IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 September 2022, accepted 21 September 2022, date of publication 28 September 2022,
date of current version 10 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3210711

RESEARCH ARTICLE

MLAN: Multi-Level Attention Network

QINXUAN WANG[1,*], PEINUAN QIN[2,*], YUE ZHANG[3], XUEYAO WEI[4], AND MEIGUO GAO[2]

[1]School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
[2]Faculty of Engineering and Information Technology, The University of Melbourne, Victoria, VIC3010, Australia
[3]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[4]School of Mechanical and Electronic Information, China University of Geosciences, Wuhan 430074, China

Corresponding author: Meiguo Gao (meiguo_g@bit.edu.cn)

This work was supported by the Radar and Countermeasure Laboratory, School of Information and Electronics, Beijing Institute of Technology, Beijing, China.

*Peinuan Qin and Qinxuan Wang are co-first authors.

ABSTRACT In this paper, we proposed a "Multi-Level Attention Network" (MLAN), which defines a multi-level structure, including layer, block, and group levels to get hierarchical attention and combines corresponding residual information for better feature extraction. We also constructed a shared mask attention module (SMA) which can significantly reduce the number of parameters compared with conventional attention methods. Based on the MLAN and SMA, we further investigated a variety of information fusion modules for better feature fusion at different levels. We conducted classification task experiments based on the ResNet backbone with different depths, and the experimental results show that our method has a significant performance improvement over the backbone on CIFAR10 and CIFAR100 datasets. Meanwhile, compared with the mainstream attention methods, our MLAN performs better with higher accuracy as well as less parameters and computation complexity. We also visualized some intermediate feature maps and explained why our MLAN performs well.

INDEX TERMS Multi-level structure, shared mask attention, hierarchical attention aggregation, information fusion.

I. INTRODUCTION

The attention mechanism is a technology widely used in natural language processing (NLP) [1], [2], [3], [4], [5], statistical learning [6], [7], image detection [8], [9], [10], [11], speech recognition [12], [13], [14], [15] and other fields since the rapid development of deep learning [16], [17]. When a scene enters our vision field, we always pay attention to some key points in the scene first, such as dynamic points or abrupt colors, and the remaining static scenes may be temporarily ignored [18]. The attention mechanism imitates this point and makes the neural network able to focus on important information with high weight and ignore irrelevant information. It can also continuously adjust the weight so that the important information can be selected in different situations. Therefore, it makes the model obtain higher scalability and robustness [19]. However, the traditional attention methods generate the attention mask only based on specific layers but ignore that the attention could also have hierarchical structures.

With the development of Computer Vision (CV), more difficult visual tasks require deeper networks as well as more complex structures to extract features. However, the complexity of the network will inevitably lead to an increase in computable parameters. Hence, how to obtain attention structures with lightweight but good effect is a hot topic in current research. MobileNet [20] introduced the depthwise separable convolution to reduce the parameters in the convolution process. SqueezeNet [21] achieved the same goal by replacing the 3×3 convolution operation with 1×1 and reducing the number of channels.

In this paper, we proposed a Multi-Level Attention Network (MLAN) to take advantage of the hierarchical attention information. Meanwhile, in order to decrease the parameters of the attention module, we introduced a new attention mechanism SMA, which is short for Shared Mask Attention. The MLAN takes the ResNet [22] as the backbone and modifies it to multi-level structures, including layer, block, and group levels. And to make full use of the information, we built a

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

correspondingly hierarchical attention mechanism that aggregates layer attentions for block attentions and then aggregates block attentions for a group one. The residual information and attention information of the same level will be fused through the fusion module, and then transmitted to the next layer. Moreover, for sake of the relatively optimal fusion method, we designed four different fusion methods and carried out experiments respectively.

The SMA is a lightweight module and unlike traditional attention mechanisms, it does not generate a mask with the same channels as the input feature. Instead, it only generates a single channel attention mask, and then uses the broadcast mechanism to multiply the input feature by this shared mask, which greatly reduces the number of parameters and computation. And to avoid the performance decrease caused by the shared mask structure, we decided to generate multiple shared masks one time, take the mean of which for the final result of the whole SMA module.

We conducted comparative experiments with the up-to-date attention methods and the experimental results show that our method has better and more stable performance. In Fig.5, we visualized the features extracted from the attention network and found that our attention method can learn more critical details. After the article is accepted, an implementation of our method can be found at https://github.com/PeinuanQin/MLAN or https://github.com/wangqinxuan/MLAN.

In summary, the main contributions in this paper are as follows:

- Proposed the MLAN, which divides the network into three levels (layer, block, and group). Both block and group levels have their own residual structure to alleviate the information loss caused by network deepening. We also generated hierarchical attention corresponding to each level through aggregation to take advantage of information extracted from different level structures.
- Designed a lightweight attention module, SMA, to decrease the parameters of the attention mechanism while maintaining the focus on effective features.
- Conducted comprehensive experimental comparisons between our method and mainstream attention mechanisms on different ResNet structures and datasets in terms of model parameters, computational complexity, feature extraction effects, etc., which proved our method is more competitive.
- Designed four feature fusion modules to fuse the multi-level residual and attention information into the backbone and carried out comparative experiments.

The rest article is organized in the following order. Section II introduces some work related to attention mechanisms and multi-level structures. In Section III, we specifically described the MLAN structure, including the SMA, attention aggregation modules, and fusion modules. Subsequently, we designed a series of experiments and verified the effectiveness of our methods in Section IV on CIFAR10 and CIFAR100. Finally, Section V makes a full review of our work and contributions.

II. RELATED WORK

A. ATTENTION MECHANISM

The conception of attention was first proposed by Bahdanau *et al.* [23] in the machine translation field for figuring out which parts the model we should pay the most attention to. After that, it developed well first in the field of natural language processing (NLP) [1] for mainly handling sequential decision tasks, and then swept the entire field of deep learning.

