Applied Intelligence (2023) 53:141–152
https://doi.org/10.1007/s10489-022-03225-9

Inner-imaging 3D attention module for residual network

Wenjie Liu[1,2] · Guoqing Wu[1] · Fuji Ren[3] · Quan Shi[2]

Accepted: 9 January 2022 / Published online: 13 April 2022
© The Author(s) 2022

Abstract

We propose an Inner-Imaging three-dimensional (3D) attentional feature fusion module for a residual network, which is a simple yet effective approach for residual networks. In our attention module, we constructed a 3D soft attention feature map to refine the input feature. The map fuses the attentional features from different dimensions, including channel and spatial axes, to create a 3D attention map. Then, we implemented a feature fusion module to further fuse the attentional features. Lastly, the attention module outputs a 3D soft attention map that is applied to the residual branch. The attention module can also model the relationship between attentional features from different dimensions and achieve the interaction between attentional features. This function allows our attention module to acquire more attentional features. To demonstrate the effectiveness of our method, extensive experiments were conducted on several computer vision benchmark datasets, including ImageNet 2012 and Microsoft COCO (MS COCO) 2017 datasets. The experimental results show that our method performed better than the baseline methods in the tasks of image classification, object detection, and instance segmentation tasks.

Keywords Attention mechanism · Feature fusion · Object recognition · Residual network

1 Introduction

Convolutional neural networks (CNNs) have created a significant improvement in representation power, in areas such as image classification [1–4, 37], object detection [5–8], and segmentation [9–11]. CNN models designed in recent years include GoogleNet [3], DenseNet [4], ResNet [1], GFNet [12], and PAG-Net [34]. The residual network (ResNet) performed well during the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015. ResNet enabled the network structure to go far deeper and achieve a higher performance with skip connection. Other strategies to improve the model's representation power include going wider, increasing cardinality, and refining features dynamically. Another method for improving performance is the attention mechanism.

The attention mechanism has been studied in areas such as natural language processing [13, 14], image classification [15–19], object re-identification [20, 21, 35, 38], and other domains [22, 23]. The attention mechanism aims at selectively focusing on specific information. Common types of attention variants include spatial attention, channel attention, and self-attention, all of which depend on different feature dimensions. Channel attention constructs various channel weight functions and is widely used due to its simplicity and effectiveness in feature modeling. In SENet [15], the authors proposed a squeeze-and-excitation architecture to model the importance of different channels, and it became a popular tool for improving model performance. However, when processing different inputs, global average pooling cannot capture rich input

✉ Guoqing Wu
wgq@ntu.edu.cn

✉ Fuji Ren
ren@is.tokushima-u.ac.jp

Wenjie Liu
lwj2014@ntu.edu.cn

Quan Shi
sq@ntu.edu.cn

1 School of Information Science and Technology, Nantong University, Seyuan Road, Nantong, 226019, Jiangsu Province, China

2 School of Transportation and Civil Engineering, Nantong University, Seyuan Road, Nantong, 226019, Jiangsu Province, China

3 Faculty of Engineering, Tokushima University, Shinkura-cho, Tokushima, 770-8506, Tokushima, Japan

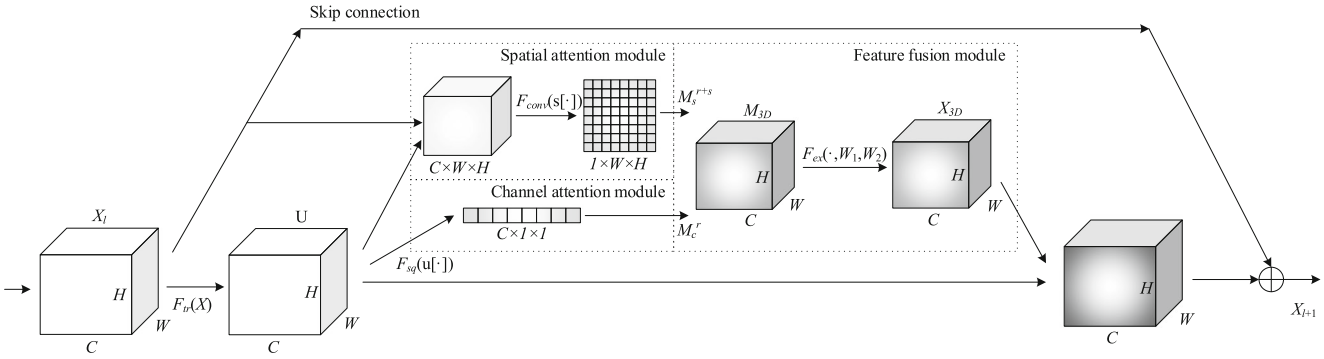


Fig. 1 Proposed inner-imaging 3D attentional feature fusion residual block

patterns and feature diversity. In this paper, we address these problems by proposing a new attention module and creating a three-dimensional (3D) soft attention feature map to refine the input feature.

In our attention module, we focus on fusing attention features from the channel and spatial dimensions to create a 3D soft attention feature map. To obtain more attentional features, we combine the information from the skip connection branch. Based on these settings, our attention module can be smoothly embedded into the residual block, as shown in Fig. 1. We called our attention module Inner-Imaging 3D Attentional Feature Fusion Residual Network (3D-AFF-ResNet). It consists of a channel attention module, a spatial attention module, and a feature fusion module. The channel attention module squeezes the input into a 1D vector, focusing on learning the object feature in channel axes. The spatial attention module compresses the input feature into a 2D vector, focusing on learning the object feature in spatial axes, and it also models the inter-spatial relationship of local spatial features. These two modules extract features in different dimensions to obtain the channel and spatial attentional feature, respectively. The function of the two modules is similar to the CBAM [16], which proposed a sequential channel and spatial attention module. However, in our attention module, we combine the channel and spatial attentional features instead of applying them sequentially. This allows our attention module to extract more features. We then fused the features from two attention modules using element-wise summation to create a 3D attention feature map. After the summation operation, the features are feed into the feature fusion module. This not only improves the non-linear representation ability but also performs attentional feature interaction from different dimensions. Then, to address the lack of feature diversity, we combined the features from the skip connection branch to extract more spatial attentional features. Evaluating the results verified the effectiveness of our method. Lastly, we applied the output of the refined 3D soft attentional feature map to the residual branch.

