# Research on the Application of Improved Attention Mechanism in Image Classification and Object Detection

Y. Y. Ding, L. Wang

**Abstract — In recent years, with the advancement of computer processing power and the rapid development of convolutional neural networks, attention models have been widely used in image classification and object detection, significantly improving the performance of networks. However, most existing methods focus on increasing the depth of networks, which inevitably leads to efficiency issues. To strike a balance between performance and complexity, this paper proposes an optimization based on Triplet Attention, introducing multi-scale attention fusion Triplet Attention and self-attention mechanisms. Firstly, the proposed model is integrated into YOLOv5s and tested on VOC2012, achieving a 3.9% increase in mean average precision (mAP). Then, the model is integrated into ResNet50 and tested on CIFAR-10 and CIFAR-100 datasets. The results show that the proposed model achieves a 1.79% improvement in Top-1 accuracy and a 0.13% improvement in Top-5 accuracy on CIFAR-10,a 3.11% improvement in Top-1 accuracy, and a 1.3% improvement in Top-5 accuracy on CIFAR-100. The experimental results demonstrate that the proposed model significantly enhances network performance.**

***Index Terms*—Triplet attention,multi-scale attention fusion,self-attention,YOLOv5s,ResNet50**

## I. Introduction

In the realm of deep learning, the attention mechanism plays a crucial role in guiding convolutional neural networks to emphasize the acquisition of relevant features while discouraging the assimilation of irrelevant ones. Leveraging its plug-and-play nature, this mechanism has evolved into an indispensable module within the convolutional network model.

Deep learning has extensively embraced the employment of attention mechanisms to steer the learning process towards targeted features and effectively suppress extraneous characteristics in convolutional neural networks. Consequently, the attention mechanism has become an essential component of convolutional network models. Additionally, apart from augmenting the depth, width, and cardinality of networks [1], network performance can be further enhanced by customizing the attention module within the architectural design. Presently, attention mechanisms are progressing towards avenues such as multiple branches and lightweight design [2]. The attention mechanism primarily focuses on crucial features while reducing attention to irrelevant ones, thereby enhancing performance. In the realm of artificial intelligence, the attention mechanism plays a crucial role in various domains such as natural language processing and computer vision. It serves as an integral component of neural network architectures and is an indispensable research area. Li et al. proposed the SK attention mechanism [3], where the output feature maps V1 and V2 in the SK module are obtained by multiplying the feature maps obtained from convolutional kernels of different sizes by the a and b vectors obtained through the Softmax function, allowing for dynamic adjustment of the receptive field size. Hu et al. proposed the SE attention mechanism [4], which can be divided into two stages in terms of model structure, namely the compression stage and the excitation stage. The compression stage embeds global information to obtain the channel attention vector, and the excitation stage enables adaptive weights between channels through inter-channel information exchange. SE has been widely used in various network models to improve the feature extraction ability of the models, although it considers different importance levels of features between channels, it neglects positional information. The CA attention mechanism proposed by Hou et al. [5] captures not just cross-channel data but also direction-aware and position-sensitive information, enhancing the model's ability to precisely identify and locate objects of interest.The Convolutional Block Attention Module (CBAM) [6] reduces the input tensor's channel dimension and employs convolution to calculate spatial attention, thus making use of positional information. However, convolution is limited to capturing local relationships and cannot effectively model critical long-range dependencies in visual tasks. Recently, $A^2$-Nets [7] introduced novel relational functions for non-local blocks, enabling these introduced non-local blocks to capture long-range dependency relationships through non-local operations. This approach makes them lightweight and readily applicable in various architectural contexts.

Despite the attention models discussed above enhancing the feature extraction capability of images from various perspectives and angles, they also increase the computational cost of the networks. In response to these challenges, this paper introduces an attention model called SCSS, which integrates multi-scale fusion of channel, spatial, and self-attention mechanisms [8]. The SCSS module is first integrated into the YOLOv5s object detection model and

Manuscript received April 19, 2023; revised September 5, 2023.This work was supported in part by the National Natural Science Foundation of China under Grant 71472081.

Y. Y. Ding is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (e-mail: 2401840089@qq.com).

L. Wang is a Professor at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (e-mail: wangli9966@163.com).

experiments are conducted on the VOC2012 dataset [9]. The results show that the improved module in this paper achieves significant performance improvement compared to the SE module and CBAM module in the object detection task. To delve deeper into the module's performance, we have also conducted analogous comparative experiments in the realm of image classification. The SE module, CBAM module, and the improved module in this paper are integrated into the Resnet50 classification model, and experiments are conducted on the CIFAR-10 [10]and CIFAR-100 [10] datasets. The experimental results demonstrate that compared to the CA module, CBAM module, and SE module, the improved module in this paper is highly effective and achieves higher classification accuracy.

## II. The Introduction of Attention Model

### A. Squeeze-and-Excitation(SE)

The SE [4] structure is divided into two parts: the compression stage and the excitation stage, as illustrated in Fig.1.

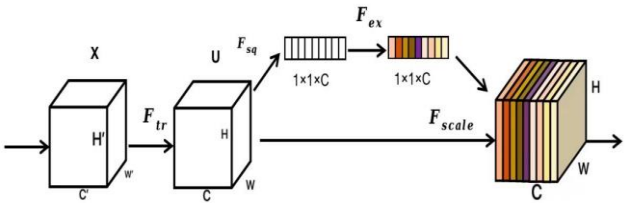



Fig. 1. The structure of Squeeze-and-Excitation(SE).

The purpose of introducing SE is to address the issue of the varying importance of different channels in the feature maps during the convolutional pooling process. The compression stage embeds global information to obtain channel attention vectors, while the excitation stage facilitates inter-channel information exchange to obtain adaptive weights between channels. The specific implementation is as follows: in the compression stage, providing a feature map $\boldsymbol{x}\in\mathrm{R}^{C\times H\times W}$, where C represents the number of feature channels, H represents height, and W represents width.