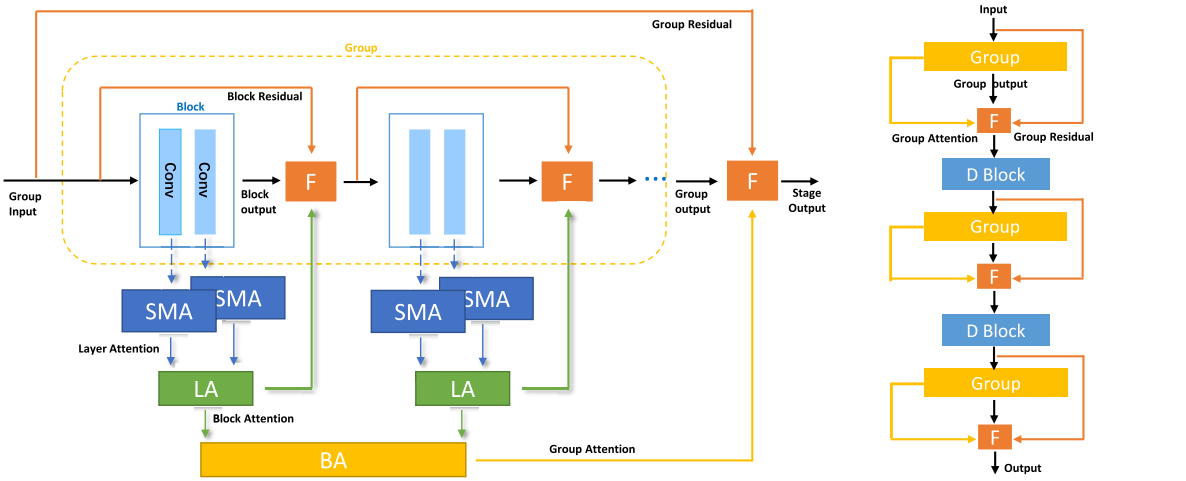
There are diverse forms of attention mechanisms, which can be roughly divided into soft attention and hard attention. Typical examples of soft attention are Spatial Transformer Networks (STN) [24], [25], [26], [27] and Residual Attention Networks [28], [29], [30], [31]. This soft-attention mechanism is differentiable and can be trained through back-propagation. Hard attention is usually trained by reinforcement learning models since the hard attention model is non-differentiable and cannot be trained end-to-end.

For classification tasks, combining basic backbone networks and attention modules to enhance the performance has also been proved to be sensible. SENet [32] put forward the Squeeze-and-Excitation (SE) Module, which uses the attention in the channel dimension to drive the model to focus on channels with the richest information while suppressing the unimportant channels. SKNet [33] merged a soft attention mechanism to choose proper receptive fields for better generalization. Furthermore, Woo *et al.* [34] proposed a Convolutional Block Attention Module (CBAM), in which the attention map of features is calculated from both channel and space dimensions. Misra *et al.* [35] investigated light-weight but effective attention mechanisms and presented triplet attention (TA), a novel method for computing attention weights by capturing cross-dimension interaction using a three-branch structure. Combining the advantages of Non-local [36] and SENet, Cao *et al.* [37] designed a global context (GC) block to construct a global context network (GCNet), which is lightweight and can effectively model the global context.

Our attention mechanism adopts an idea of sharing attention weights among all channels of the feature tensor, which reduces the network parameters a lot.

B. MULTI-LEVEL STRUCTURE

In CV tasks, many works use multi-level network structures or multi-level information extraction modules to make better use of information at different scales and levels. Shi *et al.* [38] proposed a dual branch multi-level feature dense fusion-based lightweight Convolutional Neural Network (BMDF-LCNN) to avoid the loss of shallow information due to network deepening. In [39], the Multi-level Convolutional Pyramid Semantic Fusion (MCPSF) framework was proposed to integrate multi-level semantic features extracted by bag-of-visual-words (BoVW) model and convolutional neural network (CNN) model. Zhang *et al.* [40] proposed a Multi-scale Time-Frequency Convolutional Recurrent Neural Network (MTF-CRNN) for sound time frequence map detection to improve sound event detection



(a) A decomposition of the multi-level attention structure

(b) The whole view of the MLAN framework

FIGURE 1. Whole view of the MLAN structure. SMA, LA, and BA represent the structures responsible for generating shared mask attention, block-level attention, and group-level attention, separately. F represents the fusion module. D Block represents the down-sample block whose first layer stride is 2. A group contains several normal blocks and each normal block has convolutional layers with stride = 1.

performance. Ding and He [41] proposed an Adaptive Multi-scale Detection (AdaMD) method, based on the hourglass neural network and the Gated Recurrent Unit (GRU) module, to extract different scale characteristics of time-frequency map.

The use of multi-level attention and multi-scale attention to capture key information is also an important idea to solve visual tasks. Guo *et al.* [42] proposed an attention-based network, MSANet, which applies an encoder-decoder structure for image data segmentation to aggregate contextual features from different levels and reconstruct spatial characteristics efficiently. To meet the real-time requirement of autonomous driving, Wang *et al.* [43] proposed a novel end-to-end recurrent multi-level residual learning deraining network featured with the global attention mechanism and residual network architecture. Yin *et al.* [44] introduced a new Visual Attention Dehazing Network (VADN) by proposing the multi-level refinement and fusion, which leverages a haze attention map as a haze relevant prior and learns complementary haze information among multi-level features.

Inspired by these design ideas, we divided the network into a 3-level structure (layer, block, and group), and introduced hierarchical attention at different levels. Finally, these attention maps are integrated with the residual information of the corresponding level and input to the subsequent network.

III. METHODOLOGY

As shown in Fig.1, we constructed the MLAN, which takes the ResNet as its skeleton and combines our newly proposed SMA module. In order to comprehensively utilize attention, on the basis of layer attention, we subsequently built up the block-level attention by the Layer Attention Aggregation (LA) module and the group-level attention by Block Attention Aggregation (BA) module and eventually formed a hierarchical attention structure.

To alleviate information loss, we also constructed residual branches on the block level and group level separately to assist the hierarchical attention.

Then, the residual and the attention would be fused to the backbone through the fusion module.

A. SHARED MASK ATTENTION MODULE

The generation of the attention mask depends on the feature values of different regions in the feature map, and the mask is used to magnify or suppress the original feature map for sake of effective features. We believed that the way to generate a weight for all feature values of all channels in a feature map is likely to be redundant, and too fine-grained operations might not further improve the effect of attention.

The method of SE [32] compresses the H-dimension and W-dimension feature values into a single point. After being processed by the attention mechanism, all pixels on the H-W plane share the same weight. The TA [35] module squeezes the features of all channels into two channels through the Z-Pool method and then outputs a single-channel attention mask through the convolution layer. Therefore, we considered that it is necessary to cut down the attention mask in a certain dimension. To achieve this goal, we proposed a lightweight SMA method. First, the feature map is integrated along the channel dimension to obtain a single-channel tensor containing the information of the entire feature map. Using this tensor, we can further generate a single-channel mask. Since this mask is produced based on the global information, the calculated attention weights also carry the global information, so when it is multiplied with the original feature map, it can ensure the feature extraction is reasonable and effective. For sharing the same mask among all feature channels, we implemented the broadcast operation to the single-channel mask and made it multiply with the origin feature map. By this means, the parameters and computation

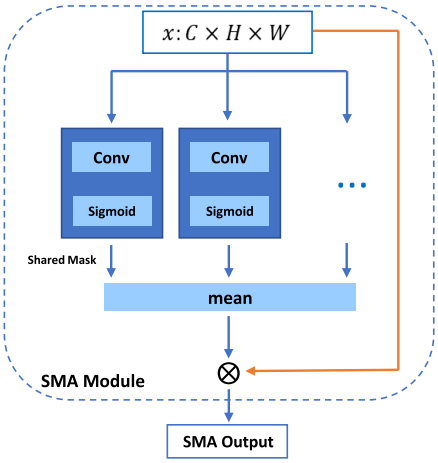


FIGURE 2. SMA module. It takes several (N) shared masks and averages them to multiply with the input feature for attention generation.

complexity would be greatly decreased. Practically, in the generation process of an attention mask, we tended to build N parallel branches of shared masks and took their average as the final attention mask. We thought this was a good way to make the model more robust.

