasizing that we report the average accuracy and standard deviation of 5 runs for each model to compare the significant differences between different models. Regardless of which backbone is considered, our method always has the optimal standard deviation compared to competitors. These comparisons reflect that our method has good stability, benefiting the network's performance to converge to the same decision boundary better.

Remarkably, when using ResNet-50 as the backbone, our method consistently achieves the best accuracy compared to other SOTA attention methods while benefiting from lower or comparable model complexity. For example, MSPANet-S-50 outperforms competitive SA by 0.3% in Top-1 Acc while saving 7.55% parameters and 5.95% FLOPs. Compared to the self-attention DA capable of capturing long-range contextual information, MSPANet-B-50 achieves a performance gain of 0.45% in Top-1 Acc with 22.51% fewer parameters and 18.49% lower FLOPs. These comparisons unequivocally reveal the efficacy and importance of the HPC module for enhancing multi-scale feature representations, the SPR module responsible for building inter-channel correlations, and the Softmax operation capable of modeling long-range channel dependencies. Similarly, compared to competitors focusing on multi-scale representations, including SKNet, EPSANet, and EMANet, our method still dominates the best performance in accuracy. For instance, the proposed MSPANet-S/B-50 can attain a margin of 0.24%/0.51% higher over the SKNet-50 in Top-1 accuracy while using 14.48%/7.25% fewer parameters and requiring 14.13%/6.75% lower FLOPs. These results illustrate that our method can improve the multi-scale representation capability of the network in an extremely efficient manner. In particular, our MSPANet-S-50 has an improvement of 0.88% in terms of Top-1 Acc over EPSANet-50 (Small) with only an increase of 1.2 M parameters and 0.086 GFLOPs. We believe that such minimal additional overhead is justified by significant improvements in model performance.

When using ResNeXt-29 as the baseline, our method achieves significant performance gains over the ResNeXt-29 ($8c \times 64\omega$) with 1.27% and 1.5% in terms of Top-1 Acc, respectively, while benefiting from cheaper model complexity. Strikingly, our proposed MSPANeXt-29 ($8c \times 30\omega \times 3s$) surpasses ResNeXt-29 ($16c \times 64\omega$) with more cardinality by 0.38% while only having about half the model complexity. Further, we also observe similar performance trends with ResNet-50 compared to other prominent attention methods. For example, our MSPANeXt-29 ($6c \times 30\omega \times 3s$) has a performance gain of 0.3% in Top-1 Acc over the SimAM-integrated counterpart while saving 24.61% parameters and 22.61% FLOPs. These comparative experimental results again show that our method is lightweight and more effective in attention inference. It is noteworthy that MSPANeXt-29 ($8c \times 30\omega \times 3s$) with more cardinality can yield a better result than MSPANeXt-29 ($6c \times 30\omega \times 3s$), indicating that the improvements induced by the MSPA module can be used in combination with increasing the base network's cardinality. In addition, similar performance trends to the above experiments can also be found in the experimental results based on PreActResNet-164.

Table 6
Comparisons of various attention methods on the CIFAR-100 test set in terms of network parameters (Parameters), floating-point operations (FLOPs), and Top-1 accuracy (Top-1 Acc), using ResNet-50, ResNeXt-29, and PreActResNet-164 as baselines, respectively. All accuracy results are the mean ± standard deviation of 5 runs.