The input x undergoes average pooling using a (H, W) kernel size, effectively performing channel-wise average pooling, allowing the convolutional network to capture global information. The output is denoted as $\boldsymbol{Z}\in\mathrm{R}^{C\times 1\times 1}$, and Z is calculated using the formula (1) as shown below:

$$Z=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{w}x_{(i,j)} \tag{1}$$

In the formula, H denotes the feature map's height, W signifies its width, and x corresponds to the feature channel. The excitation phase facilitates cross-channel information exchange through a pair of fully connected layers. The channel attention vector Z is utilized as input for these two fully connected layers, resulting in an output denoted as U. The calculation process for U is elaborated in formula (2) as follows:

$$U=\sigma(W_2\delta(W_1\cdot Z)) \tag{2}$$

In the equation, $W_1$ represents the parameters of the first fully connected layer, and $W_2$ represents the parameters of the second fully connected layer. When employing two fully connected layers, the initial one aims to minimize parameter count by channel compression, typically by a factor denoted as r, representing the compression ratio. $\sigma$ represents the ReLU activation function, $\delta$ represents the Sigmoid activation function, and the channel attention vector U is element-wise multiplied with the input x along the channel dimension to obtain the output Y.

### B. Convolutional Block Attention Module(CBAM)

CBAM [6] is a simple and effective convolutional neural network attention module. Traditional attention mechanisms based on convolutional neural networks mainly focus on analyzing interactions in the channel domain, limited to considering the relationship between feature map channels. CBAM introduces spatial attention and channel attention from two action domains, realizing a sequential attention structure from channels to space. Fig.2 illustrates the architecture.

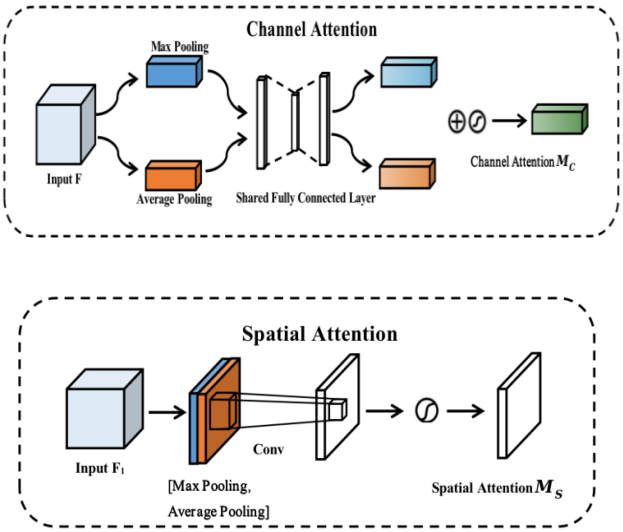



Fig. 2. The structure of Convolutional Block Attention Module(CBAM).

Spatial attention enables the neural network to prioritize essential image pixel regions for classification while ignoring irrelevant areas. Channel attention is employed for managing the distribution of feature map channels, extending attention allocation to both dimensions and thereby amplifying the enhancement in the attention mechanism's performance. In a convolutional neural network, given an intermediate feature map, CBAM incorporates attention maps independently along both the channel and spatial dimensions. These attention maps are subsequently multiplied with the input feature map to dynamically enhance the input features. Due to its nature as an end-to-end general module, CBAM can be effortlessly incorporated into various CNN architectures and trained in conjunction with base CNNs, maintaining a seamless integration throughout. When provided with an intermediate feature map $\boldsymbol{F}\in\mathrm{R}^{C\times H\times W}$ as input, the CBAM module's operation can be split into two distinct stages: initially, the input undergoes global max-pooling and average-pooling operations along the channel dimension, yielding one-dimensional vectors. These vectors are then inputted into fully connected layers and summed to produce a one-dimensional channel attention map $\boldsymbol{M_C}\in\mathrm{R}^{C\times 1\times 1}$. Then, the channel attention map is multiplied element-wise with the input to obtain the feature map $F_1$ adjusted by channel

attention. Second, $F_1$ is globally max-pooled and average-pooled along the spatial dimension, the resulting two-dimensional vectors are concatenated and convolved to generate a two-dimensional spatial attention map $\boldsymbol{M_S} \in R^{1\times H\times W}$. Finally, the spatial attention map is multiplied element-wise with $F_1$. The overall process of generating attention maps by CBAM can be described by formulas (3) and (4):

$$F_1 = M_c(F) \otimes F \tag{3}$$

$$F_2 = M_s(F_1) \otimes F_1 \tag{4}$$

### C. *Self-attention mechanism*

In the field of deep learning, the use of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) has become a common strategy for sequence encoding. However, for sequences that encompass intricate relationships and dependencies, the self-attention mechanism emerges as a powerful tool. This mechanism elegantly performs the act of attention pooling on sequences composed of word tokens. In this orchestration, a unique and fascinating feature is unveiled: the very same set of word tokens dons the roles of queries, keys, and values, all concurrently.

Delving deeper, each query embarks on a journey to engage with every key-value pair within the sequence, culminating in the generation of an insightful attention output. What adds an additional layer of intrigue is the fact that queries, keys, and values are intrinsically birthed from the same repository of input data. This symphony of interplay, encapsulating the self-attentive dance, gives rise to the term "self-attention" or, more broadly, "intra-attention."

The canvas upon which this intricate dance unfolds is vividly depicted in Fig.3. Here, the self-attention mechanism is encapsulated in a schematic representation. The input to this mechanism is a vector of substantial length, denoted as N (with N being a variable factor). As the self-attention mechanism orchestrates its intricate symphony, a transformative feat is achieved: the output mirrors the input in terms of length, emerging as a vector of equal length, N.

In essence, this self-attention mechanism is akin to a conductor orchestrating a symphony of interplay, harmonizing the inputs' inherent relationships into a melodious output of understanding. Just as a symphony's crescendo builds layer upon layer, so too does the self-attention mechanism weave a tapestry of interconnected insights, enriching the landscape of deep learning and sequence analysis.