To test the validity of our method, we evaluated 3D-AFF-ResNet on the ImageNet dataset for the task of image classification and MS COCO dataset for the task of object detection and segmentation. We used ResNet and SENet as our baseline models. The evaluations showed that our method achieved considerablely improved performance compared to the baseline models. We also constructed extensive ablation experiments to explore the properties of our attentional module.

The main contributions of this paper are summarized as follows:

1) We proposed a 3D soft attention module, which could refine features more precisely. The experimental results indicate that our attention module can be embedded into residual network seamlessly and improve the model performance significantly for the tasks of image classification, object detection, and instance segmentation.
2) To further boost the model's performance, we designed the spatial attention module and feature fusion module and had the attentional feature interaction from different dimensions.
3) To improve the model's generalization ability, we fused the attentional feature from multiple branches and modeled the relationship of these branches.

The rest of the paper is organized as follows. In Section 2, we review related works. In Section 3, we introduce our methodology. In Section 4, we discuss the experimental results and analysis on the ImageNet and Microsoft COCO datasets. In Section 5, we focus on the validity of our approach. In Section 6, we present our conclusions.

2 Related work

First, we discuss multi-branch CNNs. Then, we discuss the feature fusion method and attention mechanism in CNNs.

2.1 Multi-branch convolutional neural networks

The introduction of a bypass path can make the training less difficult. ResNet [1] adds a identity mapping in each unit, which enables the model to train with hundreds of layers. In GoogleNet [3], the author combined the feature from multiple branches with different kernel sizes. Inspired by the multiple paths method, W. Liu et al. [24] also proposed a multi-branch feature fusion residual block to learn multi-scale features from different branches. Z. Zheng et al. [25] proposed a multi-branch discriminator structure based on generative adversarial network to address imbalance learning for semantic matching. To explore local information at the final stage of the learning process, F. Hernández-Luquin et al. [26] proposed a CNN-based architecture, which is enhanced by multiple branches module. These models demonstrate that combining the features from different scales could improve model performance. In our attention module, we also utilize the information from multiple branches to enrich the extracted features. We also attempted to model the relationship of these features.

2.2 Feature fusion method

The feature fusion method is used in many works to improve the ability to extract features. In InceptionNets [3, 27, 28], models fuse features from several branches with different kernel sizes to enhance the feature diversity. In this manner, InceptionNets aggregate various features from multiple branches, each of which is equipped with customized kernel filters. This gives InceptionNets powerful generalization ability and allows them to perform well in computer vision tasks. In residual-style networks [1, 2, 24, 29, 30], models fuse the features with skip connection to alleviate the difficulty of training. Feature Pyramid Networks (FPNs) [31] fuse the features from shallow layers via skip connections to attain high-resolution and semantically strong features. In KMSA [36], the authors proposed a general framework to transform multi-view data into one channel by kernel space. These models demonstrate the effectiveness of the feature fusion method. Our attention module also uses this method to fuse features from the residual branch and the skip connection branch.

2.3 Attention mechanism in CNNs

In recent years, attention mechanisms have been used in a range of tasks, from object re-identification [20, 21] to neural machine translation [13]. The attention mechanism intensifies the useful information and simultaneously suppresses less useful information. CNNs are widely used in the field of computer vision, and the attention mechanism further improves their performance. Guan et al. [22] proposed an innovative cascade convolution neural network with a particular spatial-channel noise attention unit to separate fixed pattern noise and recover the real scene. Z. Yan et al. [39] proposed a feature attention network to refine important feature and learn the correlations among convolutional features. Other researchers have focused on designing lightweight attention architectures to increase model performance. In SENet [15], the authors introduced a compact module to exploit the inner-channel relationship with few parameters and a reduced computational burden. After SENet, more variants were proposed, such as CBAM [16], SKNet [17], GCNet [18], ECANet [19]. CBAM introduced spatial attention to improve performance. SKNet proposed an adaptive selection receptive field size of neurons with attention mechanism. GCNet proposed a spatial attention module to replace the original spatial downsampling process. To reduce the redundancy of dense connection layers, ECANet introduced a one-dimensional convolutional layer. In our attention module, we focus on constructing a 3D attention structure to refine the input features more precisely.

3 Methodology

The Inner-Imaging 3D attentional feature fusion module consists of three parts, as illustrated in Fig. 1. $X_l \in R^{C\times H\times W}$ is used to represent the l-th layer input feature map. The channel attention module is used to create a 1D attentional feature vector $M_c^r \in R^{C\times 1\times 1}$. The spatial attention module is used to generate a 2D attention feature map $M_s^{r+s} \in R^{1\times H\times W}$. Then M_c^r and M_s^{r+s} are fused by element-wise summation to generate a 3D attentional feature map $M_{3D} \in R^{C\times H\times W}$. After the summation operation, the fused feature is fed into the feature fusion module, which is used to capture the channel and spatial attentional feature dependencies with two successive 1×1 convolution blocks as F_{ex}. Lastly, we obtain a 3D soft attentional feature map $X_{3D} \in R^{C\times H\times W}$, and the weight of each pixel represents the importance in the residual branch. The overall Inner-Imaging 3D attentional feature fusion module process can be formulated as:

$$M_{3D} = M_c^r \oplus M_s^{r+s} \tag{1}$$

$$X_{3D} = \sigma\left(F_{ex}(M_{3D}, W_1, W_2)\right) \tag{2}$$

Where $\oplus$ refers to element-wise summation. During element-wise summation, the attention values are broadcast (copied) along the channel and spatial dimension. M_c^r denotes the channel attention feature from the residual branch, and M_s^{r+s} denotes the spatial attention from the residual and skip connection branches. W_1 and W_2 represent

the parameters in the feature fusion block, and σ denotes the sigmoid function.