Architecture	CIFAR-100		
	Parameters (M)	FLOPs (G)	Top-1 Acc (%)
ResNet-50	23.705	1.308	77.79 ± 0.16
ResNet-50 + SE	26.236	1.312	79.76 ± 0.17
ResNet-50 + CBAM	26.238	1.313	80.19 ± 0.31
ResNet-50 + SRM	23.766	1.308	80.77 ± 0.26
ResNet-50 + ECA	23.705	1.310	80.42 ± 0.36
ResNet-50 + TA	23.710	1.335	80.64 ± 0.28
ResNet-50 + SimAM	23.705	1.308	80.86 ± 0.12
ResNet-50 + SA	23.706	1.312	81.17 ± 0.14
ResNet-50 + SPA	31.266	1.319	80.56 ± 0.22
ResNet-50 + GC	26.253	1.312	80.25 ± 0.24
ResNet-50 + DA	30.673	1.644	81.29 ± 0.18
SKNet-50	25.628	1.437	81.23 ± 0.27
EPSANet-50 (Small)	20.710	1.148	80.59 ± 0.23
EMANet-50	21.375	1.207	80.25 ± 0.34
MSPANet-S-50 (Ours)	**21.917**	**1.234**	81.47 ± 0.09
MSPANet-B-50 (Ours)	23.769	1.340	**81.74 ± 0.07**
ResNeXt-29 ($8c \times 64\omega$)	34.519	5.413	81.33 ± 0.13
ResNeXt-29 ($16c \times 64\omega$)	68.248	10.734	82.45 ± 0.11
ResNeXt-29 ($8c \times 64\omega$) + SE	35.041	5.415	81.73 ± 0.11
ResNeXt-29 ($8c \times 64\omega$) + SRM	34.541	5.413	82.20 ± 0.18
ResNeXt-29 ($8c \times 64\omega$) + ECA	34.519	5.414	81.90 ± 0.26
ResNeXt-29 ($8c \times 64\omega$) + TA	34.522	5.428	81.92 ± 0.21
ResNeXt-29 ($8c \times 64\omega$) + SimAM	34.519	5.413	82.30 ± 0.12
MSPANeXt-29 ($6c \times 30\omega \times 3s$) (Ours)	**26.023**	**4.189**	82.60 ± 0.08
MSPANeXt-29 ($8c \times 30\omega \times 3s$) (Ours)	34.434	5.553	**82.83 ± 0.11**
PreActResNet-164	1.726	0.257	77.03 ± 0.24
PreActResNet-164 + SE	1.929	0.259	77.81 ± 0.22
PreActResNet-164 + SRM	1.759	0.257	78.39 ± 0.34
PreActResNet-164 + ECA	1.727	0.259	78.13 ± 0.32
PreActResNet-164 + TA	1.730	0.262	78.22 ± 0.29
PreActResNet-164 + SimAM	1.726	0.257	78.49 ± 0.21
PreActMSPANet-164 ($6\omega \times 3s$) (Ours)	**1.345**	**0.201**	78.89 ± 0.16
PreActMSPANet-164 ($8\omega \times 3s$) (Ours)	1.920	0.287	**79.40 ± 0.12**

In brief, all these empirical results sufficiently verify that our MSPA is low-overhead and high-performance, and its superiority is not confined to some specific networks.

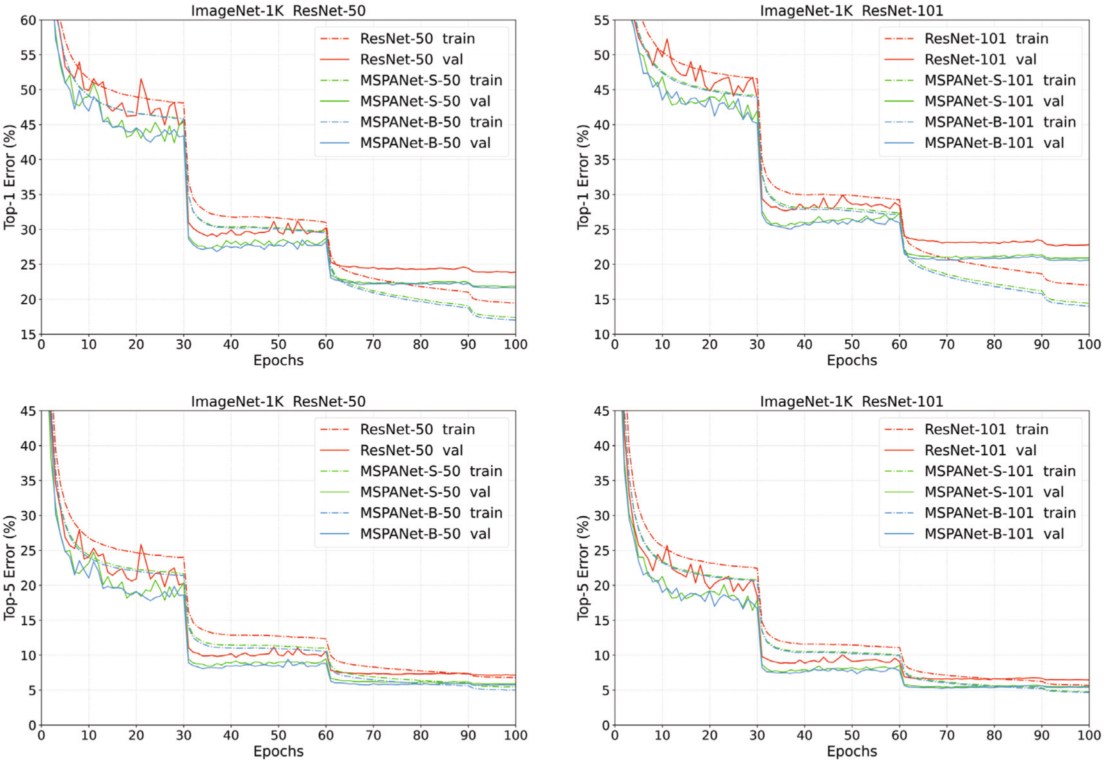
4.4. Classification performance on ImageNet-1K

