


Article

# Spatial Channel Attention for Deep Convolutional Neural Networks

Tonglai Liu <sup>1,2,3,4,5,6,†</sup> , Ronghai Luo <sup>7,†</sup>, Longqin Xu <sup>1,2,3,4,5,6</sup>, Dachun Feng <sup>1,2,3,4,5,6</sup>, Liang Cao <sup>1,2,3,4,5,6</sup>, Shuangyin Liu <sup>1,2,3,4,5,6,\*</sup> and Jianjun Guo <sup>1,2,3,4,5,6,\*</sup>

- <sup>1</sup> College of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China; tonglailiu@zhku.edu.cn (T.L.); longqinxu@zhku.edu.cn (L.X.); dachunfeng@zhku.edu.cn (D.F.); cl@zhku.edu.cn (L.C.)
  - <sup>2</sup> Smart Agriculture Engineering Technology Research Center of Guangdong Higher Education Institutes, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China
  - <sup>3</sup> Guangzhou Key Laboratory of Agricultural Products Quality & Safety Traceability Information Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China
  - <sup>4</sup> Academy of Smart Agricultural Engineering Innovations, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China
  - <sup>5</sup> Guangdong Province Key Laboratory of Waterfowl Healthy Breeding, Guangzhou 510225, China
  - <sup>6</sup> College of Mechanical and Electric Engineering, Shihezi University, Shihezi 832000, China
  - <sup>7</sup> School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China; 20032303090@mails.guet.edu.cn
- \* Correspondence: shuangyinliu@zhku.edu.cn (S.L.); guojianjun@zhku.edu.cn (J.G.)  
† These authors contributed equally to this work.

**Abstract:** Recently, the attention mechanism combining spatial and channel information has been widely used in various deep convolutional neural networks (CNNs), proving its great potential in improving model performance. However, this usually uses 2D global pooling operations to compress spatial information or scaling methods to reduce the computational overhead in channel attention. These methods will result in severe information loss. Therefore, we propose a Spatial channel attention mechanism that captures cross-dimensional interaction, which does not involve dimensionality reduction and brings significant performance improvement with negligible computational overhead. The proposed attention mechanism can be seamlessly integrated into any convolutional neural network since it is a lightweight general module. Our method achieves a performance improvement of 2.08% on ResNet and 1.02% on MobileNetV2 in top-one error rate on the ImageNet dataset.

**Keywords:** attention mechanism; image classification; deep learning; cross-dimensional interaction

**MSC:** 68T07



**Citation:** Liu, T.; Luo, R.; Xu, L.; Feng, D.; Cao, L.; Liu, S.; Guo, J. Spatial Channel Attention for Deep Convolutional Neural Networks. *Mathematics* **2022**, *10*, 1750. <https://doi.org/10.3390/math10101750>