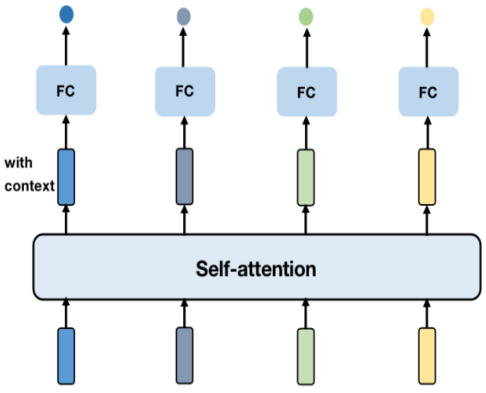



Fig. 3. The structure of self attention.

## III. Improved Strategy

In this study, we present a refined attention mechanism called SCSS, which combines channel, spatial, and self-attention mechanisms, along with a multi-scale feature fusion module [11-12] integration enables the establishment of global dependency relationships, expansion of the receptive field for capturing broader contextual information, and efficient extraction of shallow features at various scales from the original image to capture comprehensive image information. Refer to Fig.4 and Fig.5 for the schematic diagrams illustrating this concept.

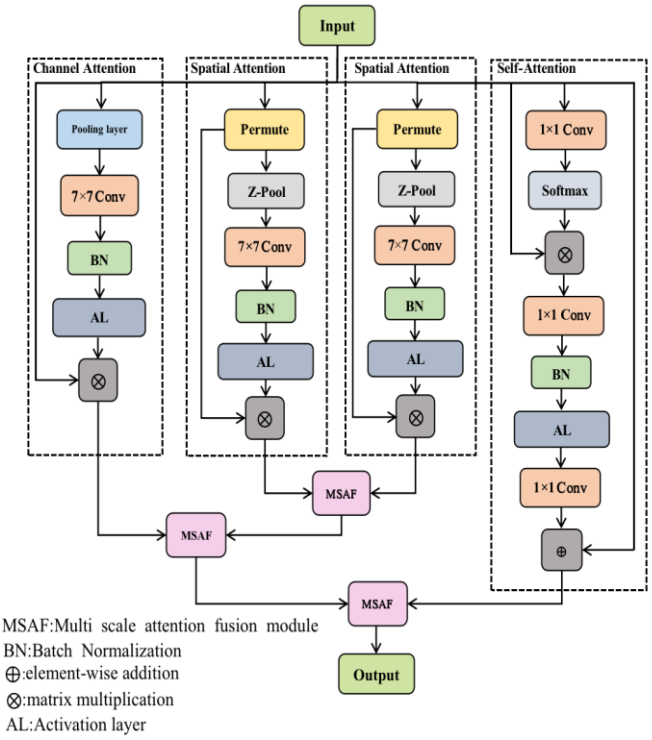



Fig. 4. The structure of improved attention mechanism.

The SCSS employs a four-branch structure that not only disregards computational overhead but also underscores the significance of multidimensional interactions without compromising dimensionality, thereby obviating indirect correspondence between channels and weights. The initial branch, dedicated to channel attention calculation, takes the input feature map and subjects it to a channel pooling layer, by performing convolutional operations and passing through a Sigmoid activation function, the required channel attention weights are ultimately generated. The second and third branches capture interactions between C and W. Initially, the input features are permuted to the H×C×W dimension, followed by a Z-Pool operation along the H dimension, repeated in successive steps. Subsequently, the features are reorganized to the C×H×W dimension for element-wise addition. To enhance feature integration amidst semantic and scale disparities, a multi-scale attention fusion module is appended at the conclusion, addressing the bottleneck in initial feature fusion. The introduction of the self-attention mechanism caters to effective modeling of extensive dependencies, reminiscent of the simplified non-local (SNL)

block, all while preserving a lightweight computational approach, akin to the SE block.

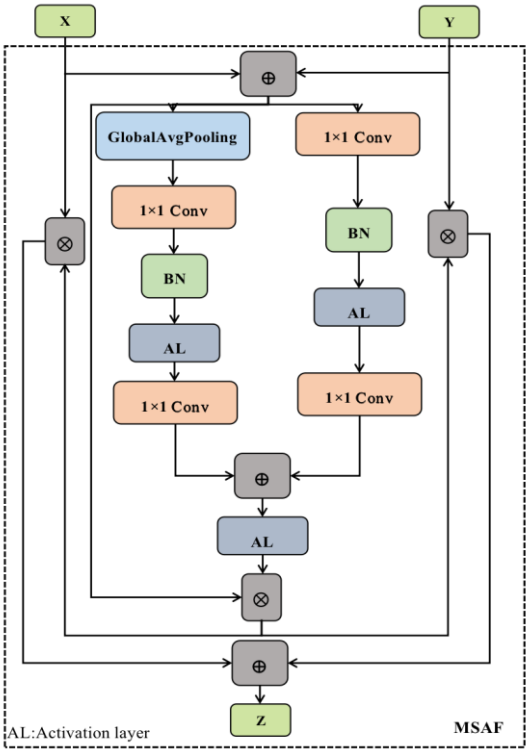



Fig. 5. Multi-scale attention fusion module diagram.

### A. *Triplet Attention*

Triplet Attention [13] is a novel approach for computing attention weights by capturing cross-dimensional interactions using a three-branch structure. For the input tensor, Triplet Attention establishes interdependencies between dimensions through rotation operations and residual transformations, encoding channel and spatial information with negligible computational overhead. Triplet Attention consists of three branches, where two branches capture cross-channel interactions between the C dimension and W/H, and the remaining branch calculates the traditional spatial attention weights, similar to CBAM. The first branch, the spatial attention calculation branch, takes the input features through the Z-Pool operation, as shown in formula (5), which performs max pooling and average pooling on the input, resulting in a 2×H×W feature map:

$$Z-Pool(x)=[MaxPool_{0d}(x), AvgPool_{0d}(x)] \quad (5)$$

In the formula, 0d represents the 0-th dimension where the max pooling and average pooling operations occur.