In this section, we further explore the practical benefits of the proposed MSPA module by comparing it with several advanced methods on the large-scale ImageNet-1K classification utilizing ResNet with 50 and 101 layers as baselines. We rigorously adhered to the network configurations elaborated in Section 3.4 and the training protocols specified in Section 4.1 throughout the experiments. The evaluation metrics include efficiency (*i.e.*, Parameters and FLOPs) and effectiveness (*i.e.*, Top-1/Top-5 accuracy). All results are reported in Table 7, from which we make the following observations.

Our method consistently matches or significantly outperforms the respective baselines across all evaluation metrics. Specifically, for ResNet-50, our MSPANet-S/B-50 achieves 2.05%/2.27% and 1.23%/1.39% performance gains in Top-1 Acc and Top-5 Acc, respectively, while maintaining cheaper or competitive overhead. For ResNet-101, we also observe similar trends in efficiency and effectiveness. Further, our method performs favorably against other SOTA attention methods across all baselines. For example, with almost the same overhead, MSPANet-B-50/101 exceeds the best-performing SA-Net-50/101 by 0.68%/0.55% and 0.45%/0.29% in Top-1 and Top-5 accuracy, respectively. Compared with the EPSANet-50/101 (Small) with the fewest overhead, the corresponding MSPANet-S-50/101 can provide a 0.69%/0.78% and 0.55%/0.4% improvement in Top-1 and Top-5 accuracy, respectively, while introducing only a minimal additional overhead. These comparisons once again validate that our approach is indeed low-overhead and high-performance and further reflect that the improvements induced by MSPA are complementary to those obtained by increasing the depth of the base network. More importantly, when compared with excellent multi-scale representation networks, including Res2Net, PyConvResNet, and CoConvResNet, no matter which backbone is considered, our method can always yield the best results with lower or almost the same overhead. These results demonstrate that our approach is powerful and efficient in improving multi-scale representation capability and has great potential to further boost the performance of CNNs. It is worth mentioning that our MSPANet-S-50 has a Top-1/5 accuracy of 78.18%/94.09%, surpassing the deeper vanilla ResNet-101 (77.37%/93.54%) by 0.81%/0.55% with almost half of the number of parameters (23.769 M vs. 44.549 M) and FLOPs (3.895 G vs. 7.850 G). Such results illustrate that while the MSPA module adds depth, it does so extremely efficiently. Besides, to intuitively illustrate the impact of MSPA on optimizing base networks, we depict the training and validation curves of ResNet with 50 and 101 layers and their corresponding MSPANet-S and MSPANet-B in Fig. 6, respectively. Throughout the training and validation process, the MSPA-incorporated networks consistently perform better than their respective baselines and exhibit improved optimization characteristics and smooth performance gains, indicating that our method has better feature representation capabilities. Our method also shows better convergence speed and stability compared to the baseline. Particularly, our method maintains stable and smooth performance during the 30th to 60th epochs, while the corresponding baseline generates significant fluctuations. Finally, compared with the experimental results recorded in Section 4.3, it can be reflected that the excellent performance of the proposed method is not constrained to CIFAR-100, which further proves that our method possesses strong robustness and generalization ability for different scale datasets.