3.1 Channel attention module

To acquire the channel attentional feature, we follow the operation in SENet [17] by squeezing the input into a 1D feature vector with a global average pooling layer. Each feature map from the input can be regarded as a feature detector that detects the object in the input image. Therefore, the channel attention is focused on extracting the global feature from an input image. Average pooling is a commonly used method for obtaining the channel information. As shown in Fig. 1, we compress global information by using an average pooling layer. This generates the channel context descriptors, and M_c^r denotes the average-pooled feature from the residual branch.

$$U = F_{tr}(X_l) \tag{3}$$

$$M_c^r = F_{sq}(U) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u^c(i, j) \tag{4}$$

In (3), F_{tr} refers to the function of successive convolutional blocks in the residual block, and U represents the output feature maps. $M_c^r \in R^{C\times 1\times 1}$ refers to the squeezed 1D feature vector, and F_{sq} denotes the function of average pooling. u^c denotes the c-th feature map in U, H represents the height, and W is the width of the feature map.

The channel attention module squeezes the channel feature map into a 1D vector. However, this operation cannot extract spatial features very well and fails to consider the inter-spatial relationship of local features among feature maps. We proposed the spatial attention module to address these problems.

3.2 Spatial attention module

The spatial attention module extracts the spatial features of the input. We use a convolution layer to compress the spatial feature into a 2D feature map and to model the inter-spatial relationship of local features among feature maps. The compressed 2D feature map from the residual branch is represented as $M_s^r \in R^{1\times H\times W}$. To diversify the extracted features, we fuse the features from the skip connection branch to improve the normalization ability of our attention module. We believe that the feature diversity could improve the model's generalization ability, and we proved this hypothesis in our experiments. To keep the number of parameters low, we fused the features from multiple branches with an element-wise summation. We found that the kernel size significantly affects model performance, which means that an appropriate receptive field could bring benefits to our attention module. The spatial attention module can be formulated as:

$$M_s^{r+s} = F_{conv}\left(U \oplus X_l, W'\right) \tag{5}$$

where F_{conv} denotes the function of convolution block and W' denotes the parameters. $M_s^{r+s} \in R^{1\times H\times W}$ represents the compressed spatial features from the residual and skip connection branches.

3.3 Feature fusion module

The channel and spatial attention modules compute the attentional feature in a complementary manner, extracting the feature in different dimensions. As described in the (1), we use element-wise summation to fuse the features from two modules, thereby obtaining the 3D fused feature map $M_{3D} \in R^{C\times W\times H}$. The pixels in M_{3D} contain the features from the channel and spatial attention modules. However, the simplified element-wise summation operation could not fully fuse these features. To address this problem, we implement the feature fusion module, which contains two successive 1×1 convolution layers with the ReLU function to improve the nonlinear representation capability. In addition, to decrease the number of parameters, we employ a bottleneck architecture, which reduces the module complexity. The formula of this module can be represented as:

$$\begin{aligned} X_{3D} &= \sigma\left(F_{ex}(M_{3D}, W_1, W_2, d)\right) \\ &= \sigma\left(W_2\delta(M_{3D}W_1)\right) \end{aligned} \tag{6}$$

Where $X_{3D} \in R^{C\times H\times W}$ denotes the output of the 3D soft attention vector, F_{ex} denotes the function of the feature fusion block, $W_1 \in R^{C\times W\times H}$ and $W_2 \in R^{\frac{C}{d}\times W\times H}$ denote the parameters of the two 1×1 convolution layers, respectively, and d refers to the reduction ratio.

Finally, the 3D soft attention vector X_{3D} is applied to the residual branch by performing the element-wise multiplication operation, and the input X_l is summarized to the output.

$$X_{l+1} = U * X_{3D} + X_l \tag{7}$$

Here, X_{l+1} represents the output of the l-th residual block.

4 Experiments and analysis

We evaluated our method on two standard benchmark datasets: ImageNet 2012 for image classification, and MS COCO for object detection and instance segmentation. First, we performed extensive ablation studies to thoroughly exploit the properties of our method. Next, we evaluated our method on ImageNet, comparing it with the baseline models, and observed the effect of our approach with our

Table 1 Comparison of the channel and spatial attention from residual and skip connection branches on ImageNet

Description	Parameters	Top-1 Error (%)	Top-5 Error (%)
ResNet-50	25.56M	23.39	6.93
ResNet-50 + M_c^r (SE-ResNet-50)	28.07M	22.84	6.35
ResNet-50 + M_c^r + M_s^r	28.24M	22.51	6.32
ResNet-50 + M_c^r + M_s^{r+s} (3D-AFF-ResNet-50)	28.38M	22.20	6.17
ResNet-50 + M_c^{r+s} + M_s^{r+s}	28.38M	22.21	6.21

attention module. We then summarized the experimental results on MS COCO. The results also demonstrated that our method outperforms the baseline models, and also verified our method's adaptability across different tasks.

4.1 Implementation details

To verify the effectiveness of our design choice, ResNet-50 was used as the base architecture, and we evaluated the proposed 3D-AFF-ResNet on the ImageNet 2012 dataset. ImageNet 2012 consists of 1.2 million images for training images and 50,000 images for validation with 1000 classes. We followed the data augmentation and hyperparameter setting in [1], cropping the input images to 224×224 with random horizontal flipping. The learning rate was set to 0.1 and dropped every 30 epochs. All models were trained for 100 epochs. We used an SGD optimizer with a momentum of 0.9, a batch size of 128, and a weight decay of 1e-4. For training efficiency and to save memory, we use a mixed-precision training method. For the setting of the reduction ratio, we followed the selection in the SENet and set $d = 16$ for ImageNet 2012. Furthermore, to demonstrate the generalization of our designing choices, we also conducted ablation study on CIFAR-100 dataset. The CIFAR-100 dataset consists of 50,000 training images and 10,000 testing images with 100 classes. We used the standard data augmentation strategies as described in [1]. For the reduction ratio, we set $d = 4$ for CIFAR-100 dataset.

To evaluate our method for the task of object detection task on MS COCO, we used the Faster R-CNN [8] and Mask R-CNN architectures. For the task of instance segmentation, we evaluated our method on MS COCO using the Mask R-CNN architecture. We tested the performance on the MMDetection toolkit platform, using its default settings. The short side of the input image was resized to 800 pixels during the training period. The learning rate was set to 0.005 and was reduced by a factor of 10 at the eighth and 11th epochs, respectively. SGD was used to optimize with a weight decay of 1e-4, a momentum of 0.9, and a batch size of 2 per GPU within 12 epochs.