Academic Editor: Catalin Stoean

Received: 18 April 2022

Accepted: 19 May 2022

Published: 20 May 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

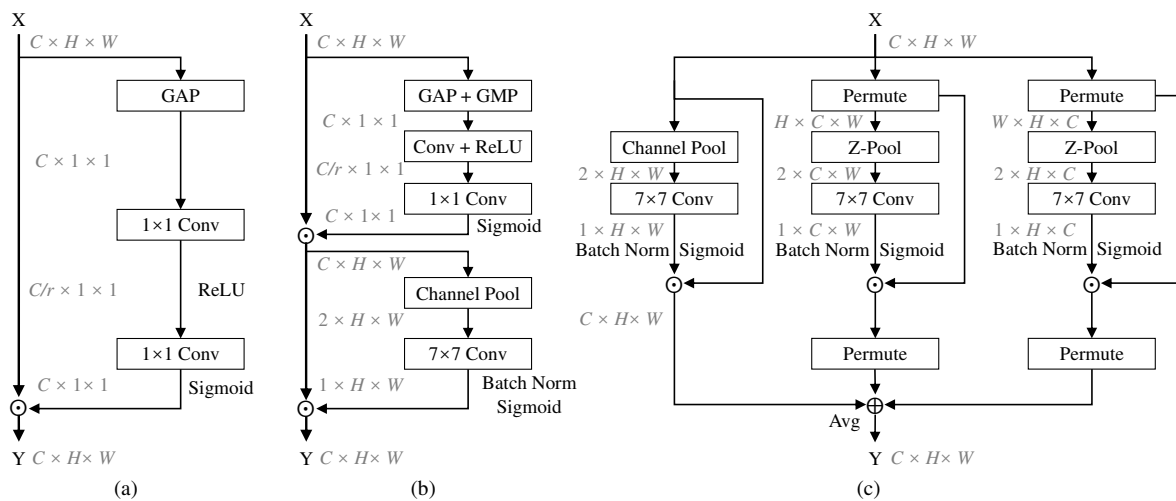
Recently, attention mechanisms have attracted extensive research in natural language processing (NLP) [1], computer vision (CV) [2], and speech signal processing (SSP) [3], including spatial attention [1,2,4–6], channel attention [7,8], and spatial and channel attention [9–13], as they can easily improve the performance of neural networks by recalibrating the weights of features. These mechanisms enable the network to learn where or what to pay attention to by explicitly building cross-channel dependencies or weighted spatial regions.

SENet [7], as one of the state-of-the-art channel attention mechanisms, provides significant performance gains at an extremely low computational cost. However, SE attention only considers the encoded inter-channel information and ignores the importance of spatial information, which is crucial for image classification. The convolutional block attention module (CBAM) [10] provides robust representative attention by combining channel information and spatial information. Compared with SENet, the CBAM offers significant

performance improvements with a small computational overhead. However, in CBAM, spatial attention is obtained by simply global average pooling (GAP) and global max pooling (GMP), compressing channel  $C$  into a single channel to obtain important spatial information. Similarly, coordinate attention (CA) [12] uses GAP to compress, respectively, height  $H$  and width  $W$  to capture the interaction between spatial information and channel information.

Wang [8] pointed out that it is important to avoid dimensionality reduction and proper cross-channel interaction when learning channel attention. Therefore, Misra [13] proposed a lightweight non-dimensionality reduction attention mechanism, triplet attention (TA). TA captures the interaction between the spatial dimension and channel dimension through a rotation operation to enhance feature representations.

Provided by Misra [13], the SE module, CBAM, and TA are shown in Figure 1a–c, respectively.



**Figure 1.** Comparisons with different attention modules [13]: (a) squeeze excitation (SE) module; (b) convolutional block attention module (CBAM); (c) triplet attention (TA).  $\oplus$  denotes broadcast elementwise addition, and  $\odot$  denotes broadcast elementwise multiplication.

However, triplet attention uses two branches to capture the cross-dimensional interaction of channel  $C$  with height  $H$  and width  $W$ , respectively, and uses the third branch to capture spatial information, which is unnecessary; furthermore, it increases the complexity of the model. Therefore, we propose a simpler, but better performing attention mechanism to capture cross-dimensional interaction information, which does not involve dimensionality reduction, namely Spatial channel attention (SCA).

Specifically, to capture cross-dimensional interaction and alleviate the spatial information loss caused by GAP or GMP, we aggregate the cross-dimensional interaction features between the spatial dimensions  $H$  or  $W$  with the channel dimension  $C$  by simply permuting the input tensors. Then, we feed them into convolutional layers and Sigmoid activation layers to generate two interaction attention maps, respectively. Finally, we permute back attention maps to the original input shape and apply them to the input tensors via multiplication.