Table 7
Comparisons of efficiency (*i.e.*, Parameters and FLOPs) and effectiveness (*i.e.*, Top-1/Top-5 Acc) of various attention methods and different multi-scale representation architectures on the ImageNet-1K validation set when taking ResNet with 50 and 101 layers as backbones, respectively. Top-1 Acc and Top-5 Acc stand for top-1 and top-5 classification accuracy, respectively.

Methods	Backbones	ImageNet-1K			
		Parameters (M)	FLOPs (G)	Top-1 Acc (%)	Top-5 Acc (%)
ResNet-50	ResNet-50	25.557	4.122	76.13	92.86
SENet		28.088	4.130	76.71	93.38
CBAM		28.090	4.128	77.34	93.69
SRM		25.617	4.122	77.13	93.51
ECA-Net		25.557	4.128	77.40	93.61
TA-Net		25.562	4.169	77.48	93.68
SimAM		25.557	4.122	77.45	93.66
SA-Net		25.558	4.130	77.72	93.80
SKNet		26.154	4.185	77.55	93.82
EPSANet (Small)		22.562	3.632	77.49	93.54
Res2Net ($26\omega \times 4s$)		25.699	4.293	77.95	93.85
PyConvResNet		24.848	3.883	77.88	93.80
CoConvResNet		25.557	4.122	77.27	93.51
MSPANet-S (Ours)		**23.769**	**3.895**	78.18	94.09
MSPANet-B (Ours)		25.621	4.218	**78.40**	**94.25**
ResNet-101	ResNet-101	44.549	7.850	77.37	93.54
SENet		49.327	7.863	77.62	93.93
CBAM		49.330	7.861	78.49	94.31
SRM		44.679	7.850	78.47	94.20
ECA-Net		44.549	7.859	78.62	94.28
TA-Net		44.559	7.946	78.03	93.85
SimAM		44.549	7.850	78.65	94.11
SA-Net		44.551	7.863	78.95	94.34
SKNet		45.682	7.978	78.84	94.29
EPSANet (Small)		38.900	6.839	78.43	94.11
Res2Net ($26\omega \times 4s$)		45.207	8.122	79.19	94.43
PyConvResNet		42.308	7.314	79.01	94.47
CoConvResNet		44.549	7.850	78.71	94.28
MSPANet-S (Ours)		**41.376**	**7.338**	79.21	94.51
MSPANet-B (Ours)		44.923	7.993	**79.50**	**94.63**



(a) ResNet-50

(b) ResNet-101

Fig. 6. Comparisons of training and validation curves on ImageNet-1K for ResNet, MSPANet-S, and MSPANet-B architectures of 50 and 101 layers, respectively. Dash-dot lines depict training errors, and solid lines indicate validation errors. It is best viewed in color.