Next, convolutional operations are performed, followed by a Sigmoid activation function, ultimately generating the required attention weights. The second branch captures cross-channel and spatial interactions between channel dimension C and spatial dimension W through a series of operations. The input features are first permuted to the H×C×W dimensions, followed by Z-Pool along the H dimension, and subsequent operations are similar to the previous branch. Finally, the permute operation is applied to transform the features into C×H×W dimension for convenient element-wise addition. Similarly, the third branch captures cross-channel and spatial interactions between channel dimension C and spatial dimension H. The input features are first permuted to the W×H×C dimension, followed by Z-Pool along the W dimension, and subsequent operations are similar to the previous branches. Finally, it needs to be changed into a C×H×W dimension feature by permute, as illustrated in Fig.6.

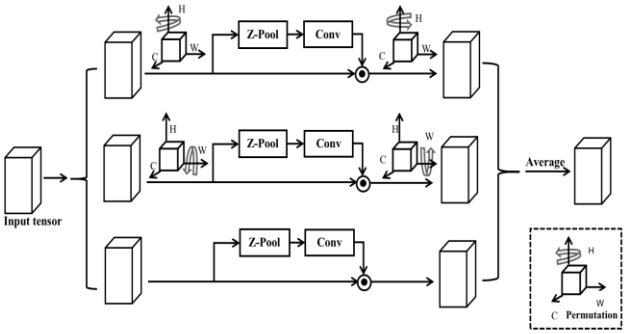



Fig. 6. The structure of Triplet Attention.

The Triplet Attention consists of three branches, with two branches dedicated to capturing cross-channel interactions between the channel dimension and the spatial dimension W/H, while the remaining branch performs traditional spatial attention weight calculation.

### B. *AFF*

Feature fusion, the integration of features from various layers or branches, is a commonplace element in contemporary network architectures. Traditionally, this is accomplished through basic linear operations like summation or concatenation. However, these methods may not always be the most efficient choices. As a solution, an all-encompassing and adaptable approach, known as Attention-based Feature Fusion (AFF), has been introduced. AFF has a wide range of applicability, including Inception layers [14] and feature fusion generated through both short and long skip connections. In the AFF module, the fusion of initial features X and Y is simply performed by element-wise addition. The fused features are subsequently fed into the attention module, where they contribute to the computation of the final fusion weights. In order to have a comprehensive perception of the input feature maps, it is necessary to apply an attention-based fusion mechanism to the fusion of initial features as well, by using another attention module to fuse the input features. However, the initial feature fusion may become a bottleneck, hence the proposed Iterative Attention-based Feature Fusion (iAFF) module. The architecture is illustrated in Fig.7 and Fig.8.

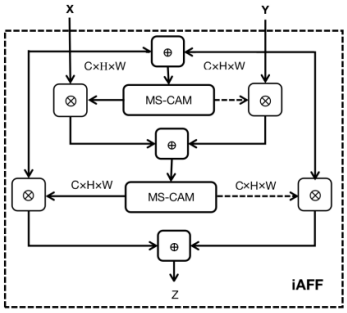



Fig. 7. iAFF Module.

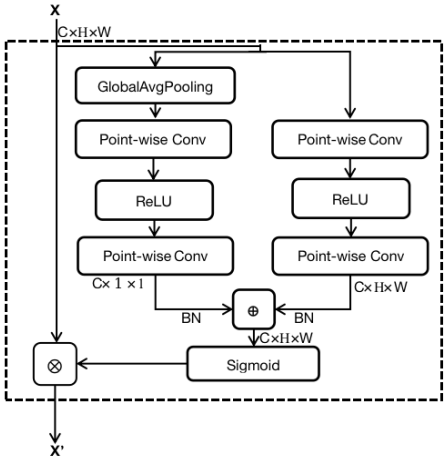



Fig. 8. MS-CAM Module.

The Iterative Attention-based Feature Fusion (iAFF) module distinguishes itself from the Partial Context-aware Method [15] by addressing the inherent challenge in fully context-aware methods—specifically, the initial integration of input features. The quality of initial fusion, as input to the attention module, may directly affect the final fusion weights. To address this feature fusion problem, another attention module is used to fuse the input features, as illustrated in Fig.7.

As seen in Fig.8, channel attention is achieved across multiple dimensions by altering the spatial pooling size. Weight reduction is achieved by introducing local context into the global context as needed. The use of point-wise convolution in this study is aimed at aggregating local channel context, achieved by channel interactions based on spatial positions.

Calculating the local channel context $L(X)\in R^{C\times H\times W}$ on the basis of the characteristic of parameter preservation using a bottleneck structure, as shown in Formula (6):

$$L(X)=\mathfrak{B}(PWConv_2(\delta(\mathfrak{B}(PWConv_1(X))))) \tag{6}$$

The kernel sizes of $PWConv_1$ and $PWConv_2$ are $(C/r)\times C\times 1\times 1$ and $C\times(C/r)\times 1\times 1$, respectively, L(X) maintains an identical shape to the input feature, allowing it to preserve and accentuate fine details within the low-level features. Given the global channel context g(X) and local channel context L(X), the formula of refined feature $X'\in R^{C\times H\times W}$ can be obtained by MS-CAM as shown in (7). $M(X)\in R^{C\times H\times W}$ represents the attention weights generated by MS-CAM.

$$X'=X\otimes M(X)=X\otimes\sigma(L(X)\oplus g(X)) \tag{7}$$

Given two feature maps X and Y, it is assumed that Y has a larger receptive field by default, as shown in Formulas (8) and (9):

$$Z=M(X\uplus Y)\otimes X+(1-M(X\uplus Y))\otimes Y \tag{8}$$

$$X\uplus Y=M(X+Y)\otimes X+(1-M(X+Y))\otimes Y \tag{9}$$

$Z\in R^{C\times W\times H}$ is the fusion feature, $\uplus$ represents the initial feature integration;the dashed line in the iAFF structure diagram represents $1-M(X\uplus Y)$ ;the fusion weight $M(X\uplus Y)$ should be composed of real numbers between 0 and 1;$1-M(X\uplus Y)$ can make the network perform soft selection or perfrom weighted average between X and Y.