We implemented 3D-AFF-ResNet using PyTorch [32]. All models used in this paper were trained on two Nvidia Titan RTX GPUs.

4.2 Ablation study

The process of designing our attention module consisted of three parts. First, we fused the spatial attention feature from the residual branch to compare it with the baseline models. Next, we fused the feature from the residual branch and the skip connection branch to explore the effect of feature diversity for our attention module. Finally, we searched for the best kernel size for the spatial attention module.

Feature fusion In the spatial attention module, we compressed the input into a 2D spatial feature map. Experiments demonstrated that fusing the feature from the spatial attention module could achieve finer attention inference. The experimental results are summarized in Tables 1 and 2. Table 1 shows the ResNet-50 has top-1 and top-5 test error rates of 23.39% and 6.93%, respectively, on ImageNet 2012. The ResNet-50 + M_c^r outperforms ResNet-50 significantly. Notably, the architecture of ResNet-50 + M_c^r is the same as SE-ResNet-50, which only utilizes the channel attentional feature from the residual branch. However, the channel attention feature could not represent the spatial feature, nor could it consider the inter-spatial relationship of spatial feature among input feature maps. To address these problems,

Table 2 Comparison of the channel and spatial attention from residual and skip connection branches on CIFAR-100

Description	Parameters	Test Error (%)
ResNet-164	1.7M	22.72
ResNet-164 + M_c^r (SE-ResNet-164)	2.52M	22.00
ResNet-164 + M_c^r + M_s^r	2.59M	21.65
ResNet-164 + M_c^r + M_s^{r+s} (3D-AFF-ResNet-164)	2.66M	21.23
ResNet-164 + M_c^{r+s} + M_s^{r+s}	2.66M	21.52

Table 3 Comparison of kernel size in the spatial attention module on ImageNet

Description	Params	Top-1 Error (%)	Top-5 Error (%)
3D-AFF-ResNet($k = 1$)	28.13M	22.38	6.35
3D-AFF-ResNet($k = 3$)	28.38M	22.20	6.17
3D-AFF-ResNet($k = 5$)	28.86M	22.35	6.27

Table 4 Comparison of kernel size in the spatial attention module on CIFAR-100

Description	Params	Test Error (%)
3D-AFF-ResNet($k = 1$)	2.52M	22.38
3D-AFF-ResNet($k = 3$)	2.59M	22.20
3D-AFF-ResNet($k = 5$)	2.72M	22.35

we constructed the spatial attention module. To demonstrate the critical nature of the spatial attention feature, we fused the spatial attention from the residual branch, as ResNet-50 + M_c^r + M_s^r in Table 1. The top-1 test error rate was reduced by 0.33%, with a slight increase in the number of parameters compared with SE-ResNet-50. Furthermore, comparing the experimental results in Table 2, the model also has a lower test error rate, as it fuses the spatial attention from the residual branch. The experiments demonstrate that combining spatial attention with channel attention can enhance the model's performance.

Feature diversity We explored a more effective method for attention-refined features by aggregating the features from the skip connection branch to reduce the difficulty. Combining the features from the skip connection branch enriched the feature diversity and improved the model's generalization ability. We fused the features from two branches by element-wise summation without adding parameters. From Table 1, we can see that fusing the features from the skip connection branch, as ResNet-50 + M_c^r + M_s^{r+s} (3D-AFF-ResNet-50), further decreased the test error rate. The 3D-AFF-ResNet-50 has a top-1 test error rate of 22.20%, outperforming ResNet-50 + M_c^r + M_s^r by 0.31%. The experimental results in Table 2 also show that ResNet-164 + M_c^r + M_s^{r+s} achieved the best results. The experimental results on ImageNet 2012 and CIFAR-100 datasets verified that fusing the features from the skip connection branch in the spatial attention module was beneficial.

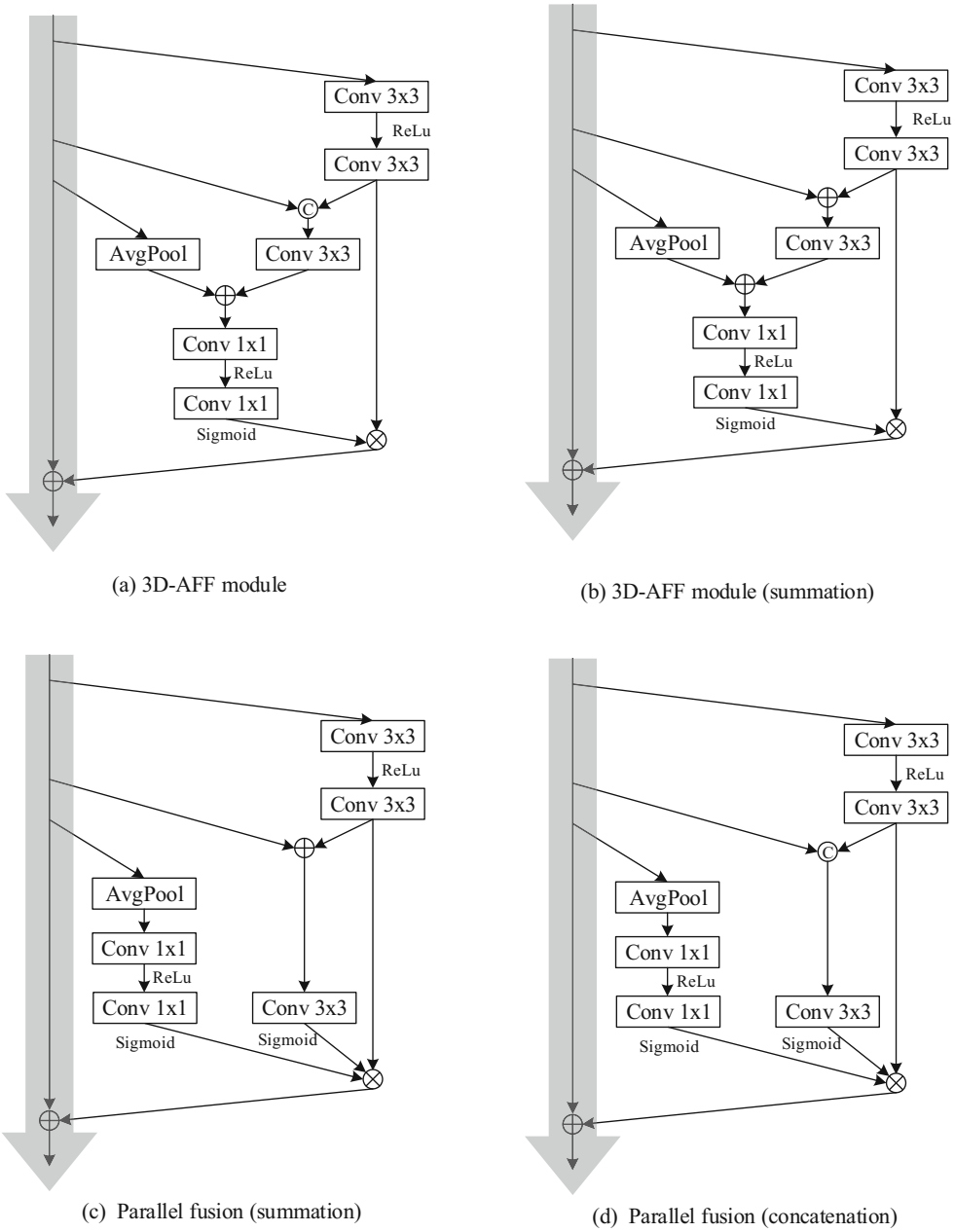
We also explored the effect of feeding the skip connection into the channel attention modules, as ResNet-50 + M_c^{r+s} + M_s^{r+s} in Table 1 and ResNet-164 + M_c^{r+s} + M_s^{r+s} in Table 2. The tested performance of this was slightly higher than 3D-AFF-ResNet-50 and 3D-AFF-ResNet-164. Therefore, we only fed the features from the skip connection branch into the spatial attention module.