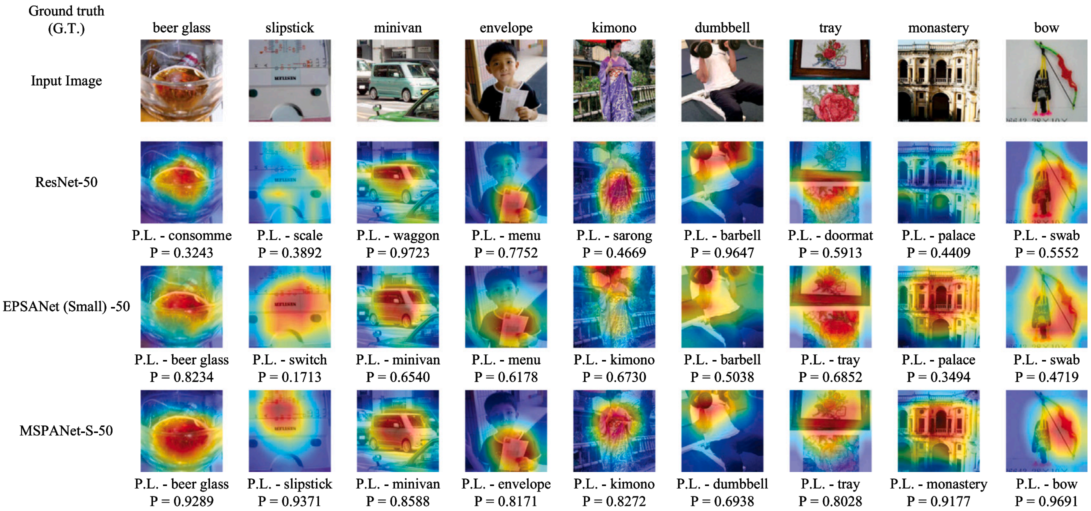


Fig. 7. Visualization of feature maps generated by different models in the last building block. We use GradCAM++ as a visualization tool and compare the visualization results of ResNet-50, EPSANet (Small)-50, and MSPANet-S-50. Ground truth (G.T.) labels for images are provided on the top of the original samples. Prediction labels (P.L.) and prediction scores are provided below the corresponding visualizations. It is best viewed in color.

4.5. Network visualization results with Grad-Cam++

In this section, to intuitively showcase the outstanding performance of the proposed MSPA module and gain insights into its intrinsic behavior and multi-scale representation capability, we utilize the Grad-Cam++ technique to visualize several random samples from the ImageNet-1K validation set. The Grad-Cam++ visualization method is commonly used to localize discriminative regions for image classification. It visualizes the gradients of the top-class prediction concerning the input sample as a colored overlay and interprets how the network makes classification decisions in the form of a heatmap. In this experiment, we compare and analyze the visualization results of the original ResNet-50 with its corresponding EPSANet-50 (Small) and MSPANet-S-50. All results are plotted in Fig. 7, where we gain the following insights.

As shown in Fig. 7, from the comparative visualization results, our method can outperform other competitors by a large margin and more accurately focus on relevant regions of the objects of interest, which indicates that our method is beneficial for capturing more informative and meaningful multi-scale contextual information for a particular target class. Specifically, in columns 3, 4, 5, and 8 in Fig. 7, other competitive methods either make incorrect class predictions or only produce correct labels with lower accuracy. In contrast, our MSPANet-S-50 can predict the correct labels with 85.88%, 81.71%, 82.72%, and 91.77% accuracy, respectively. These results show that our method has a more robust multi-scale representation capability and tends to focus on important regions with more object details. The most significant visualization occurs in column 9, where ResNet-50 and EPSANet-50 (Small) predict the wrong “swab” label with 55.52% and 47.19% probability, respectively, while our method can predict the correct “bow” label with an accuracy of 96.91%. This comparison demonstrates that our method can concentrate on richer and more discriminative features while suppressing irrelevant regions, allowing for better localization of target objects.

As a brief conclusion, these visualization results align perfectly with our design expectations and further validate that the proposed MSPA module is practical and can enhance the feature representation power of the network.

4.6. Applications

Image recognition is an advanced technology that utilizes deep learning algorithms to determine the category of a given image. It is used in various fields, such as medical imaging diagnosis, security and surveillance, autonomous driving, and e-commerce. Especially in autonomous driving engineering applications, image recognition technology can identify vehicles, pedestrians, traffic signs, obstacles, and other information on the road, thereby providing intelligent vehicles with a visual perception of the surrounding environment. In this section, to improve the accuracy and real-time performance of image recognition in autonomous driving applications, we conduct a series of empirical studies on the proposed method in campus road scenes. Specifically, we collected many images from the Northwestern Polytechnical University campus and meticulously organized them into a labeled dataset, which we called the campus scene dataset. This dataset consists of 2000 32 × 32 color images from campus road scenes, of which 1500 images are for training and 500 for testing. This dataset has 10 classes: automobile, pedestrian, plant, signpost, animal, building, bicycle, road, motorcycle, and traffic light. Each class contains 200 images, with 150 images for training and 50 for testing. The test set includes 50 randomly selected images from each class, while the training set contains the remaining images in random order. For all experiments on this dataset, we consistently adopt the data augmentation rules for CIFAR-100 classification specified in Section 4.1 to allow for fair and consistent comparisons. For optimization, we utilize a synchronous SGD optimizer with a weight decay of 5e-4, momentum of 0.9, and mini-batch size of 32, and employ a cosine learning schedule with an initial learning rate of 0.005 to optimize all models within 50 epochs. It is worth noting that all models are pre-trained on CIFAR-100 and then fine-tuned on the proposed dataset. When testing, all models only accept the original images.