### C. GCNet

GCNet [4] integrates the advantages of SENet [16] and NLNet [17], combining the global contextual modeling capability of NLNet with the lightweight design of SENet. GCNet is an improvement upon the non-local network (NLNet) attention mechanism, where "non-local" refers to the ability to capture long-range dependencies instead of just local neighborhoods. Although stacking convolutional layers can enlarge the receptive field, it's important to note that the receptive field of individual convolutional kernels on the original image remains constrained. This limitation arises from the inherent local operations. However, certain tasks require more global information from the original image, such as attention mechanisms. Introducing global information into certain layers can address the issue of local operations not capturing the global context, thereby providing richer information to subsequent layers. Nevertheless, non-local connections also come with challenges, as they involve global fully connected operations that result in a large number of parameters, making optimization difficult. The proposed Simplified NLNet in GCNet greatly reduces the computational cost of NLNet. The GCNet's Global Context (GC) block is structured as shown in Fig.9.

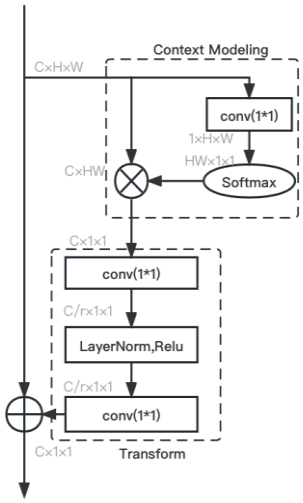



Fig. 9. Global context(GC) block.

The GC block consists of global attention pooling for context modeling, bottleneck transformations to capture channel correlations and element-wise addition for feature fusion. It amalgamates the benefits of the Simplified Non-Local (SNL) block for proficient long-range dependency modeling with the computationally lightweight characteristics of the Squeeze-and-Excitation (SE) block. The global context provided by the GC block can benefit a wide range of visual recognition tasks. Due to its flexibility, the GC block can be inserted into various network architectures used for computer vision problems. The formula is shown as (10).

$$z_i=x_i+W_{v2}ReLU(LN(W_{v1}\sum_{j=1}^{N_p}\frac{e^{W_k x_j}}{\sum_{m=1}^{N_p}e^{W_k x_m}}x_j)) \tag{10}$$

（1）Relationship with non-local blocks

Since non-local blocks actually learn global context that is irrelevant to the queries, the reason why the GC block can mimic the same global contextual information as non-local blocks is due to its global attention pooling, and it can significantly reduce computational costs. The bottleneck transformations in the GC block reduce redundancy in the global context features, resulting in further reduction in the number of parameters and floating point operations (FLOPs). The feasibility of applying GC blocks to multiple layers stems from the fact that the FLOPs and parameter count of GC blocks are noticeably lower than those of non-local blocks, and they only require a slight increase in computational cost. This approach effectively captures long-range dependencies, making it highly beneficial for network training.

（2）Relationship with the squeeze-and-excitation module

The main differences between the SE module and the GC module lie in their fusion mechanisms, reflecting the different objectives of the two modules. The SE module uses channel-wise scaling to recalibrate the importance of channels, but it may not fully capture long-range dependencies. On the other hand, the GC block follows the approach of the non-local block, aggregating global context to all positions using addition, which helps capture long-range dependencies. One more distinction lies in the utilization of layer normalization within the bottleneck transformation. Given that the GC block employs addition for fusion, the inclusion of layer normalization aids in streamlining the optimization process for the two-layer architecture of the bottleneck transformation, ultimately enhancing overall performance.

## IV. Experiment and Result Analysis

### A. Experimental data sets and evaluation metrics

This experiment involves integrating SCSS modules into YOLOv5s and utilizing the VOC2012 dataset, which includes 20 categories. This dataset includes people, animals, vehicles, and household items, grouped into 4 main categories, with 5717 images in the training set and 5823 images in the validation set. Subsequently, this study integrated the improved attention mechanism into the ResNet50 classification model for validation, using the CIFAR-10 and CIFAR-100 public datasets. CIFAR-10 has 10 categories and 50000 training images, while CIFAR-100 has 100 categories with the same number of training images. These datasets are essential for refining and assessing computer vision models and algorithms.

In this experiment, we have chosen to use Average Precision (AP) as our evaluation metric. AP offers a comprehensive measure, quantifying detection accuracy across various recall rates, providing insights into precision and recall trade-offs. It's commonly used for evaluating object detection models, typically assessed per class. Mean Average Precision (mAP), the average of AP for all classes, serves as the final performance measure for comparing detection performance across different object categories on the entire dataset.

In the context of object detection, Intersection over Union (IoU) is used to measure the accuracy of the predicted bounding box about the ground truth. In this experiment, IoU is set to 0.5, and the predicted results are categorized as follows: if the IoU is greater than or equal to 0.5 and the predicted class is correct, it is considered a good prediction; if the predicted class is incorrect, it is considered a bad prediction; if the IoU is less than 0.5, it is considered a bad prediction. To focus more on the accuracy of the bounding box location, this experiment also sets mAP@.5:.95, where AP values are averaged over multiple IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05.

### B. Model training environment configuration

In image classification, the learning rate is configured as 0.0001, the image size is cropped to 32x32, and the batch size is set to 64. The chosen optimizer is SGD.

For object detection, the learning rate is also set to 0.0001. The image size is resized to 640x640, and the batch size is set to 32. Similar to image classification, SGD is employed as the optimizer.