Our Spatial channel attention has the following advantages. First, it emphasizes the importance of cross-dimensional interaction to capture not only orientation-aware channel information, but also channel-sensitive spatial information, which helps the model to locate and identify objects of interest more accurately. Second, our method is flexible and lightweight, capturing rich discriminative feature representations with negligible computational overhead, classic convolutional neural network building blocks that can be easily inserted, such as ResNet [14] and MobileNetV2 [15], by emphasizing information representation to enhance functionality.

The proposed method considers cross-dimensional dependencies, which are computationally efficient and inexpensive. For example, for ResNet-50 [14] with 25.557M parameters and 4.122 GFLOPs, the proposed Spatial channel attention mechanism achieves a 2.08% improvement on top-one error rate at the cost of 4.6K more parameters and  $4.1 \times 10^{-2}$  more GFLOPs.

We propose a Spatial channel attention mechanism to improve the performance of CNNs for image classification. The remainder of this paper is organized as follows. In Section 2, recent efforts in attention mechanisms are described for image classification. In Section 3, the working details of the proposed method SCA are explicitly introduced. In Section 4, extensive experiments are conducted to assess the performance of SCA. In Section 5, our work is summarized.

## 2. Related Work

Attention mechanisms originate from the human visual system where humans selectively concentrate on regions of interest while ignoring the rest. Therefore, attention mechanisms are extensively studied in computer vision tasks, such as image classification [16–18], object detection [4,19], and image segmentation [5,6,11], aiming to tell a model *where* and *what* to attend to for boosting the performance of deep convolutional neural networks (CNNs). In this section, we review some attention mechanisms that are closely related to our work.

Attention mechanisms adaptively recalibrate the weights of features to improve the information perception ability of the model. According to the feature dimension applied, attention can be categorized into various types of variants, such as spatial attention and channel attention.

To improve the capability of modeling spatial information in CNNs, spatial attention is widely used with great success. The non-Local module [4] computes the relationship between a pixel and all other pixels to capture the long-range dependencies in images. However, the computational overhead of the non-local module is expensive. In order to reduce the amount of computation, GCNet [19] uses  $1 \times 1$  convolution and scaling operations, and CCNet [5] uses criss-cross attention modules in a cascading manner to aggregate the information on rows and columns of pixels. DANet [11] compresses the 3D tensor to 2D and captures spatial information through matrix multiplication. Similarly, SPNet [6] uses strip pooling on the height and width of features separately and generates spatial attention maps through matrix multiplication.

Channel attention assigns weights to different channels to tell the model what to focus on, which is simple and effective. SENet [7] was the first to propose an efficient method for channel attention, providing significant performance improvements at a minimal additional computational cost. SENet compresses each 2D spatial feature to generate channel weights, explicitly establishing the interdependencies between channels. To balance the paradox between model performance and complexity, Wang [8] proposed an attention without dimensionality reduction, efficient channel attention (ECA), which can bring significant performance gains by adding only a handful of parameters.

However, these channel attention methods only consider inter-channel interdependencies and ignore spatial information. Therefore, the CBAM [10] combines the channel attention and spatial attention to recalibrate the weights of features. The CBAM sequentially extracts attention maps along channels and spatial information and multiplies them with input feature maps to achieve adaptive feature augmentation.

Similarly, CA [12] utilizes two branches in parallel to extract cross-dimensional interaction between channel  $C$  with height  $H$  and width  $W$  to generate an attention map. Furthermore, TA [13] proposes three-branch attention without dimensionality reduction, where two branches capture cross-dimensional interaction and the third branch is used to build spatial attention.

### 3. Method

Spatial channel attention can be viewed as a computational unit that aims to enhance the expressive power of the learned features for mobile networks. It can take any intermediate feature tensor  $\mathbf{X}$  as the input and outputs a transformed tensor with augmented representations  $\mathbf{Y}$ . To provide a clear description of the proposed Spatial channel attention, we first revisit the channel attention in the CBAM, which is widely used in convolutional neural networks.