To rigorously evaluate the practical benefits of the proposed MSPA module in campus road scenes, we compare it with three attention-based competitors, SPA, GC, and EPSA, using ResNet-50 as the backbone. Throughout the experiments, we employ parameters, FLOPs, and inference speed (tested on the Jetson AGX Orin platform) to evaluate efficiency and utilize Top-1 classification accuracy to assess effectiveness. All experimental results are listed in Table 8, from which the following insights can be obtained. Firstly, the experimental results

Table 8
Comparisons of different attention methods on the campus scene dataset in terms of parameters, FLOPs, inference speed (frame per second, FPS), and Top-1 classification accuracy (Top-1 Acc). All accuracy results are the mean ± standard deviation of 5 runs.

Architecture	Parameters (M)	FLOPs (G)	Inference speed (FPS)	Top-1 Acc (%)
ResNet-50	23.705	1.308	125	90.88 ± 0.19
ResNet-50 + SPA	31.266	1.319	54	92.76 ± 0.31
ResNet-50 + GC	26.253	1.312	87	92.24 ± 0.39
EPSANet-50 (Small)	20.710	1.148	152	92.98 ± 0.27
MSPANet-S-50 (Ours)	**21.917**	**1.234**	**139**	94.71 ± 0.16
MSPANet-B-50 (Ours)	23.769	1.340	98	**95.04 ± 0.13**

show that our method can significantly outperform the backbone in classification accuracy while having cheaper overhead, faster inference speed, and better stability. Remarkably, compared with ResNet-50, MSPANet-S-50 can achieve a 3.83% improvement in classification accuracy while saving 7.54% parameters and 5.66% FLOPs and increasing the inference speed by 11.2%. Secondly, our method always yields better results than SPA-based channel attention or GC-based spatial attention in almost all evaluation metrics. For example, MSPANet-B-50 achieves a Top-1 accuracy of 95.04%, an improvement of 2.28% over SPA-ResNet-50 (92.76%), while reducing parameters by 23.98% and increasing inference speed by 81.48%. Finally, compared to EPSANet-50, MSPANet-S-50 provides a 1.73% performance gain and better standard deviation with only a marginal additional cost. These comparisons show that our method can generalize well to campus road scenes and achieve superior performance with cheaper overhead, faster inference speed, and better stability, which aligns perfectly with our motivations and expectations.

In summary, empirical studies on the engineering-related dataset demonstrate that our method can effectively improve image recognition's accuracy and real-time performance in practical applications, thus contributing to the further advancement of autonomous driving applications. Meanwhile, these experimental results also reflect that our method has both essential theoretical significance and extensive practical value.

5. Conclusion

This study proposes a new technique to improve the performance and representation capability of CNNs. This technique, termed MSPA, is characterized by a low-overhead yet high-performance multi-scale spatial pyramid attention module with strong generalization ability for various network architectures and datasets. The proposed MSPA module mainly incorporates three core components: HPC module, SPR module, and Softmax operation, which can extract multi-scale spatial information, fully integrate structural regularization and structural information, and efficiently build long-range channel dependencies in attention learning. For the HPC module, we effectively extract multi-scale spatial informative features at a more granular level by constructing hierarchical residual-like connections in a single block, thereby increasing the range of receptive fields for each layer and further strengthening multi-scale representations. In the SPR module, we innovatively integrate structural regularization and structural information through an adaptive combination mechanism and explore inter-channel relationships, thus enhancing feature representations. Moreover, we efficiently establish long-range dependencies between channels by employing Softmax operation. We apply MSPA to the bottleneck residual blocks of ResNets and propose a series of novel backbones named MSPANet, which have powerful multi-scale representation capability and can recalibrate cross-dimension channel weights adaptively. We performed ablation experiments to comprehensively evaluate the MSPA performance and verify our design scheme's correctness and effectiveness. We also thoroughly compare MSPA with other SOTA counterparts on the small-scale CIFAR-100 and large-scale ImageNet-1K benchmarks. Extensive experimental results demonstrate that our method consistently outperforms other competitors, verifying the superiority of MSPA and supporting the inherent philosophy of our design. In addition, the visualization results obtained using the Grad-CAM++ technique show that our method significantly improves network performance in locating and recognizing objects of interest with greater precision. Finally, the experimental results on an engineering-related dataset confirm that our approach can effectively improve the accuracy and real-time performance of image recognition in autonomous driving at a cheaper cost, indicating the significant potential of our approach in engineering applications.