Kernel size in spatial attention module An appropriate receptive field is necessary to obtain enough features for our tasks. In our attention module, the spatial attention module was applied to extract spatial features and model the inter-spatial relationship of the local spatial features. Accordingly, we determined that an appropriate receptive field was also critical for our attention module. The experiments proved this hypothesis.

In our experiments, we compared three different kernel sizes in spatial attention module: $k = 1$, $k = 3$, and $k = 5$. We also evaluated these options on ImageNet 2012 and CIFAR-100 datasets. Tables 3 and 4 summarize the experimental results. The results show that using $k = 3$ generated the best results. Therefore, the experiments empirically demonstrated that an appropriate receptive field is significant for deciding local spatial important regions and modeling the inter-spatial relationship of local features. As a result, we used the $k = 3$ kernel size for our spatial attention module.

Concatenation or summation In the original 3D-AFF module, we used the concatenation operation to combine the features from the residual and skip connection branches. However, the concatenation operation needed more parameters and memory during training. Therefore, we also explored replacing the concatenation operation with the element-wise summation, as illustrated in Fig. 2(b). Compared with our original module, this option could save parameters, but the test error rate (Table 5, 22.35%) was substantially higher than 3D-AFF-ResNet-50. Therefore, we chose the concatenation operation to combine features in our attention module.

Parallel fusion The 3D-AFF module first fused the channel and spatial features, then used the feature fusion module to improve the nonlinear ability and fuse the channel and spatial attentional features. To evaluate the importance of the feature fusion module for our attention module, we implemented two parallel fusion architectures, which performed the multiplication operation to the residual branch respective, as shown in Fig. 2(c) and (d). We also compared two different fusing methods: element-wise summation and concatenation. The two options achieved a top-1 test error rate of 22.49% and 22.46%, respectively, which was also higher than 3D-AFF-ResNet-50 but lower than SENet. Therefore, these ablation experiments empirically showed the validity of the feature fusion module, which could fuse the channel and spatial attentional features even further.



(a) 3D-AFF module

(b) 3D-AFF module (summation)

(c) Parallel fusion (summation)

(d) Parallel fusion (concatenation)

Fig. 2 Various types of feature fusion used in Table 5. The grey arrows indicate the identity mapping. For simplicity, the BN layers are omitted here

Final module design Our design of the channel and spatial attentional feature fusion module was based on the results of the ablation studies. Figure 1 shows our final module, in which we fused the channel and spatial attentional features; aggregated the spatial attentional feature from the skip connection branch to enrich the extracted features; and used convolution with a kernel size of 3 in the spatial attention module. Our final module (i.e. 3D-AFF-ResNet-50) has a

Table 5 Testing results on ImageNet dataset with different fusing methods

Description	Fig.	Top-1 Error (%)	Top-5 Error (%)
SENet-50	–	22.84	6.35
3D-AFF-ResNet-50	Fig. 2(a)	22.20	6.17
3D-AFF-ResNet-50(summation)	Fig. 2(b)	22.35	6.24
Parallel fusion (summation)	Fig. 2(c)	22.49	6.28
Parallel fusion (concatenation)	Fig. 2(d)	22.46	6.27

Table 6 Comparison of baseline methods on ImageNet

Method	Backbone	Params	FLOPs	Top-1 Error (%)	Top-5 Error (%)
ResNet [1]	ResNet-34	21.80M	3.68G	26.69	8.60
SENet [15]		21.95M	3.68G	26.13	8.35
ECANet [19]		21.80M	3.68G	25.79	8.17
3D-AFF-ResNet		22.00M	3.73G	25.32	8.04
ResNet [1]	ResNet-50	25.56M	4.12G	23.39	6.93
SENet [15]		28.07M	4.13G	22.84	6.35
CBAM [16]		28.07M	4.14G	22.61	6.31
ECANet [19]		25.56M	4.13G	22.70	6.32
3D-AFF-ResNet		28.24M	4.64G	22.20	6.17
ResNet [1]	ResNet-101	44.55M	7.85G	23.17	6.52
SENet [15]		49.29M	7.86G	22.38	6.07
CBAM [16]		49.30M	7.88G	21.51	5.69
ECANet [19]		44.55M	7.86G	21.35	5.68
3D-AFF-ResNet		49.65M	8.87G	21.19	5.73

top-1 error rate of 22.20%, which is much lower than that of the baseline models.