#### 3.1. Revisiting CBAM

We first revisit the channel attention module and spatial attention module used in the CBAM [10] in this subsection. Let  $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$  be the input of the CBAM channel attention module, where  $C$ ,  $H$ , and  $W$  denote the number of channels and the height and width of the feature map, respectively.

The channel attention weight in the CBAM can be expressed by the following equation:

$$\omega_c = \sigma(\mathbf{W}_c^1 \text{ReLU}(\mathbf{W}_c^0 \text{GAP}_c(\mathbf{X})) + \mathbf{W}_c^1 \text{ReLU}(\mathbf{W}_c^0 \text{GMP}_c(\mathbf{X}))) \quad (1)$$

where  $\omega_c$  represents the channel attention weight,  $\sigma$  is the Sigmoid activation function, ReLU is another activation function, and  $\mathbf{W}_c^1$  and  $\mathbf{W}_c^0$  are weight matrices, whose sizes are defined as  $C \times C/r$  and  $C/r \times C$ , respectively.  $\text{GAP}_c$  and  $\text{GMP}_c$  are the global average pooling function and global max pooling function of the channel, respectively.

Similarly, the spatial attention weight in the CBAM can be expressed by the following equation:

$$\omega_s = \sigma(\text{Conv}^{7 \times 7}[\text{GAP}_s(\mathbf{X}); \text{GAP}_s(\mathbf{X})]) \quad (2)$$

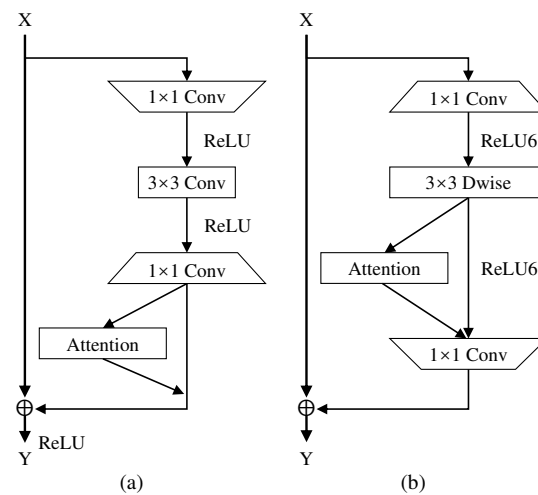
where  $\text{Conv}^{7 \times 7}$  represents a convolution operation with a filter size of  $7 \times 7$ ,  $[\cdot; \cdot]$  represents the tensor concatenation operation, and  $\text{GAP}_s$  and  $\text{GMP}_s$  are the global average pooling function and global max pooling function of the channel, respectively.

The CBAM is widely used in convolutional neural networks and has proven to be efficient. However, the CBAM compresses channel  $C$  into a single channel to obtain important spatial information and compresses spatial dimension  $H \times W$  into a single pixel to obtain channel importance, ignoring the cross-dimensional interaction information between spatial and channel dimensions. Therefore, in the following, we introduce a novel attention block that considers both channel and spatial cross-dimensional information to enhance feature representation.

#### 3.2. Spatial Channel Attention

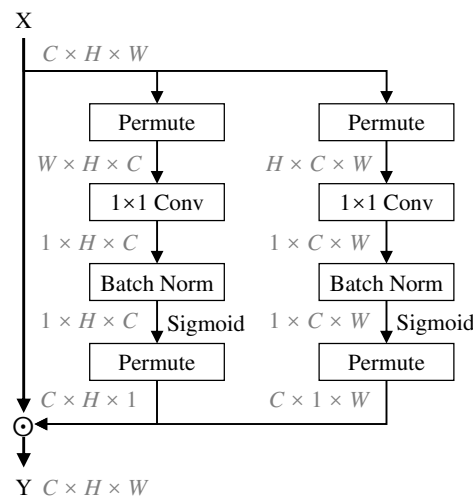
As shown in Section 1, we propose an attention mechanism with few parameters to capture the interaction between spatial and channel dimensions, namely Spatial channel attention. Figure 2a,b show how we insert the attention mechanism into the residual block in ResNet and the inverted residual block in MobileNetV2, respectively.