The SMA module is shown in Fig.2

$$M_s^{(i)} = \sigma(Conv^{(i)}(x)) \in \mathbb{R}^{1\times H\times W} \tag{1}$$

$$M_{mean} = \frac{\Sigma_{i=1}^{N} M_s^{(i)}}{N} \in \mathbb{R}^{1\times H\times W} \tag{2}$$

$$O_{SMA} = M_{mean} \cdot x \tag{3}$$

where x represents the input feature of the SMA module. C, H, and W respectively represent the channel number, height, and width of the feature. $Conv(\cdot)$ indicates a convolution operation with a single-channel kernel, $\sigma(\cdot)$ is the sigmoid activation function. $M_s^{(i)}$ represents the $i-th$ shared mask in the SMA, and N means the number of shared masks in an SMA. M_{mean} is our averaged shared mask by averaging all these N shared masks, and O_{SMA} is the output of the SMA module. Traditionally, attention is implemented by producing a mask of the same size as x ($C\times H\times W$) and then multiplying it by x. However, this method may cause parameter redundancy and a great computation increase. SMA solves this problem to some extent. Besides, the way of averaging multiple shared masks improves the stability and effect of SMA. Different from the traditional attention method, the number of our shared masks N is controllable, and $N \ll C$. We finally multiplied the averaged shared mask by the input x. In this way, we achieved the goal of sharing attention weights among all channels of x, and greatly reduced parameters.

B. MULTI-LEVEL ATTENTION AGGREGATION

To make full use of higher-level attention and features, we adopted hierarchical processing. Layer and block aggregation modules (LA and BA) are established to obtain block-level attention and group-level attention respectively. LA and BA modules are shown in Fig.3.

$$X_l = Cat(x_l^{(1)}, x_l^{(2)}, \cdots, x_l^{(p)}) \in \mathbb{R}^{(p\cdot C_l)\times W_l\times H_l} \tag{4}$$

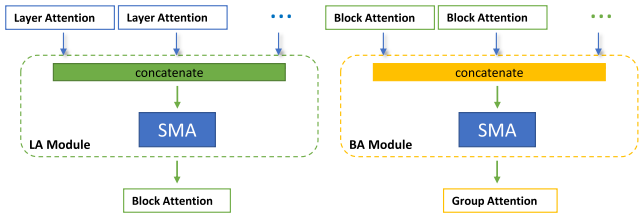


FIGURE 3. LA and BA module. The attention aggregation module concatenates the attention of the previous level to obtain a more informative tensor and then passes it to an SMA module to generate attention for the next level.

$$O_{LA} = SMA(X_l) \tag{5}$$

where $x_l^{(j)}$ means the $j-th$ layer attention with a shape of $C_l \times W_l \times H_l$, used for aggregation in a LA module. $Cat(\cdot)$ represents the concatenation operation, and X_l is the concatenated tensor of all p layer attentions. $SMA(\cdot)$ represents being processed by the SMA module, introduced in Section.III-A. O_{LA} indicates the output of the LA module. We spliced these layer attentions along the channel dimension and made them be processed by the SMA again. In this way, the layer attentions can be aggregated to a block-level one. Moreover, for the process of producing group-level attention, we still took this strategy, concatenating block-level attention and finally using SMA for aggregation.

$$X_b = Cat(x_b^{(1)}, x_b^{(2)}, \cdots, x_b^{(q)}) \in \mathbb{R}^{(q\cdot C_b)\times W_b\times H_b} \tag{6}$$

$$O_{BA} = SMA(X_b) \tag{7}$$

where $x_b^{(k)}$ is the $k-th$ block attention with a shape of $C_b \times W_b \times H_b$, used for aggregation in a BA module. X_b is the concatenated tensor of all q block attentions. O_{BA} indicates the BA module output.

C. FUSION MODULE

Since our model designs a variety of attentions hierarchically, how to effectively integrate them into the backbone network is also what we focus on. So in this part, we tried to propose different fusion modules to integrate attention and residual information at different levels.

1) PLUS FUSION

$$\begin{aligned} O_{plus} &= F_{plus}(x_O, x_A, x_R) \\ &= x_O + x_A + x_R \end{aligned} \tag{8}$$

where $F_{plus}(\cdot)$ is the plus fusion module. x_O, x_A and x_R respectively represent backbone features, block (or group) attention, and block (or group) residual information. By adding directly, the attention and residual information can be integrated into the backbone network and continue to be transmitted to the next layer.

2) CONCATENATION FUSION

$$\begin{aligned} O_{cat} &= F_{cat}(x_O, x_A, x_R) \\ &= Conv(Cat(x_O, x_A, x_R)) \end{aligned} \tag{9}$$

Splicing the three kinds of information together along the channel dimension minimizes the information loss, even if it brings a rapid increase in the number of parameters. Then the whole tensor is extracted through a convolution layer and restored to the channel numbers before concatenation.

3) 1×1 CONVOLUTION FUSION

$$\begin{aligned} O_{conv_{1\times1}} &= F_{conv_{1\times1}}(x_O, x_A, x_R) \\ &= Conv_{1\times1}(Cat(x_O, x_A, x_R)) \end{aligned} \tag{10}$$

where $Conv_{1\times1}(\cdot)$ is the 1×1 convolution layer used for compressing the concatenated tensor generated by $Cat(\cdot)$ operations. The kernel number of $Conv_{1\times1}(\cdot)$ is consistent with the number of channels of the input features, which means its parameters when doing convolution are only $\frac{1}{9}$ of the Concatenation method (Section III-C2).

4) WEIGHTED PLUS FUSION

$$\begin{aligned} O_{wp} &= F_{wp}(x_O, x_A, x_R) \\ &= w_1 \cdot x_O + w_2 \cdot x_A + w_3 \cdot x_R \end{aligned} \tag{11}$$

where w_1, w_2, and w_3 represent the weights of the three input information respectively. The weighted plus fusion module is an optimization of plus fusion. These weights are calculated by average pooling and linear mapping, followed by a sigmoid function. These weights are learnable and determine the contribution of the three different components.