4.3 Image classification on ImageNet

To comprehensively evaluate our method comprehensively, we employed three widely used CNNs as backbone models: ResNet-34, ResNet-50, and ResNet-101. We compared 3D-AFF-ResNet with the baseline methods on ImageNet. As shown in Table 6, our approach outperformed the baseline models and other state-of-the-art methods. Our method outperformed SENet by a large margin with only a few additional parameters. 3D-AFF-ResNet has better performance than SENet by 0.81%, 0.64%, and 1.19% in terms of Top-1 test error rate with three different backbones, respectively. Figure 3 shows the top-1 test error curves

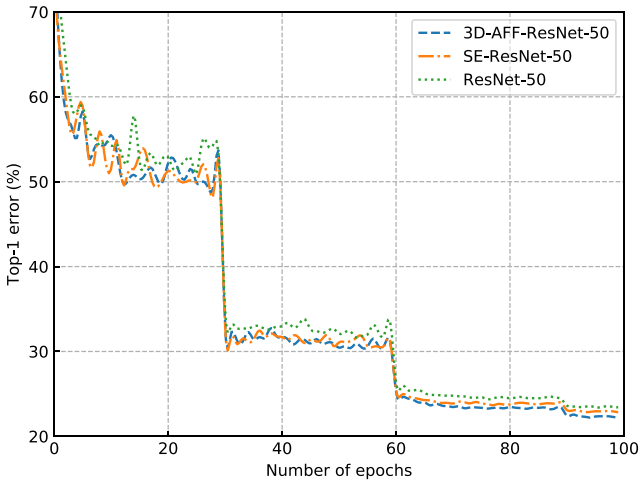


Fig. 3 Top-1 test error curves on ImageNet 2012 dataset by ResNet-50, SE-ResNet-50, and 3D-AFF-ResNet-50

of the baseline models and 3D-AFF-ResNet-50 during different epochs. Furthermore, our method achieved a better performance than CBAM and ECANet. These experimental results show the effectiveness of our method and the 3D soft attention feature map. Besides, we also tested the speed at inference on CPU (AMD Ryzen 7 3700X) with different method, and the time consumption for ResNet-50, SE-ResNet-50, and 3D-AFF-ResNet-50 are 185, 188, and 209 ms per image, respectively.

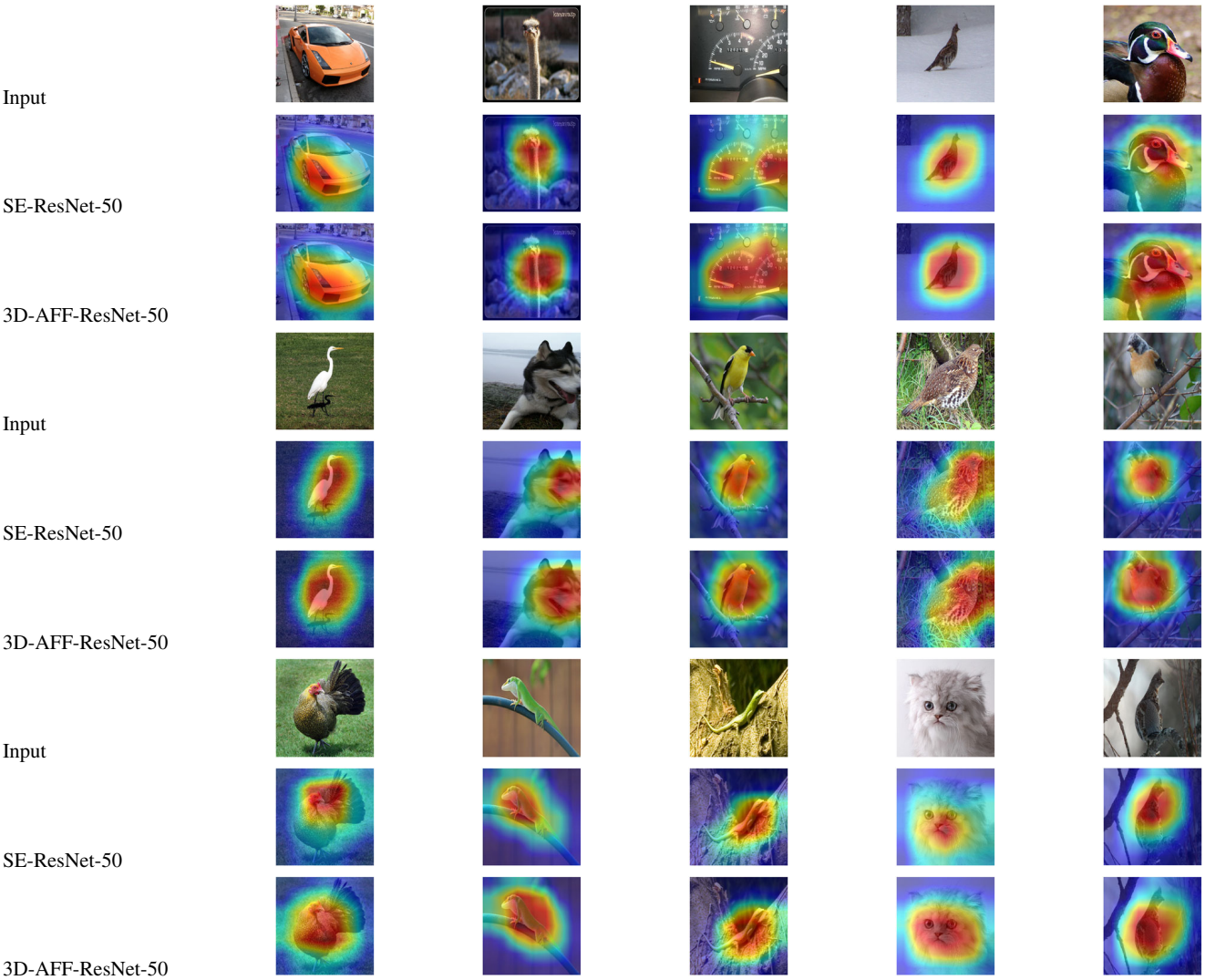
4.4 Network visualization with Grad-Cam

To analyze the effectiveness of our method, we used the Grad-Cam [33] to highlight the important regions for the task of image classification. Grad-Cam is a visualization method that uses a gradient to calculate the importance of the spatial location in convolution layers. Some images were randomly selected from the ImageNet validation set. By visualizing the regions that the model regards as important for predicting a class, we could clearly observe the impact of our method on model performance. To evaluate the effect of our attention module, we compared the visualization results of 3D-AFF-ResNet-50 with SE-ResNet-50. Table 7 shows these results. We observed that the highlighted region generated by 3D-AFF-ResNet-50 was larger than SE-ResNet-50, which indicated that our attention module could enable the model to focus on a wider important region for image classification.