The traditional way to compute spatial attention is to compress the channels of the input tensor to generate a weight for each region on the spatial dimension. This can lead to incorrectly assigning higher weights to non-target regions. Similarly, the traditional way to compute channel attention is to compress the spatial information of the input tensor to generate a weight for each channel via global average pooling. This results in a severe loss of spatial information. Furthermore, the interdependence between these spatial attention and channel attention methods is non-existent. In simple terms, channel attention tells *which channel to focus on*, while spatial attention tells *where to focus on in the channel*. The disadvantage of this process is that channel attention and spatial attention are separated and computed independently of each other, without considering any relationship between them. Therefore, we propose a Spatial channel attention mechanism that captures cross-dimensional interaction.



**Figure 2.** Connection implementation between the proposed attention mechanism and CNNs, where  $\oplus$  denotes broadcast elementwise addition.

The proposed Spatial channel attention is shown in Figure 3. There are two branches in Spatial channel attention, which are responsible for capturing the cross-dimensional interaction between channel dimension  $C$  and spatial dimension  $H$  or  $W$ , respectively. SCA will rearrange the input tensors through the permutation operation. Then, they are sequentially input to the convolutional layer and the Sigmoid activation layer to generate two attention maps for cross-dimensional interaction, respectively. Finally, attention maps are permuted again and applied to the input tensor via multiplication, obtaining feature representations that are interactively enhanced across dimensions.



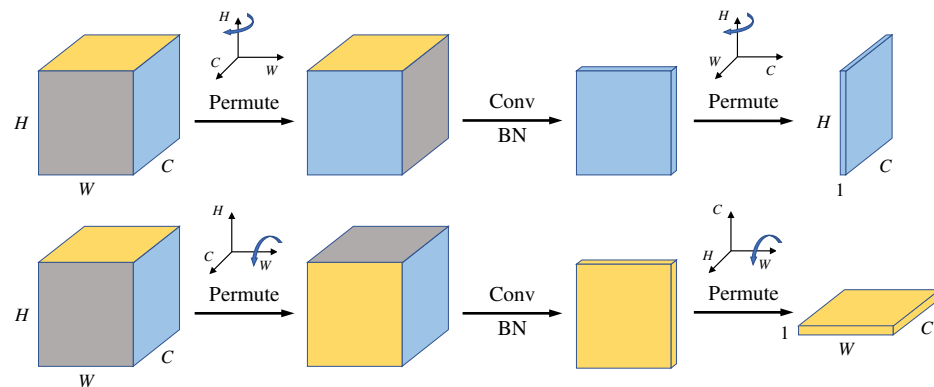
**Figure 3.** Spatial channel attention mechanism (SCA), where  $\odot$  denotes broadcast elementwise multiplication.

As shown in Figure 4, given an input tensor  $X$  with dimension  $C \times H \times W$ , we first pass it to the two branches in the proposed attention module. In the top branch, we construct the interaction between the height dimension  $H$  and the channel dimension  $C$ . To do this, we permute  $X$  and rearrange to  $W \times H \times C$ . Next, the rearranged tensor is successively fed to a convolutional layer with a kernel size of  $1 \times 1$  and a batch normalization layer, and an intermediate output with dimension  $1 \times H \times C$  is obtained. Then, we input it in

the Sigmoid activation function and obtain the attention weights. Finally, we permute the newly generated attention weights again to rearrange them as  $C \times H \times 1$ .

$$F_{ch}(\mathbf{X}) = \text{Permute}_{ch}(\sigma(\text{Conv}^{1 \times 1} \text{Permute}_{ch}(\mathbf{X}))) \tag{3}$$

where  $\text{Permute}_{ch}$  represents the operation of permuting the channel dimension  $C$  and the spatial dimension  $H$  and  $\text{Conv}^{1 \times 1}$  represents a convolution operation with a filter size of  $1 \times 1$ .