IV. RESULT AND DISCUSSION

In this section, we conducted systematic experiments on CIFAR10 and CIFAR100. We implemented the same data enhancement method for CIFAR10 and CIFAR100: images were padded to $(36, 36)$ and randomly cropped back to $(32, 32)$, and then the image was randomly flipped horizontally. In addition, $batch\_size = 128$ was adopted in all training processes. *SGD* optimizer was used with the same configuration ($lr = 0.1, weight\_decay = 5e-4$) for all experiments. The optimizer attenuated the learning rate at the three milestones $[50, 100, 150]$ by $gamma = 0.1$ and we set a total of 200 epochs to train models. In the follow-up experiments, we uniformly selected $N = 3$ (E.q.(2)) for each attention generation and Plus Fusion as the information fusion method (E.q.(8)).

A. CIFAR10

CIFAR10 [45] is a common dataset, containing 60,000 images which are 32×32 in size and have RGB color channels. These images cover 10 different categories. Additionally, the entire dataset is split into a training set and a test set for 50,000 images and 10,000 images, separately.

In CIFAR10 experiments, we adopted ResNet14, ResNet20, ResNet26, and ResNet32 to verify the effectiveness of our MLAN method in networks with different depths. Taking ResNet14 as an example, we divided the network into 3 groups with each group containing 2 blocks, and each block composed of 2 convolution layers. This design corresponded to our subsequent multi-level attention structure. The frameworks and their parameter configurations are described in detail in Table.1. Different from other papers, in which each convolution layer always takes hundreds of kernels for extreme performance, our work concentrates more on testing the efficiency of structures and comparison with other benchmarks instead of only pursuing the absolute accuracy performances.

1) MLAN EXPERIMENTS

To verify our MLAN framework performance, we designed four different ResNet structures (Table.1) to conduct controlled experiments. The results are shown in Table.2 and Fig.4. We uniformly chose all models to use the Plus Fusion module for variable control. It is clear that when the training epoch exceeds about 50, the accuracy-epoch curve of the MLAN method dominates the accuracy metric. Finally, the average error rate of our model is 8.6% lower than that of the backbone method.

After that, we visualized the heat map of the last convolution layer output in the network to study the feature capture ability of different models, shown in Fig.5. By comparison, the features learned by the ResNet backbone are divergent, while our MLAN method can focus on more critical and convergent areas.

2) MULTI-LEVEL STRUCTURE EXPERIMENTS

In this experiment, we tried to demonstrate the superiority of our proposed multi-level structure. We conducted a set of comparative experiments, one of which was our MLAN, while the other only took the single-level structure, which means the layer attention and the block residual are fused in the fusion module and directly flow to the next layer. Table.3 shows that our MLAN performs better than the single-level attention structure at different ResNet depths, which proves that the simple layer-level attention information is not comprehensive, and there is also critical knowledge at block and group levels, which is supposed to be fully utilized.

3) SMA EXPERIMENTS

Full-Mask Attention (FMA) refers to the original attention design that has not been lightweight. It first generates weights through a convolutional layer of the same size as the input feature. After being activated by the sigmoid function, the input feature is multiplied point by point with these generated weights. In this part we compared our SMA with FMA and tended to demonstrate that the meaning of excess parameters is not very great. Experiments show that our SMA module not only exceeds the backbone (Table.2) but also surpasses the FMA (Table.4) on the CIFAR10. From this, we can conclude that overly detailed attention operations may not be completely conducive to the improvement of model results.

4) FUSION MODULE EXPERIMENTS

In this experiment, we carried out comparative experiments on 4 fusion methods proposed in Section.III-C. We made

TABLE 1. Configurations of ResNet variants on CIFAR10.

Groups	ResNet14	ResNet20	ResNet26	ResNet32
Group 1	$\begin{bmatrix} 3\times3, 8 \\ 3\times3, 8 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 8 \\ 3\times3, 8 \end{bmatrix} \times 3$	$\begin{bmatrix} 3\times3, 8 \\ 3\times3, 8 \end{bmatrix} \times 4$	$\begin{bmatrix} 3\times3, 8 \\ 3\times3, 8 \end{bmatrix} \times 5$
Group 2	$\begin{bmatrix} 3\times3, 16 \\ 3\times3, 16 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 16 \\ 3\times3, 16 \end{bmatrix} \times 3$	$\begin{bmatrix} 3\times3, 16 \\ 3\times3, 16 \end{bmatrix} \times 4$	$\begin{bmatrix} 3\times3, 16 \\ 3\times3, 16 \end{bmatrix} \times 5$
Group 3	$\begin{bmatrix} 3\times3, 32 \\ 3\times3, 32 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 32 \\ 3\times3, 32 \end{bmatrix} \times 3$	$\begin{bmatrix} 3\times3, 32 \\ 3\times3, 32 \end{bmatrix} \times 4$	$\begin{bmatrix} 3\times3, 32 \\ 3\times3, 32 \end{bmatrix} \times 5$
Average pool, Fc, Softmax				
Parameters	0.04M	0.06M	0.09M	0.12M