### C. Experimental test of YOLOv5s fused with SCSS

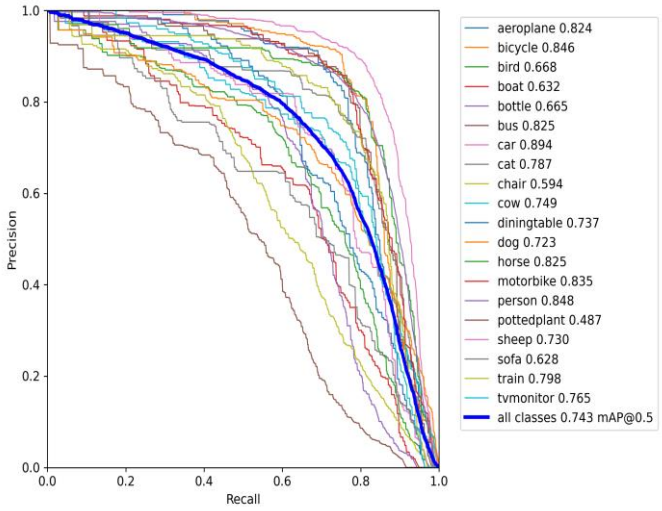
In this study, four types of attention modules, namely SE (referred to as YOLOv5s+SE), CBAM (referred to as YOLOv5s+CBAM), CA (referred to as YOLOv5s+CA), and SCSS (referred to as YOLOv5s+SCSS), are incorporated into the YOLOv5s model after the 17th, 20th, and 23rd C3 layers for comparative experiments. The outputs of these modules are used as inputs to the final detection head. The results of YOLOv5s and the four attention mechanism models on the VOC2012 dataset are shown in Table I. The addition of the SCSS module leads to a significant improvement of 3.9% in mAP.

TABLE I
COMPARISON OF YOLOv5s EXPERIMENT RESULTS INTEGRATED WITH ATTENTION MODEL

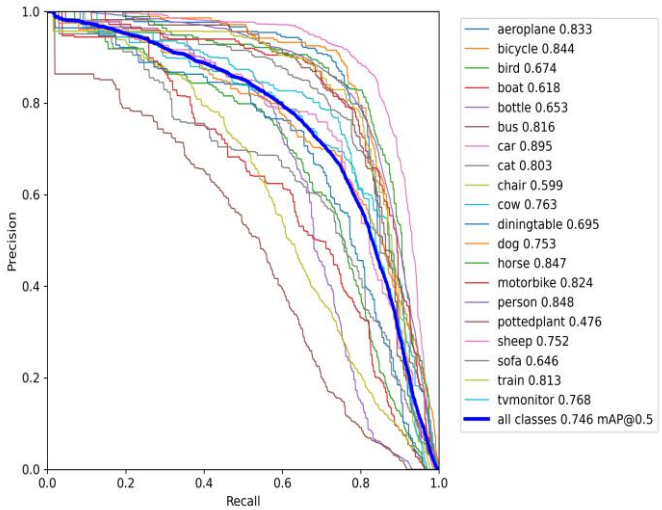
| Model | mAP/% |
|---|---|
| YOLOv5s | 74.3 |
| YOLOv5s+SE | 74.6 |
| YOLOv5s+CBAM | 74.6 |
| YOLOv5s+CA | 74.8 |
| YOLOv5s+SCSS | 78.2 |

According to the predictions of the classifier, the samples are sorted in descending order, with samples that the classifier considers most likely to be positive at the top, and samples that the classifier considers least likely to be positive at the bottom. In this study, mAP@0.5 and mAP@[.5:.95] are selected as evaluation metrics. Precision is plotted on the y-axis and Recall is plotted on the x-axis to draw the Precision-Recall (P-R) curves. Five types of P-R curves, namely YOLOv5s, YOLOv5s+SE, YOLOv5s+CBAM, YOLOv5s+CA, and YOLOv5s+SCSS, are plotted separately as shown in Fig.10.

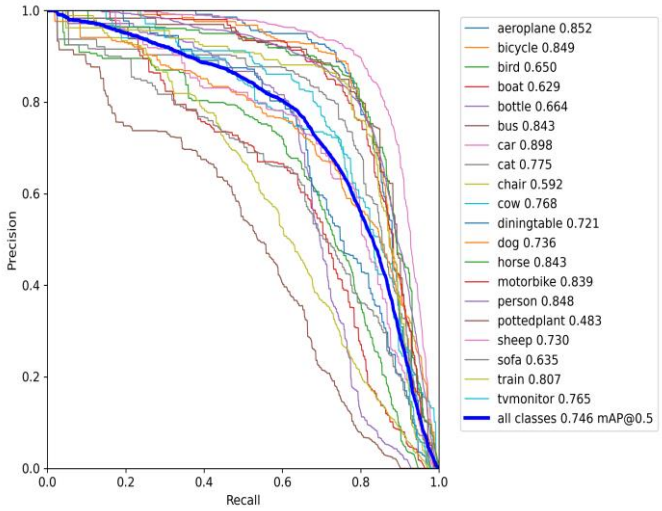
From the following five plots, it can be observed that the predicted probabilities of the "aeroplane" category for YOLOv5s with different attention modules are 82.4%, 83.3%, 85.2%,85.7%, and 87% respectively, while the predicted probabilities for the "bicycle" category are 84.6%, 84.4%, 84.9%,86.3%, and 87.3% respectively. It can be seen that YOLOv5s+SCSS has the highest predicted probabilities, indicating the best performance among the five models.