4.5 Object detection and instance segmentation on MS COCO

We also conducted the task of object detection on the MS COCO dataset is also conducted. The dataset includes

Table 7 Highlighted important regions



118,000 training images ("2017 train") and 5000 validation images ("2017 val"). We used our 3D-AFF-ResNet with FPN [31] as the backbone (ResNet-50) of Faster rcnn and Mask rcnn. All the models were tested on the MS COCO validation dataset. Tables 8 and 9 show the experimental results.

Table 8 Objection detection results on the COCO val 2017

Method	Detector	AP	AP50	AP75	APS	APM	APL
ResNet-50	Faster rcnn	37.6	58.4	40.9	21.5	41.4	48.6
SE-ResNet-50		39.0	60.5	42.3	23.4	43.0	49.8
CBAM-50		39.3	60.8	42.4	24.5	43.1	50.5
ECANet-50		39.3	60.8	42.9	23.6	43.0	50.1
3D-AFF-ResNet-50		39.4	60.8	42.6	22.9	43.4	50.8
ResNet-50	Mask rcnn	38.3	59.1	41.8	22.3	41.6	50.2
SE-ResNet-50		39.5	60.7	42.8	23.5	43.3	51.2
CBAM-50		40.0	61.2	43.6	24.3	43.7	52.1
ECANet-50		40.0	61.4	43.4	23.8	43.6	51.1
3D-AFF-ResNet-50		40.3	61.4	43.9	24.0	43.7	52.4

Table 9 Instance segmentation results with different methods on the COCO dataset

Method	AP	AP50	AP75
ResNet-50	34.8	55.9	37.0
SE-ResNet-50	35.6	57.6	37.8
CBAM-50	36.1	57.7	38.6
ECANet-50	36.1	58.0	38.4
3D-AFF-ResNet-50	36.3	58.1	38.5

As shown in Table 8, Faster rcnn and Mask rcnn were used as our detection method. Here we focused on demonstrating the effectiveness of plugging the 3D-AFF module into the baseline network. Because the same detection method was used in all models, the gains can only be due to the enhanced representation power, given by the 3D-AFF module or the spatial attention module. The experimental results show that 3D-AFF-ResNet-50 achieves significant improvements over the baseline models, which demonstrates the generalization ability of the 3D-AFF module on the task of object detection.

We also evaluated our method for the task of instance segmentation. Table 9 shows that our approach outperformed the baseline models by a considerable margin. Our approach outperformed ResNet-50 by 1.5% AP and SE-ResNet-50 by 0.7%, and our model also had better performance than ECANet-50 and CBAM-50. These results demonstrate the validity of our approach.

5 Discussion

In this section, we discuss the validity of our approach - first the role of the spatial attention module and then the feature interaction in our attention module.

Role of spatial attention module As shown in Fig. 2, we compared several variants of the 3D-AFF module. In these modules, we explored different methods of fusing the channel and spatial attentional features. The experimental results showed that all the variants achieved better performance than SENet. Therefore, these experiments demonstrated the hypothesis that local spatial attention has a critical impact on our attention module, which could improve the model's performance significantly. We also compressed the feature from the residual and skip connection branches into a 2D vector. Therefore, the vector not only contained the features from the residual and skip connection branches but also extracted the dependencies of the local spatial features. These settings enabled for the feature fusion module to extract more of the inter-spatial relationship of local features. Based on this analysis, the empirical experiments verified that the spatial attention module could do a good job of extracting the spatial features and the inter-spatial relationship of local features.

Feature interaction The impact of feature interaction in our attention module can be looked at from two aspects. The first is the feature interaction between channel and spatial attentional features. Figure 2(c) and (d) show how the proposed parallel fusing attention modules apply the channel and spatial attentional features to the residual branch, respectively. Therefore, these modules are implemented without feature interaction. Compared to the other cases in Fig. 2, the modules, which were implemented with the feature fusion module, achieve lower top-1 and top-5 test error rates, as shown by the experimental results in Table 5. Therefore, these experiments empirically demonstrate that the feature interaction between channel and spatial attentional features boost the model's performance. The second aspect is the feature interaction in the spatial attention module, which extracts the spatial feature and the inter-spatial relationship of local spatial features from the residual and skip connection branches. Based on the above analysis, the empirical experiments demonstrated that feature interaction plays a critical role in our attention module.

6 Conclusion

In this work, we propose a light-weight 3D attention module for residual network. The experiments demonstrated that our attention module could fuse the channel and spatial features from the residual and skip connection branches. Furthermore, our attention module could extract the spatial features and model the inter-spatial relationship of local spatial features among input feature maps. Extensive ablation studies empirically verified the properties of our attention module. To evaluate the effectiveness of our method, we tested 3D-AFF-ResNet on the ImageNet 2012 dataset, and the experimental results showed that our method could achieve better performance than the baseline methods. We also tested the effect of our method on other computer vision tasks, including object detection and instance segmentation. The experimental results also showed that our method could achieve better performance than the baseline models in these tasks as well.

In the future, we will try to implement more effective attention method and apply the method to different domains.