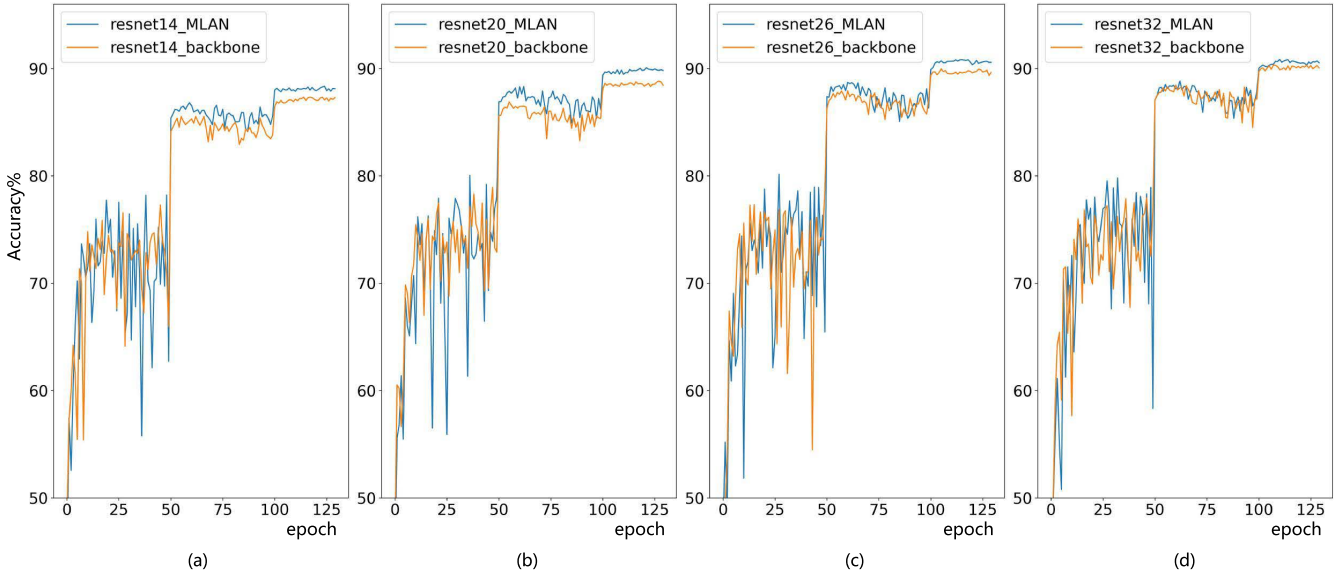


FIGURE 4. Training curve of MLAN on CIFAR10.

TABLE 2. MLAN experiments on CIFAR10 with ResNet variants of different depth.

methods	ResNet14	ResNet20	ResNet26	ResNet32
backbone	87.58%	89.00%	90.02%	90.34%
ours	**88.37%**	**90.23%**	**90.94%**	**90.84%**

TABLE 3. Multi-level structure experiments on CIFAR10 with ResNet variants of different depth.

structure	ResNet14	ResNet20	ResNet26	ResNet32
single-level	87.74%	88.84%	89.33%	89.94%
multi-level	**88.37%**	**90.23%**	**90.94%**	**90.84%**

TABLE 4. FMA v.s. SMA experiments on CIFAR10.

indicators	ResNet14	ResNet20	ResNet26	ResNet32
FM Acc	89.18%	90.21%	90.86%	90.83%
SM Acc	**88.37%**	**90.23%**	**90.94%**	**90.84%**
FM Paras	0.39M	1.04M	1.98M	3.21M
SM Paras	**0.06M**	**0.1M**	**0.14M**	**0.18M**

ResNet20 implement all these methods while keeping other conditions unchanged. Table.5 displays that the Concatenation Fusion gets the best accuracy (about 0.5% increase), but

TABLE 5. Fusion method experiments on CIFAR10.

indicators	plus	concatenation	1×1 conv	weighted plus
Acc	90.23%	91.00%	90.72%	90.11%
Paras	0.10M	0.21M	0.11M	0.10M

it also suffers from the parameter explosion (almost 100% increase). The other 3 fusion methods don't have a significant difference in performance. After consideration, we chose the Plus Fusion as the standard operation in the following experiments since it wouldn't produce any computed parameters.

5) COMPARISON WITH BENCHMARKS

Based on the ResNet backbone in Table.1, we reproduced the benchmarks of attention methods, including CBAM [34], SEnet [32], TAnet [35], and GCnet [37]. The experimental results are shown in Table.7, and our MLAN method is demonstrated to have better performance and the heatmap visualization (Fig.5) also supports this viewpoint.

To further highlight the advantages of our SMA in terms of parameters and computational complexity, we made the mainstream attention methods with the same multi-level structure as our MLAN. Table.6 shows that although the SE

TABLE 6. Comprehensive comparison results among various attention mechanisms based on the same multi-level architecture on CIFAR10.

	ResNet14			ResNet20			ResNet26			ResNet32		
methods	paras	times	acc	paras	times	acc	paras	times	acc	paras	times	acc
SMA	**0.06M**	**18.12ms**	**88.37%**	**0.10M**	**29.84ms**	**90.23%**	**0.14M**	**43.61ms**	**90.94%**	**0.18M**	**51.54ms**	**90.84%**
SMAv2	**0.05M**	**14.12ms**	**88.61%**	**0.08M**	**21.28ms**	**89.67%**	**0.11M**	**27.43ms**	**90.36%**	**0.14M**	**35.67ms**	**90.84%**
CBAM [34]	0.07M	43.85ms	88.37%	0.14M	65.15ms	87.88%	0.22M	89.42ms	89.50%	0.32M	117.55ms	89.04%
TA [35]	0.06M	72.01ms	88.41%	0.10M	106.74ms	89.31%	0.13M	141.24ms	90.24%	0.16M	172.72ms	89.96%
SE [32]	0.07M	18.23ms	88.09%	0.13M	27.33ms	89.20%	0.21M	36.43ms	89.49%	0.31M	47.61ms	89.91%
GC [37]	0.07M	29.64ms	87.55%	0.14M	43.21ms	89.15%	0.22M	54.81ms	89.79%	0.32M	74.37ms	90.51%

TABLE 7. Comparison of different attention mechanisms for ResNet of different depths on CIFAR10.

methods	ResNet14	ResNet20	ResNet26	ResNet32
Ours	**88.37%**	**90.23%**	**90.94%**	**90.84%**
CBAM [34]	87.10%	88.47%	89.01%	89.46%
SE [32]	87.36%	87.94%	88.92%	88.33%
TA [35]	87.74%	89.27%	90.06%	89.85%
GC [37]	87.18%	89.21%	89.39%	88.87%

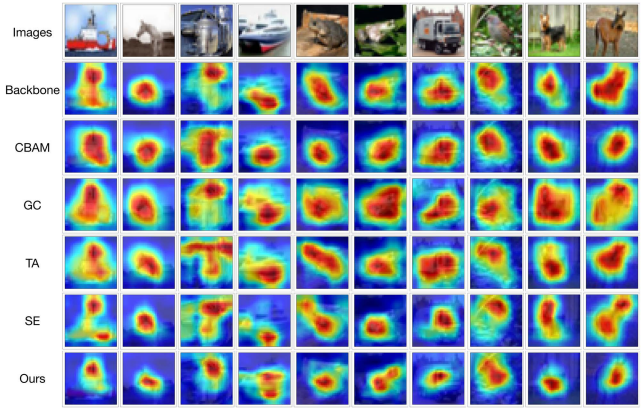


FIGURE 5. Feature visualization of the last feature layer based on different networks which take ResNet20 as the backbone on CIFAR10. The bright areas of the heatmap of our method are significantly more clustered than other methods, indicating that our method pays attention to more critical details.