Although the MSPA module sheds light on the shortcomings of previous methods in attention learning and shows extraordinary performance, it still suffers from some limitations in modeling spatial attention, especially the inability to capture long-range contextual information in the spatial dimension, which is crucial for further improving learning ability. In the future, we intend to investigate the incorporation of MSPA with spatial attention module to develop a lightweight and efficient attention module capable of capturing long-range contextual information in both spatial and channel dimensions. In addition, we are confident that MSPA can handle even larger datasets and more complex computer vision tasks, including but not limited to semantic segmentation, object detection, instance segmentation, and depth estimation. We believe that the potential for future work in these areas is immense, and we are excited about the possibilities.

CRediT authorship contribution statement

Yang Yu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Yi Zhang:** Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Zeyu Cheng:** Formal analysis, Investigation, Data curation, Writing – review & editing. **Zhe Song:** Formal analysis, Investigation, Data curation, Writing – review & editing. **Chengkai Tang:** Formal analysis, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank editors for rigorous work and the anonymous reviewers for their comments and suggestions. This work was supported in part by National Natural Science Foundation of China under Grant 62171735, 62271397, 62173276, 62101458, 62001392, 61803310 and 61801394, in part by Natural Science Basic

Research Program of Shaanxi under Grant 2023-JC-QN-0760, 2022GY-097, 2021JQ-122 and 2021JQ-693, in part by Shenzhen Science and Technology Innovation Program under Grant JCYJ20220530161615033 and in part by China Postdoctoral Science Foundation under Grant 2020M673482 and 2020M673485.

References

Bakr, E.M., El-Sallab, A., Rashwan, M., 2022. EMCA: Efficient multiscale channel attention module. IEEE Access 10, 103447–103461.

Çalışkan, A., 2023. Finding complement of inefficient feature clusters obtained by metaheuristic optimization algorithms to detect rock mineral types. Trans. Inst. Meas. Control 45 (10), 1815–1828.

Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H., 2019. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop. ICCVW, IEEE, pp. 1971–1980.

Çoğalmiş, K.N., Bulut, A., 2022. Generating ad creatives using deep learning for search advertising. Turk. J. Electr. Eng. Comput. Sci. 30 (5), 1882–1896.

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 839–847.

Chen, C.-F.R., Fan, Q., Panda, R., 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 357–366.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.

Duta, I.C., Georgescu, M.I., Ionescu, R.T., 2021. Contextual convolutional neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 403–412.

Duta, I.C., Liu, L., Zhu, F., Shao, L., 2020. Pyramidal convolution: Rethinking convolutional neural networks for visual recognition. arXiv preprint arXiv:2006.11538.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3146–3154.

Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., Torr, P., 2019a. Res2net: A new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. 43 (2), 652–662.

Gao, Z., Xie, J., Wang, Q., Li, P., 2019b. Global second-order pooling convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3024–3033.

He, K., Zhang, X., Ren, S., Sun, J., 2015a. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2015b. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1904–1916.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer, pp. 630–645.

Hou, Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13713–13722.

Hu, J., Shen, L., Albanie, S., Sun, G., Vedaldi, A., 2018. Gather-excite: exploiting feature context in convolutional neural networks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 9423–9433.

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2020. Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. 42 (8), 2011–2023.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.

Krizhevsky, A., et al., 2009. Learning multiple layers of features from tiny images.

Lee, H., Kim, H.-E., Nam, H., 2019. Srm: A style-based recalibration module for convolutional neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1854–1862.

Li, G., Fang, Q., Zha, L., Gao, X., Zheng, N., 2022. HAM: Hybrid attention module in deep convolutional neural networks for image classification. Pattern Recognit. 129, 108785.

Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 510–519.