Acknowledgements This work was partially supported by the National Natural Science Foundation of China (Grant no. 61872425 and no. 61771265) and the Research Clusters program of Tokushima University (No. 2003002).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
2. He K, Zhang X, Ren S et al (2016) Identity mappings in deep residual networks. In: European conference on computer vision. Springer, Cham, pp 630–645
3. Szegedy C, Liu W, Jia Y et al (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
4. Huang G, Liu Z, Van Der Maaten L et al (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
5. Pang Y, Zhao X, Zhang L et al (2020) Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9413–9422
6. Carion N, Massa F, Synnaeve G et al (2020) End-to-end object detection with transformers. In: European conference on computer vision. Springer, Cham, pp 213–229
7. Li X, Lai S, Qian X (2021) DBCFace: Towards pure convolutional neural network face detection. IEEE Trans Circ Syst Video Technol
8. Ren S, He K, Girshick R et al (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28:91–99
9. Wang Y, Xu Z, Wang X et al (2021) End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8741–8750
10. Lin K, Wang L, Luo K et al (2020) Cross-domain complementary learning using pose for multi-person part segmentation. IEEE Trans Circ Syst Video Technol 31(3):1066–1078
11. Dong J, Cong Y, Sun G et al (2020) Weakly-supervised cross-domain adaptation for endoscopic lesions segmentation. IEEE Trans Circ Syst Video Technol 31(5)
12. Rao Y, Zhao W, Zhu Z et al (2021) Global filter networks for image classification. Adv Neural Inf Process Syst:34
13. Yang B, Wang L, Wong DF et al (2021) Context-aware self-attention networks for natural language processing. Neurocomputing 458:157–169
14. Galassi A, Lippi M, Torroni P (2020) Attention in natural language processing. IEEE Trans Neural Netw Learn Syst
15. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
16. Woo S, Park J, Lee JY et al (2018) Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
17. Li X, Wang W, Hu X et al (2019) Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 510–519
18. Cao Y, Xu J, Lin S et al (2019) Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, pp 0–0
19. Wang Q, Wu B et al (2020) ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition
20. Wu L, Wang Y, Gao J et al (2018) Where-and-when to look: Deep siamese attention networks for video-based person re-identification. IEEE Trans Multimed 21(6):1412–1424
21. Chen G, Lu J, Yang M et al (2020) Learning recurrent 3D attention for video-based person re-identification. IEEE Trans Image Process 29:6963–6976
22. Guan J, Lai R, Xiong A et al (2020) Fixed pattern noise reduction for infrared images based on cascade residual attention CNN. Neurocomputing 377:301–313
23. Li J, Jin K, Zhou D et al (2020) Attention mechanism-based CNN for facial expression recognition. Neurocomputing 411:340–350
24. Liu W, Wu G, Ren F (2020) Deep multi-branch fusion residual network for insect pest recognition. IEEE Trans Cogn Dev Syst
25. Zheng Z, Yu Z, Wu Y et al (2021) Generative adversarial network with multi-branch discriminator for imbalanced cross-species image-to-image translation. Neural Netw 141:355–371
26. Hernández-Luquin F, Escalante HJ (2021) Multi-branch deep radial basis function networks for facial emotion recognition. Neural Comput and Appl:1–15
27. Szegedy C, Vanhoucke V, Ioffe S et al (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
28. Szegedy C, Ioffe S, Vanhoucke V et al (2017) Inception-v4 inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence
29. Liu W, Wu G, Ren F et al (2020) DFF-ResNet: An insect pest recognition model based on residual networks. Big Data Min Analytics 3(4):300–310
30. Ren F, Liu W, Wu G (2019) Feature reuse residual networks for insect pest recognition. IEEE Access 7:122758–122768
31. Lin TY, Dollár P, Girshick R et al (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125
32. Paszke A, Gross S, Massa F et al (2019) Pytorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:8026–8037
33. Selvaraju RR, Cogswell M, Das A et al (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
34. Wu Y, Jiang X, Fang Z et al (2021) Multi-modal 3D object detection by 2D-guided precision anchor proposal and multi-layer fusion. Appl Soft Comput 108:107405
35. Wang H, Peng J, Chen D et al (2020) Attribute-guided feature learning network for vehicle reidentification. IEEE MultiMedia 27(4):112–121
36. Wang H, Wang Y, Zhang Z et al (2020) Kernelized multiview subspace analysis by self-weighted learning. IEEE Trans Multimed

37. Wang H, Peng J, Zhao Y et al (2020) Multi-path deep CNNs for fine-grained car recognition. IEEE Trans Veh Technol 69(10):10484–10493
38. Wang H, Peng J, Jiang G et al (2021) Discriminative feature and dictionary learning with part-aware model for vehicle re-identification. Neurocomputing 438:55–62
39. Yan Z, Liu W, Wen S et al (2019) Multi-label image classification by feature attention network. IEEE Access 7:98005–98013

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Wenjie Liu was born in Hunan province, China, in 1989. He received the B.S. degree in information engineering from Nanhang Jincheng college, China, in 2011, the M.S. degree in information and communication engineering from Nantong University, China, in 2014, the Ph.D in intelligent information system engineering from Tokushima University. His research interests include Image Analysis, Computer Vision and Artificial Intelligence.

Guoqing Wu received the B.S. and M.S. degree in mechatronics from Jiangsu University, China, in 1983 and 1993 respectively, and the Ph.D. degree in mechanical design and theory from Shanghai University, China, in 2006. He is currently a Professor Sciences, Nantong University, China. His research interests are in the area of Mechanical Engineering, Laser Technology Application, and Artificial Intelligence.

Fuji Ren received his Ph. D. degree in 1991 from the Faculty of Engineering, Hokkaido University, Japan. From 1991 to 1994, he worked at CSK as a chief researcher. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an Associate Professor. Since 2001, he has been a Professor of the Faculty of Engineering, Tokushima University. His current research interests include Natural Language Processing, Artificial Intelligence, Affective Computing, Emotional Robot. He is the Academician of The Engineering Academy of Japan and EU Academy of Sciences. He is a senior member of IEEE, Editor-in-Chief of International Journal of Advanced Intelligence, a vice president of CAAI, and a Fellow of The Japan Federation of Engineering Societies, a Fellow of IEICE, a Fellow of CAAI. He is the President of International Advanced Information Institute, Japan.

Quan Shi was born in Nantong, China, in 1973. He is currently a Professor in the School of Transpiration, Nantong University, China. He has authored more than 60 papers, since 2007, of which more than 40 are peer-reviewed and well-known journal papers. His research is focused on the development of signal and image processing and big data techniques for computer vision.