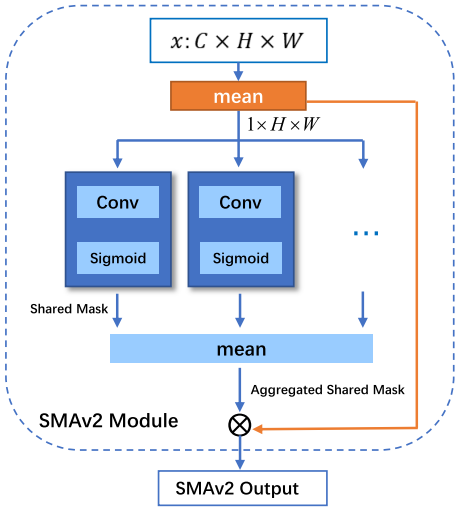


FIGURE 6. SMAv2 module. It modifies the SMA module by firstly averaging along the channel dimension for shared mask generation.

method has a slight advantage in inference speed over our SMA, its parameter quantity is much higher. The reason for this phenomenon is that the SE module first obtains information between channels through global average pooling, and then uses the fully connected network to generate channel weights. The global average operation replaces the feature extraction of convolution, so the SE method can directly attain the feature value of each channel, thus it is very fast. However, due to the use of fully connected operations when generating channel weights, the parameters explode with the increase of SE module numbers. In contrast, our SMA increases almost linearly with the change of network depth with a higher accuracy performance.

The TA module outperforms the SMA module slightly from the parameter aspect, but the inference speed is much slower. In detail, the TA module calculates three attention weights through three parallel branches. In each branch, the channel dimension of the feature map is pooled to reduce the channel number, and then the convolution is implemented for the pooled results to generate weights. The reduction of channel number further decreases the parameters of the convolutional layer, meanwhile, the structure of multiple branches causes more convolution operations, thus reducing the inference speed. Therefore, under multiple measures, we can preliminarily conclude that the SMA method has the advantages of better accuracy, fewer parameters, and faster inference speed at the same time.

Furthermore, by analyzing the fact that the SMA cannot surpass SE and TA roundly, we found that there was still room for improvement. Concretely, the parameters of our SMA method mainly come from the convolution operations when obtaining the shared mask. In order to further decrease the model complexity, we tried to modify the generation steps of shared masks and formed SMAv2 (shown in Fig.6).

SMAv2 reduced the channel number of the input feature before convolution operations, which cut down the computational parameters a lot. As shown in Table.6, although the accuracy is inferior compared with the original SMA, it still transcends all other benchmarks with a brilliant performance on parameter and computation complexity.

B. CIFAR100

The CIFAR100 [45] is a commonly used dataset in computer vision. It consists of 60,000 images in 100 categories, which is suitable for a variety of tasks. All images in CIFAR100 have a similar format to CIFAR10. Different from experiments designed for CIFAR10, in this part, we increased the

TABLE 8. Configurations of ResNet variants on CIFAR100.

Groups	ResNet14	ResNet20	ResNet26	ResNet32
Group 1	$\begin{bmatrix} 3\times3, 16 \\ 3\times3, 16 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 16 \\ 3\times3, 16 \end{bmatrix} \times 3$	$\begin{bmatrix} 3\times3, 16 \\ 3\times3, 16 \end{bmatrix} \times 4$	$\begin{bmatrix} 3\times3, 16 \\ 3\times3, 16 \end{bmatrix} \times 5$
Group 2	$\begin{bmatrix} 3\times3, 32 \\ 3\times3, 32 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 32 \\ 3\times3, 32 \end{bmatrix} \times 3$	$\begin{bmatrix} 3\times3, 32 \\ 3\times3, 32 \end{bmatrix} \times 4$	$\begin{bmatrix} 3\times3, 32 \\ 3\times3, 32 \end{bmatrix} \times 5$
Group 3	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 4$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 5$
Average pool, Fc, Softmax				
Parameters	0.18M	0.27M	0.37M	0.47M

TABLE 9. Comprehensive comparison results among various attention mechanisms based on the same multi-level architecture on CIFAR100.

	ResNet14			ResNet20			ResNet26			ResNet32		
methods	paras	times	acc	paras	times	acc	paras	times	acc	paras	times	acc
SMA	**0.22M**	**25.43ms**	**68.91%**	**0.36M**	**39.05ms**	**70.51%**	**0.50M**	**51.85ms**	**70.64%**	**0.63M**	**66.43ms**	**71.78%**
SMAv2	**0.20M**	**16.26ms**	**68.12%**	**0.32M**	**24.80ms**	**71.02%**	**0.44M**	**36.57ms**	**71.44%**	**0.56M**	**44.76ms**	**71.96%**
CBAM [34]	0.28M	42.94ms	67.72%	0.54M	77.50ms	67.77%	0.87M	117.09ms	68.13%	1.26M	148.33ms	69.79%
TA [35]	0.21M	84.12ms	67.67%	0.34M	123.04ms	69.39%	0.46M	185.91ms	70.13%	0.58M	215.48ms	70.25%
SE [32]	0.28M	17.28ms	68.69%	0.53M	32.87ms	69.24%	0.86M	49.69ms	70.11%	1.24M	61.57ms	70.66%
GC [37]	0.28M	46.96ms	68.42%	0.54M	71.30ms	69.05%	0.87M	92.85ms	69.63%	1.26M	115.07ms	70.18%