**Figure 4.** The attention-map-generation process of SCA. The top branch represents the cross-dimensional interaction between  $C$  and  $H$ , and the bottom branch represents the cross-dimensional interaction between  $C$  and  $W$ .

Likewise, in the bottom branch, we rearrange  $\mathbf{X}$  into a dimensional representation of  $H \times C \times W$ . Then, we feed it sequentially to a convolutional layer with a filter size of  $1 \times 1$ , a batch normalization layer, and a Sigmoid activation function, obtaining attention weights of shape  $1 \times C \times W$ . Finally, we permute the newly generated attention weights again, rearranging them as  $C \times 1 \times W$ .

$$F_{cw}(\mathbf{X}) = \text{Permute}_{cw}(\sigma(\text{Conv}^{1 \times 1} \text{Permute}_{cw}(\mathbf{X}))) \tag{4}$$

Finally, we apply the two newly generated attention weights to the input tensor  $\mathbf{X}$  via broadcast elementwise multiplication. A tensor  $\mathbf{Y}$  weighted with cross-dimensional interaction can be obtained.

$$\mathbf{Y} = \mathbf{X} \odot F_{ch}(\mathbf{X}) \odot F_{cw}(\mathbf{X}) \tag{5}$$

where  $F_{ch}(\mathbf{X})$  represents the interaction function between the channel dimension  $C$  and the height dimension  $H$  and  $F_{cw}(\mathbf{X})$  represents the interaction function between the channel dimension  $C$  and the width dimension  $W$ .

#### 4. Results and Analysis

In this section, we first introduce our experimental settings. Then, the proposed method is evaluated on ImageNet-1K [20] for image classification based on ResNet-50 [14] and MobileNetV2 [15]. Ablation experiments were conducted to verify the effectiveness of cross-dimensional interaction. Finally, sample visualizations are provided from Grad-CAM to demonstrate the effectiveness of the proposed method in locating and identifying objects of interest.

##### 4.1. Experimental Setup

For the fairness of comparisons, we followed the training configuration of ResNet. Likewise, we followed the training configuration and data augmentation method in [15] to implement our MobileNetV2-based architecture. We used the Adam [21] optimizer and cosine learning schedule, training on a 1 Nvidia Tesla P100 GPU.



#### 4.2. Comparative Experiment

The results of the comparative experiments are shown in Table 1. Spatial channel attention introduces the fewest parameters while consistently outperforming other attentions.

**Table 1.** Comparative experimental results.

Method	Backbone	Parameters	FLOPs	Top-1(%)	Top-5(%)
ResNet-50		<b>25.56M</b>	<b>4.12G</b>	24.56	7.50
+ SENet		28.07M	4.13G	23.14	6.70
+ CBAM	ResNet-50	28.09M	4.13G	22.66	6.31
+ TA		25.56M	4.17G	22.52	6.32
+ SCA (ours)		25.56M	4.16G	<b>22.48</b>	<b>6.30</b>
MobileNetV2		<b>3.51M</b>	<b>0.32G</b>	28.36	9.80
+ SENet		3.53M	0.32G	27.58	9.33
+ CBAM	MobileNetV2	3.54M	0.32G	30.07	10.67
+ TA		3.51M	0.32G	27.38	9.23
+ SCA (ours)		3.51M	0.32G	<b>27.34</b>	<b>9.21</b>

The ResNet-50-based model improved the top-one error rate on ImageNet by 2.08%, while the number of parameters only increased by 0.02%, and the FLOPs increased by about 1%. Spatial channel attention outperformed SENet and CBAM with 0.66% and 0.18% improvements in top-one error rates, respectively. The main reason is that the proposed method considers spatial information and does not use GAP or GMP to reduce dimensionality, which avoids information loss. Furthermore, our method also outperformed TA by a small margin.