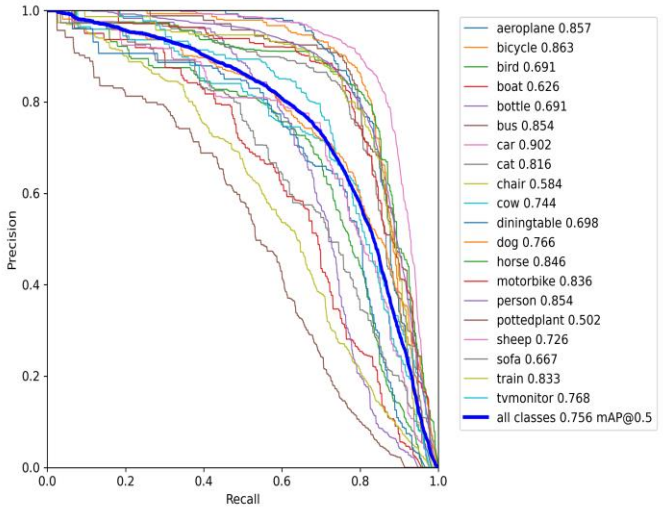

（a）YOLOv5s



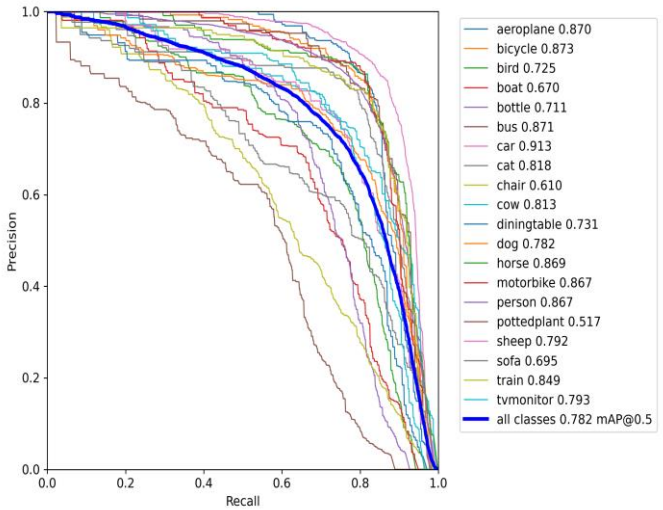

（b）YOLOv5s+SE




（c）YOLOv5s+CBAM




（d）YOLOv5s+CA




（e）YOLOv5s+SCSS

Fig. 10. Five fusion results with YOLOv5s as baseline.

To provide further substantiation and solidify the already compelling body of evidence supporting the effectiveness of the proposed optimization scheme, exhaustive and methodically executed testing procedures were meticulously carried out utilizing the VOC2012 dataset within the framework of this comprehensive study. The results derived from the empirical evaluation, thoughtfully encapsulated in the informative Fig.11, unequivocally illuminate the remarkable capabilities of the YOLOv5s model, which has been enriched by the seamlessly integrated SCSS module. This augmentation indisputably culminates in an unprecedented level of precision within the domain of object detection when compared to its contemporaries.

These contemporaries can be grouped into four categories: the first group lacks an attention mechanism, the second combines with SE, the third with CA, and the fourth with CBAM. This comparison underscores that the YOLOv5s model with the integrated SCSS module excels among its algorithmic peers.

(a) YOLOv5s (b)YOLOv5s+SE (c)YOLOv5s+CBAM (d)YOLOv5s+CA (e)YOLOv5s+SCSS

Fig. 11. Five prediction results.

### D. *Experimental testing of ResNet50 fused with SCSS*

Based on the findings presented in Table II and Table III from the experiments, this study conducted a comparison experiment by incorporating SE (denoted as ResNet50+SE), CBAM (denoted as ResNet50+CBAM), CA(denoted as ResNet50+CA), and SCSS (denoted as ResNet50+SCSS) modules before the pooling layer in ResNet50 on the CIFAR-10 and CIFAR-100 datasets. After adding the SCSS module, the Top-1 accuracy on the CIFAR-10 dataset improved significantly by 1.79%, and the Top-5 accuracy improved by 0.13%. On the CIFAR-100 dataset, the Top-1 accuracy improved by 3.11%, and the Top-5 accuracy improved by 1.3%. These results indicate that incorporating the SCSS module into ResNet50 improves the classification performance on both CIFAR-10 and CIFAR-100 datasets, surpassing the other three attention mechanisms (SE, CBAM, and CA).

TABLE II
COMPARISON OF RESNET50 EXPERIMENT RESULTS INTEGRATED WITH ATTENTION MODEL(CIFAR-10)

| Model | Top-5(%) | Param.(M) | FLOPs(M) |
|---|---|---|---|
| ResNet50 | 99.25 | 23.521 | 89.22 |
| ResNet50+SE | 99.32 | 23.525 | 89.27 |
| ResNet50+CBAM | 99.28 | 24.047 | 89.25 |
| ResNet50+CA | 99.38 | 23.524 | 89.25 |
| ResNet50+SCSS | 99.38 | 24.103 | 89.28 |

TABLE III
COMPARISON OF RESNET50 EXPERIMENT RESULTS INTEGRATED WITH ATTENTION MODEL(CIFAR-100)

| Model | Top-1(%) | Top-5(%) | Param.(M) | FLOPs(M) |
|---|---|---|---|---|
| ResNet50 | 55.28 | 82.3 | 27.317 | 83.31 |
| ResNet50+SE | 55.78 | 83.11 | 27.321 | 84.51 |
| ResNet50+CBAM | 55.53 | 82.75 | 27.289 | 84.47 |
| ResNet50+CA | 55.86 | 83.15 | 26.984 | 84.25 |
| ResNet50+SCSS | 58.39 | 83.6 | 26.972 | 84.27 |

### E. *Ablation experiment*

Self-attention

In the experiments, the channel attention, spatial attention, and multi-scale feature fusion were retained, collectively referred to as the SCSS module with self-attention mechanism. To validate the effectiveness of integrating CIFAR-10 and CIFAR-100 datasets with a self-attention mechanism, Scenario 1 involves the inclusion of self-attention, while Scenario 2 does not include self-attention.