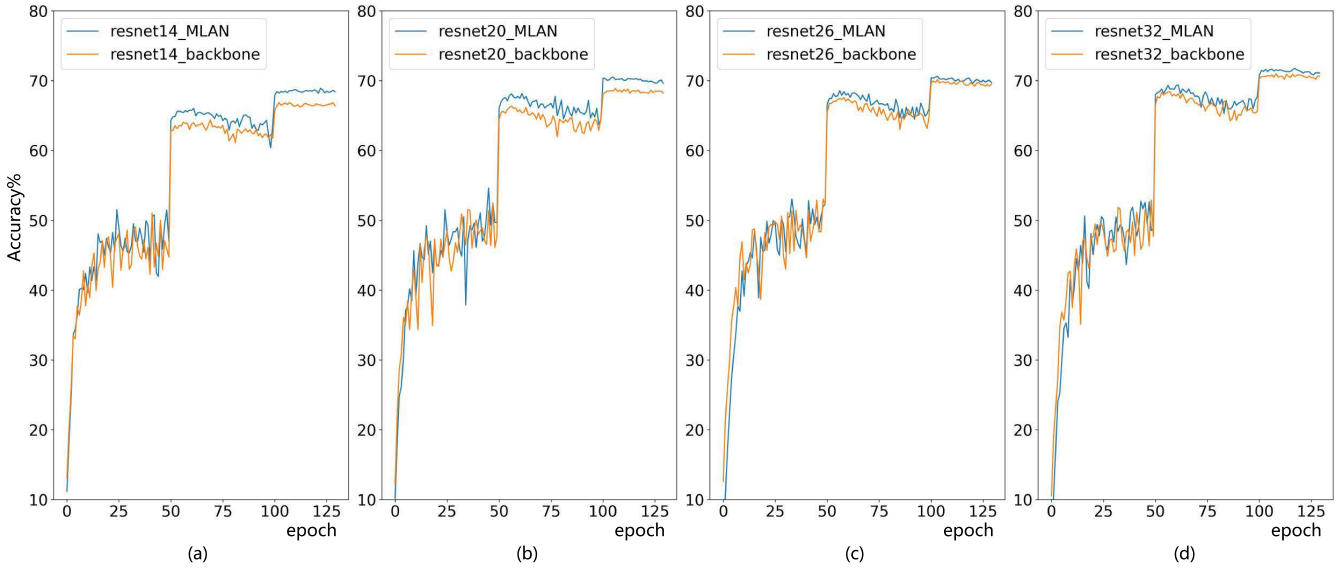


FIGURE 7. Training curve of MLAN on CIFAR100.

TABLE 10. MLAN experiments on CIFAR100 with ResNet variants of different depth.

methods	ResNet14	ResNet20	ResNet26	ResNet32
backbone	67.08%	68.78%	70.17%	71.01%
ours	**68.91%**	**70.51%**	**70.64%**	**71.78%**

kernel numbers of each convolutional layer but maintained the network depth unchanged.

1) MLAN EXPERIMENTS

Similar to experiments of CIFAR10, we also adopted ResNet14, ResNet20, ResNet26, and ResNet32. Their framework configurations are described in detail in Table.8. The results are shown in Table.10 and Fig.7.

2) MULTI-LEVEL STRUCTURE EXPERIMENTS

This experiment compares the performance of our multi-level attention structure with that of the single-level attention structure. Table.11 also shows that the classification performance of our MLAN is better than the single-level attention.

3) SMA EXPERIMENTS

This experiment is used to compare the performance of FMA and SMA. The results show that on the CIFAR100, our SMA module also surpasses the FMA (Table.12).

TABLE 11. Multi-level structure experiments on CIFAR100 with ResNet variants of different depth.

structure	ResNet14	ResNet20	ResNet26	ResNet32
single-level	67.06%	69.39%	70.10%	70.67%
multi-level	**68.91%**	**70.51%**	**70.64%**	**71.78%**

TABLE 12. FMA v.s. SMA experiments on CIFAR100.

indicators	ResNet14	ResNet20	ResNet26	ResNet32
FM Acc	70.00%	70.39%	71.24%	72.11%
SM Acc	**68.91%**	**70.51%**	**70.64%**	**71.78%**
FM Paras	1.58M	4.17M	7.93M	12.84M
SM Paras	**0.22M**	**0.36M**	**0.50M**	**0.63M**

TABLE 13. Fusion method experiments on CIFAR100.

indicators	plus	concatenation	1×1 conv	weighted plus
Acc	70.51%	72.31%	71.31%	70.74%
Paras	0.36M	0.81M	0.41M	0.36M

TABLE 14. Comparison of different attention mechanisms for ResNet of different depths on CIFAR100.

works	ResNet14	ResNet20	ResNet26	ResNet32
Ours	**68.91%**	**70.51%**	**70.64%**	**71.78%**
CBAM [34]	67.37%	68.76%	69.89%	69.45%
SE [32]	67.69%	67.57%	68.93%	68.38%
TA [35]	66.85%	68.70%	69.45%	69.34%
GC [37]	67.01%	68.28%	68.71%	67.87%

4) FUSION MODULE EXPERIMENTS

In this experiment, we took ResNet20 to implement the SMA with our proposed fusion methods, and the result is shown in Table.13.

5) COMPARISON WITH BENCHMARKS

Based on the ResNet backbone in Table.8, we also reproduced the classic attention methods, CBAM [34], SEnet [32], TAnet [35], and GCnet [37] to make a comparison. The experimental results are shown in Table.14.

Comprehensive comparison experiment results are shown in Table.9, which indicate the same conclusion similar to Section.IV-A5, and the performance gaps between our proposed SMA and other attention benchmarks become larger.

V. CONCLUSION

In this article, we designed and constructed an MLAN framework with a multi-level structure. Specifically, block and group levels have their own residual structure and aggregated attention. And to reduce the parameters of the attention mechanism, we proposed a lightweight SMA module. Besides, we explored 4 different fusion methods to better integrate information. Based on CIFAR10 and CIFAR100 datasets, we compared our methods with several acknowledged attention methods and the results verified that our method has lower computational complexity, fewer parameters, and higher accuracy.

ACKNOWLEDGMENT

Peinuan Qin and Qinxuan Wang contributed equally to this work and should be considered co-first authors. And both authors are listed in no particular order. This work was supported by the Radar & Countermeasure Laboratory, School of Information and Electronics, Beijing Institute of Technology, Beijing, China.