Liu, J.-J., Hou, Q., Cheng, M.-M., Wang, C., Feng, J., 2020. Improving convolutional networks with self-calibrated convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10096–10105.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022a. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

Ma, X., Guo, J., Sansom, A., McGuire, M., Kalaani, A., Chen, Q., Tang, S., Yang, Q., Fu, S., 2021. Spatial pyramid attention for deep convolutional neural networks. IEEE Trans. Multimed. 23, 3048–3058.

Misra, D., Nalamada, T., Arasanipalai, A.U., Hou, Q., 2021. Rotate to attend: Convolutional triplet attention module. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3139–3148.

Öztürk, Ş., Alhudhaif, A., Polat, K., 2021. Attention-based end-to-end CNN framework for content-based X-ray image retrieval. Turk. J. Electr. Eng. Comput. Sci. 29 (8), 2680–2693.

Öztürk, Ş., Turalı, M.Y., Çukur, T., 2023. HydraViT: Adaptive multi-branch transformer for multi-label disease classification from chest X-ray images. arXiv preprint arXiv:2310.06143.

Park, J., Woo, S., Lee, J.-Y., Kweon, I.-S., 2018. BAM: Bottleneck attention module. In: British Machine Vision Conference. BMVC, British Machine Vision Association (BMVA).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. pp. 8026–8037.

Qin, Z., Zhang, P., Wu, F., Li, X., 2021. Fcanet: Frequency channel attention networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 783–792.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J., 2019. Stand-alone self-attention in vision models. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. pp. 68–80.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. pp. 4278–4284.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826.

Trockman, A., Kolter, J.Z., 2023. Patches are all you need? Trans. Mach. Learn. Res..

Wan, D., Lu, R., Shen, S., Xu, T., Lang, X., Ren, Z., 2023. Mixed local channel attention for object detection. Eng. Appl. Artif. Intell. 123, 106442.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017. Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3156–3164.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al., 2020a. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43 (10), 3349–3364.

Wang, C., Wang, H., 2023. Cascaded feature fusion with multi-level self-attention mechanism for object detection. Pattern Recognit. 138, 109377.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020b. ECA-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11534–11542.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1492–1500.

Yang, H., Yang, M., He, B., Qin, T., Yang, J., 2022. Multiscale hybrid convolutional deep neural networks with channel attention. Entropy 24 (9), 1180.

Yang, L., Zhang, R.-Y., Li, L., Xie, X., 2021. Simam: A simple, parameter-free attention module for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 11863–11874.

Yang, Z., Zhu, L., Wu, Y., Yang, Y., 2020. Gated channel transformation for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11794–11803.

Yu, Y., Zhang, Y., Cheng, Z., Song, Z., Tang, C., 2023a. MCA: Multidimensional collaborative attention in deep convolutional neural networks for image recognition. Eng. Appl. Artif. Intell. 126, 107079.

Yu, Y., Zhang, Y., Song, Z., Tang, C.-K., 2023b. LMA: lightweight mixed-domain attention for efficient network design. Appl. Intell. 53 (11), 13432–13451.

Yu, W., Zhou, P., Yan, S., Wang, X., 2023c. Inceptionnext: When inception meets convnext. arXiv preprint arXiv:2303.16900.

Yuan, P., Lin, S., Cui, C., Du, Y., Guo, R., He, D., Ding, E., Han, S., 2020. HS-ResNet: Hierarchical-split block on convolutional neural network. arXiv preprint arXiv:2010.07621.

Zhang, X., Du, B., Wu, Z., Wan, T., 2022a. LAANet: lightweight attention-guided asymmetric network for real-time semantic segmentation. Neural Comput. Appl. 34 (5), 3573–3587.

Zhang, Q.-L., Yang, Y.-B., 2021. Sa-net: Shuffle attention for deep convolutional neural networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 2235–2239.

Zhang, H., Zu, K., Lu, J., Zou, Y., Meng, D., 2022b. EPSANet: An efficient pyramid squeeze attention block on convolutional neural network. In: Proceedings of the Asian Conference on Computer Vision. pp. 1161–1177.

Zhou, H., Qi, L., Huang, H., Yang, X., Wan, Z., Wen, X., 2022. CANet: Co-attention network for RGB-D semantic segmentation. Pattern Recognit. 124, 108468.