We observed similar performance trends in the smaller MobileNetV2 model. Compared with MobileNetV2, using Spatial channel attention improved top-one error rate by 1.02%, while increasing parameter complexity by only 0.03%. Spatial channel attention outperformed SENet with a 0.24% improvement in top-one error rate, respectively. We also observed that in the case of MobileNetV2, the CBAM hurt the model performance: it reduced the accuracy by 1.71%. Experimental results showed that the proposed Spatial channel attention worked in both heavyweight and lightweight models with negligible increases in parameters and computation. Furthermore, our method also slightly outperformed TA. The main reason is that the proposed method utilizes multiplication to fuse the interaction of different dimensions, rather than simple addition, which results in the information of different dimensions being treated equally.

#### 4.3. Ablation Studies

To investigate the importance of the interaction between different spatial dimensions and channel dimensions, we observed changes in performance by clipping different branches. In Table 2, CHA represents the top branch and CWA represents the bottom branch in Figure 4. As shown in Table 2, the experimental results showed that Spatial channel attention consistently outperformed the baseline model and its two counterparts.

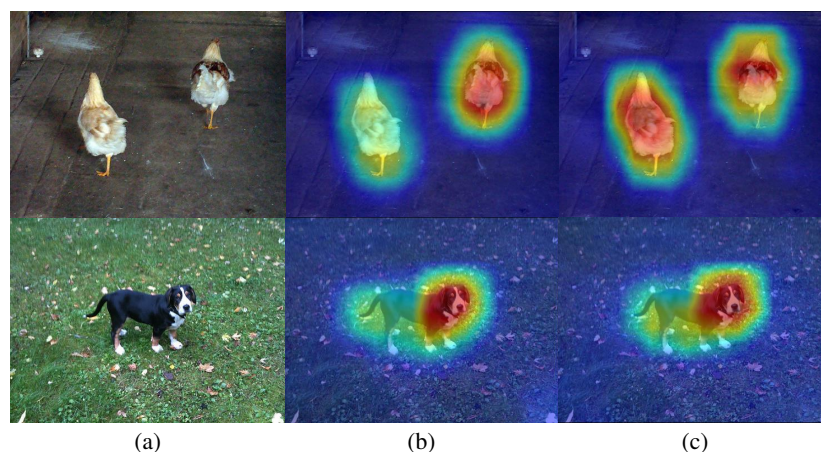
**Table 2.** Effectiveness experiment.

Method	Parameters	FLOPs	Top-1(%)	Top-5(%)
ResNet-50	<b>25.56M</b>	<b>4.12G</b>	24.56	7.50
+ CHA	28.07M	4.14G	23.15	6.92
+ CWA	25.92M	4.14G	23.09	6.85
+ SCA (ours)	25.56M	4.16G	<b>22.48</b>	<b>6.30</b>
MobileNetV2	<b>3.51M</b>	<b>0.32G</b>	28.36	9.80
+ CHA	3.53M	0.32G	27.62	9.47
+ CWA	3.54M	0.32G	27.68	9.53
+ SCA (ours)	3.51M	0.32G	<b>27.34</b>	<b>9.21</b>

#### 4.4. Grad-CAM Visualization

To evaluate the effectiveness of the proposed method in locating and identifying objects of interest, sample visualizations are provided utilizing the Grad-CAM [22] techniques. Figure 5 is the gradient visualization based on Resnet-50, where the ground-truth label at the top is cock and the bottom is the Greater Swiss Mountain dog.

The visualizations shows that the proposed method can help to locate target objects and improve the performance of CNNs.



**Figure 5.** Visualization of Grad-CAM results based on ResNet-50; (a) original images; (b) ResNet-50; (c) ResNet-50 with SCA.