REFERENCES

[1] D. Hu, "An introductory survey on attention mechanisms in NLP problems," in *Proc. of SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2019, pp. 432–448.
[2] D. P. Morgan and C. L. Scofield, *Natural Language Processing*. New York, NY, USA: Springer, 1991, pp. 245–288, doi: 10.1007/978-1-4615-3950-6.
[3] D. Hu, "An introductory survey on attention mechanisms in NLP problems," in *Proc. SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2019, pp. 432–448.
[4] A. H. Zadeh, I. Edo, O. M. Awad, and A. Moshovos, "GOBO: Quantizing attention-based NLP models for low latency and energy efficient inference," in *Proc. 53rd Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2020, pp. 811–824.
[5] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui, "Attention interpretability across NLP tasks," 2019, *arXiv:1909.11218*.
[6] A. Schapiro and N. Turk-Browne, "Statistical learning," *Brain Mapping*, vol. 3, pp. 501–506, Jan. 2015.
[7] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning. 2001," *J. Roy. Stat. Soc.*, vol. 167, no. 1, p. 192, 2004.
[8] S. M. R. Hashemi and A. Broumandnia, "A review of attention models in image protrusion and object detection," *J. Math. Comput. Sci.*, vol. 15, no. 4, pp. 273–283, 2015.
[9] X. Yao, D. Li, and X. Sun, "Detection of small target in infrared image sequences using attention mechanism," in *Proc. 1st Int. Symp. Syst. Control Aerosp. Astronaut.*, 2006, pp. 456–460.
[10] S. Liu and Z. Cao, "SAR image target detection in complex environments based on improved visual attention algorithm," *EURASIP J. Wireless Commun. Netw.*, vol. 2014, no. 1, pp. 1–8, Dec. 2014.
[11] Y. Yan, J. Ren, G. Sun, H. Zhao, J. Han, X. Li, S. Marshall, and J. Zhan, "Unsupervised image saliency detection with gestalt-laws guided optimization and visual attention based refinement," *Pattern Recognit.*, vol. 79, pp. 65–78, Jul. 2018.
[12] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4945–4949.
[13] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4835–4839.
[14] H. Taherian, "End-to-end attention-based distant speech recognition with highway LSTM," 2016, *arXiv:1610.05361*.
[15] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," 2018, *arXiv:1805.03294*.
[16] H. Xing, G. Zhang, and M. Shang, "Deep learning," *Int. J. Semantic Comput.*, vol. 10, no. 3, pp. 417–439, 2016.
[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, p. 436, Nov. 2016.
[18] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–32, Oct. 2021.
[19] S. Hao, D.-H. Lee, and D. Zhao, "Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system," *Transp. Res. C, Emerg. Technol.*, vol. 107, pp. 287–300, Oct. 2019.
[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[21] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[24] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. Annu. Conf. Neural Inf. Process. Syst.*, vol. 28. Montreal, QB, Canada, Dec. 2015, pp. 2017–2025.
[25] L. Tan, Z. Li, and Q. Yu, "Deep face attributes recognition using spatial transformer network," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Aug. 2016, pp. 1928–1932.
[26] D. Park and S. Y. Chun, "Classification based grasp detection using spatial transformer network," 2018, *arXiv:1803.01356*.
[27] C. Shu, X. Chen, Q. Xie, and H. Han, "Hierarchical spatial transformer network," 2018, *arXiv:1801.09467*.
[28] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," 2017, *arXiv:1704.06904*.
[29] L. Ning, A. Wang, L. Zhao, W. Xue, and D. Bu, "MRANet: Multi-atrous residual attention network for stereo image super-resolution," *J. Vis. Commun. Image Represent.*, vol. 77, May 2021, Art. no. 103115.
[30] D. Qiu, Y. Cheng, and X. Wang, "Gradual back-projection residual attention network for magnetic resonance image super-resolution," *Comput. Methods Programs Biomed.*, vol. 208, Sep. 2021, Art. no. 106252.
[31] Z. Liu, J. Huang, C. Zhu, X. Peng, and X. Du, "Residual attention network using multi-channel dense connections for image super-resolution," *Appl. Intell.*, vol. 51, no. 1, pp. 1–15, 2021.
[32] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[33] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.
[34] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, *CBAM: Convolutional Block Attention Module*. Cham, Switzerland: Springer, 2018.
[35] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3139–3148.
[36] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. CVPR*, Jun. 2018, pp. 7794–7803.
[37] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.
[38] C. Shi, X. Zhang, J. Sun, and L. Wang, "Remote sensing scene image classification based on dense fusion of multi-level features," *Remote Sens.*, vol. 13, no. 21, p. 4379, Oct. 2021.
[39] X. Sun, Q. Zhu, and Q. Qin, "A multi-level convolution pyramid semantic fusion framework for high-resolution remote sensing image scene classification and annotation," *IEEE Access*, vol. 9, pp. 18195–18208, 2021.
[40] K. Zhang, Y. Cai, Y. Ren, R. Ye, and L. He, "MTF-CRNN: Multiscale time-frequency convolutional recurrent neural network for sound event detection," *IEEE Access*, vol. 8, pp. 147337–147348, 2020.
[41] W. Ding and L. He, "Adaptive multi-scale detection of acoustic events," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 294–306, 2020.
[42] J. Guo, Z. Jiang, and D. Jiang, "Multi-level spatial attention network for image data segmentation," *Int. J. Embedded Syst.*, vol. 14, no. 3, pp. 289–299, 1 2021.
[43] M. Wang, C. Li, and F. Ke, "Recurrent multi-level residual and global attention network for single image deraining," *Neural Comput. Appl.*, pp. 1–12, Jan. 2022.
[44] S. Yin, X. Yang, Y. Wang, and Y.-H. Yang, "Visual attention dehazing network with multi-level features refinement and fusion," *Pattern Recognit.*, vol. 118, Oct. 2021, Art. no. 108021.
[45] A. Krizhevsky, *Learning Multiple Layers of Features From Tiny Images*. Toronto, ON, Canada: Univ. Toronto, May 2012.

QINXUAN WANG was born in Linyi, Shandong, China, in 1998. He received the B.S. degree in electronic information engineering from the China University of Geosciences, Wuhan, in 2020. He is currently pursuing the M.S. degree in information and communication engineering with the Beijing Institute of Technology.

From 2018 to 2020, he was a Research Assistant with the Intelligent Image Processing Laboratory, China University of Geosciences. Since 2020, he has been a Research Assistant with the Institute of Radar Technology Countermeasure, Beijing Institute of Technology. His research interests include artificial intelligence, image processing, signal processing, and target detection and recognition (e-mail:wqinxuan4@gmail.com).

PEINUAN QIN was born in Rizhao, Shandong, China, in 1998. He received the B.S. degree in electronic information engineering from the China University of Geosciences, Wuhan, in 2020. He is currently pursuing the M.S. degree in software engineering with The University of Melbourne.

In summer of 2020, he was a Research Assistant with Sun Yat-sen University. In 2021 winter, he was a Research Assistant with the Institute of Software, Chinese Academy of Sciences. His research interests include computer vision, deep learning with medical care, and natural language processing (e-mail: peinuanq@student.unimelb.edu.au).

YUE ZHANG was born in Rizhao, Shandong, China, in 1998. She received the B.S. degree in finance from Shandong University, in 2020. She is currently pursuing the M.S. degree in public management with Tsinghua University. Her research interests include health economics, health policy evaluation, and health care services research.

XUEYAO WEI was born in Baoding, Hebei, China, in 2001. She is currently pursuing the B.S. degree in communication engineering with the China University of Geosciences, Wuhan.

From 2019 to 2022, she was a Research Assistant with the Department of Communication Engineering, China University of Geosciences. She was in the top 2% with a professional weighted grade and was recommended for a master's degree. Her research interests include artificial intelligence, image processing, and signal processing.

MEIGUO GAO received the B.S. and Ph.D. degrees in electronic engineering from the Beijing Institute of Technology, Beijing, China, in 1988 and 1993, respectively. He was a Lecturer with the School of Information and Electronic Engineering, Beijing Institute of Technology, in 1994. He was a Senior Visiting Scholar with the University of Calgary, Canada, in 2002, for a period of six months. He is currently a Full Professor with the Beijing Institute of Technology. His research interests include real-time signal processing with applications to communication, image processing, artificial intelligence, and radar systems.