TABLE IV
ABLATION EXPERIMENT OF SELF ATTENTION

| Model | Top-1(%) (CIFAR-10) | Top-1(%) (CIFAR-100) |
|---|---|---|
| ResNet50+SCSS-SELF | 85.93 | 56.55 |
| ResNet50+SCSS | 86.75 | 58.39 |

Based on Table IV, it can be observed that the model with self-attention outperforms the model without self-attention by 0.82% in terms of Top-1 accuracy on the CIFAR-10 dataset. Similarly, on the CIFAR-100 dataset, the model with self-attention surpasses the model without self-attention by 1.84% in terms of Top-1 accuracy. These results provide evidence that incorporating self-attention is effective and necessary.

Channel Attention

In the experiments, we integrated spatial attention, multi-scale feature fusion, and self-attention mechanisms, collectively known as the SCSS module. We conducted tests using two versions of the SCSS module: one incorporating channel attention (referred to as CHANNEL) and another without channel attention. The effectiveness of including channel attention was evaluated on the CIFAR-10 and CIFAR-100 datasets, respectively.

Based on the findings presented in Table V, the model incorporating channel attention demonstrates a 1.1% improvement in Top-1 accuracy compared to the model without channel attention on the CIFAR-10 dataset. Likewise, on the CIFAR-100 dataset, the model with channel attention

exhibits a 1.16% enhancement in Top-1 accuracy over the model without channel attention.

TABLE V
ABLATION EXPERIMENT OF CHANNEL ATTENTION

| Model | Top-1(%) (CIFAR-10) | Top-1(%) (CIFAR-100) |
|---|---|---|
| ResNet50+SCSS-CHANNEL | 85.93 | 56.55 |
| ResNet50+SCSS | 86.75 | 58.39 |

These results serve as compelling evidence that the inclusion of channel attention is not only effective but also necessary for improved performance.

Spatial Attention

In the experiments, we preserved the channel attention, multi-scale feature fusion, and self-attention mechanisms, collectively known as the SCSS module. We conducted tests using two variations of the SCSS module: one incorporating spatial attention (referred to as SPATIAL) and another without spatial attention. To demonstrate the effectiveness of spatial attention, CIFAR-10 and CIFAR-100 datasets were integrated.

TABLE VI
ABLATION EXPERIMENT OF SPATIAL ATTENTION

| Model | Top-1(%) (CIFAR-10) | Top-1(%) (CIFAR-100) |
|---|---|---|
| ResNet50+SCSS-SPATIAL | 85.23 | 57.12 |
| ResNet50+SCSS | 86.75 | 58.39 |

Based on Table VI, it can be observed that the model with spatial attention outperforms the model without spatial attention by 1.52% in terms of Top-1 accuracy on the CIFAR-10 dataset. Similarly, on the CIFAR-100 dataset, the model with channel attention surpasses the model without channel attention by 1.27% in terms of Top-1 accuracy. These results provide evidence that incorporating spatial attention is effective and necessary.

## V. CONCLUSION

This study integrates channel, spatial, and self-attention mechanisms, along with a multi-scale feature fusion module, to establish global dependencies, increase the receptive field for capturing more contextual information, and extract diverse shallow features from original images to capture rich image information. Experimental results demonstrate a significant improvement in the accuracy of both object detection and image classification when incorporating the proposed attention model. In object detection, integrating the attention model leads to a 3.9% improvement in the mean Average Precision (mAP). For image classification on the CIFAR-10 dataset, the Top-1 accuracy is improved by 1.79%, and the Top-5 accuracy is improved by 0.13%. Similarly, on the CIFAR-100 dataset, the Top-1 accuracy is improved by 3.11%, and the Top-5 accuracy is improved by 1.3%. Future research will further explore the application of the SCSS module in other areas of convolutional neural networks (CNNs).

## REFERENCES


[1] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations fordeep neural networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 5987-5995, 2017.

[2] Xu Z, Chen Q, Chen Y, "Tree Species Recognition Using Combined Attention and ResNet for Unmanned Aerial Vehicle Images," *Laser & Optoelectronics Progress*, 2023, 60(2): 0210004.

[3] Li X, Wang W, Hu X, et al., "Selective kernel networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 510-519, 2019.

[4] HU J, SHEN L, SUN G, et al., "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 2011-2023, 2020.

[5] Hou Q, Zhou D, Feng J, "Coordinate attention for efficient mobile network design," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 13713-13722, 2021.

[6] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[J].arXiv: 1807.06521, 2018.

[7] Chen Y, Kalantidis Y, Li J, et al., "Aˆ 2-nets: Double attention networks ," *Advances in Neural Information Processing Systems,* pp. 352-361, 2018.

[8] Wenhao Pan, and Kai Yang, "Enhanced Multi-Head Self-Attention Graph Neural Networks for Session-based Recommendation," *Engineering Letters*, vol. 30, no.1, pp37-44, 2022.

[9] Everingham M, Van Gool L, Williams C K I, et al, "The Pascal visual object classes(voc)challenge," *International Journal of Computer Vision*, pp. 303-338,2010.

[10] Krizhevsky A, Hinton G, "Learning multiple layers of features from tiny images," *Handbook of Systemic Autoimmune Diseases,* 2009, 1(4):1.

[11] Longlei Cui, and Ying Tian, "Facial Expression Recognition by Regional Attention and Multi-task Learning," *Engineering Letters*, vol. 29, no.3, pp919-925, 2021.

[12] Yue-Hua, L. I. , et al., "Video Smoke Detection Based on Multi-feature Fusion and Modified Random Forest," *Engineering Letters,* pp. 1115-1122, 2021.

[13] Misra D, Nalamada T, Arasanipalai A U, et al., "Rotate to Attend: Convolutional Triplet Attention Module," *Proceedings of the IEEE/CVF Winter Applications of Computer Vision,* pp. 3139-3148, 2020.

[14] Dai Y, Gieseke F, Oehmcke S, Wu Y, Barnard K, "Attentional feature fusion," *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3560-3569, 2021.

[15] Li H, Xiong P, An J, et al., "Pyramid attention network for semantic segmentation," *British Machine Vision Conference (BMVC),* pp. 1-13, 2018.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[17] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2018.