## 5. Conclusions

In this paper, we proposed an attention mechanism that does not involve dimensionality reduction, namely Spatial channel attention (SCA). SCA can capture cross-dimensional interaction information, including direction-aware channel information and channel-sensitive spatial information, which can help to improve the performance of the model image classification. Because SCA is a lightweight general module, it can be flexibly plugged into any convolutional neural network.

In the future, we plan to replace standard convolution with dilated convolution in the proposed method, aiming to reduce the computational cost while increasing the receptive field. We intend to apply SCA to object detection and other visual tasks and consider adding time-aware information to adapt to video-related tasks.

**Author Contributions:** Conceptualization, L.X.; Data curation, L.C.; Investigation, D.F.; Methodology, J.G.; Validation, S.L.; Writing—original draft, T.L. and R.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This paper was supported partly by the National Natural Science Foundation of China under Grant 61871475, in part by the Special Project of Laboratory Construction of Guangzhou Innovation Platform Construction Plan under Grant 201905010006, Guangzhou Key Research and Development Project under Grant 202103000033, 201903010043, Guangdong Science and Technology Project under Grant 2020A1414050060, 2020B0202080002, Innovation Team Project of Universities in Guangdong Province under Grant 2021KCXTD019, Characteristic Innovation Project of Universities in Guangdong Province under Grant KA190578826, Guangdong Province Enterprise Science and Technology Commissioner Project under Grant GDKTP2021004400, Guangdong Science and Technology Planning Project under Grant 2016A020210122, Meizhou City S&T Planed Projects under Grant 2021A0305010, Rural Science and Technology Correspondent Project of Zengcheng District, Guangzhou City under Grant 2021B42121631, Educational Science Planning Project of Guangdong Province under Grant 2020GXJK102, and Grant 2018GXJK072.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.



**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; pp. 5998–6008.
2. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021.
3. Chen, S.; Zhang, M.; Yang, X.; Zhao, Z.; Zou, T.; Sun, X. The Impact of Attention Mechanisms on Speech Emotion Recognition. *Sensors* **2021**, *21*, 7530. [[CrossRef](#)] [[PubMed](#)]
4. Wang, X.; Girshick, R.B.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803. [[CrossRef](#)]
5. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. CCNet: Criss-Cross Attention for Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, 27 October–2 November 2019; pp. 603–612. [[CrossRef](#)]
6. Hou, Q.; Zhang, L.; Cheng, M.; Feng, J. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 4002–4011. [[CrossRef](#)]
7. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [[CrossRef](#)]
8. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539. [[CrossRef](#)]
9. Park, J.; Woo, S.; Lee, J.; Kweon, I.S. BAM: Bottleneck Attention Module. In Proceedings of the British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, 3–6 September 2018; BMVA Press: Newcastle, UK, 2018; p. 147.
10. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Lecture Notes in Computer Science, Proceedings of the Computer Vision—ECCV 2018—15th European Conference, Munich, Germany, 8–14 September 2018*; Proceedings, Part VII; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Berlin, Germany, 2018; Volume 11211, pp. 3–19. [[CrossRef](#)]
11. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154. [[CrossRef](#)]
12. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; pp. 13713–13722.
13. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to Attend: Convolutional Triplet Attention Module. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, 3–8 January 2021; pp. 3138–3147. [[CrossRef](#)]
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
15. Sandler, M.; Howard, A.G.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520. [[CrossRef](#)]
16. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, NV, USA, 3–6 December 2012; Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; pp. 1106–1114.
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; Conference Track Proceedings; Bengio, Y., LeCun, Y., Eds.
18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; pp. 1–9. [[CrossRef](#)]
19. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea, 27–28 October 2019; pp. 1971–1980. [[CrossRef](#)]
20. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; pp. 248–255. [[CrossRef](#)]

21. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; Conference Track Proceedings; Bengio, Y., LeCun, Y., Eds.
22. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 618–626. [[CrossRef](#)]