IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 August 2023, accepted 8 September 2023, date of publication 12 September 2023, date of current version 19 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3314573

RESEARCH ARTICLE

HAT: A Visual Transformer Model for Image Recognition Based on Hierarchical Attention Transformation

XUANYU ZHAO[1], TAO HU[1,2,3], (Member, IEEE), CHUNXIA MAO[1], YE YUAN[4], AND JUN LI[1]
[1]College of Intelligent Systems Science and Engineering, Hubei Minzu University, Enshi 445000, China
[2]Hubei Engineering Research Center of Selenium Food Nutrition and Health Intelligent Technology, Hubei Minzu University, Enshi 445000, China
[3]Key Laboratory of Performing Art Equipment and System Technology, Ministry of Culture and Tourism, Beijing 100007, China
[4]Enshi Audit Office, Enshi 445000, China

Corresponding author: Tao Hu (hutao_es@hbmzu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61962019 and Grant 61972017, in part by the Natural Science Foundation of Hubei Province of China under Grant 2023AFD061, in part by the Science and Technology Program of Enshi of China under Grant D20220005, and in part by the High-Level Cultivation Program of Hubei Minzu University under Grant PY22011.

ABSTRACT In the field of image recognition, Visual Transformer (ViT) has excellent performance. However, ViT, relies on a fixed self-attentive layer, tends to lead to computational redundancy and makes it difficult to maintain the integrity of the image convolutional feature sequence during the training process. Therefore, we proposed a non-normalization hierarchical attention transfer network (HAT), which introduces threshold attention mechanism and multi head attention mechanism after pooling in each layer. The focus of HAT is shifted between local and global, thus flexibly controlling the attention range of image classification. The HAT used the smaller computational complexity to improve it's scalability, which enables it to handle longer feature sequences and balance efficiency and accuracy. HAT removes layer normalization to increase the likelihood of convergence to an optimal level during training. In order to verify the effectiveness of the proposed model, we conducted experiments on image classification and segmentation tasks. The results shows that compared with classical pyramid structured networks and different attention networks, HAT outperformed the benchmark networks on both ImageNet and CIFAR100 datasets.

INDEX TERMS Visual transformer, attention transfer mechanism, hierarchical network, image feature, image recognition.

I. INTRODUCTION

Visual Transformer [1], [2], [3], [4], [5] shows perfect performance in image classification and segmentation, NPL tasks based on stronger global modeling ability, which can solve the problem of elemental dependencies with large spanning in sequence models effectively. With the gradual fusion of the local sensing capability in CNNs with the global coding capability of ViT networks, the hierarchical structure has begun to receive widespread attention in the problem of reducing the computational effort. However, if the CNN focuses too much on local correlations, it may instead impair the model's ability to capture long-term dependencies. If we just use the global relations, the Transformer networks will be unconstrained by local neighborhoods with lacking the structural inductive bias. The result is that the Transformer is more data-dependent and more susceptible to over-fitting on the small datasets.

The self-attention layer [6] is used as the complement of backbone or header network, which can enhance the interaction ability of heterogeneous model. The single attention mechanism is modeled using a strict induction bias, which affects the performance and convergence speed of the model. Due to the acquisition of more adaptive inductive biases, the switching mechanism of multiple attention can greatly improve the efficiency and generalization of data, such as Funnel-Transformer [7] and Swin Transformer [8] etc. It is because of the ViT networks have strong performance in

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

computer vision, we propose a new visual transfer network based on hierarchical attention transformation. The proposed network, denoted as ''HAT'', is used to further enhance the performance of computer vision. HAT use the hierarchical pooling to adjust the feature map. We purify the feature to construct a multilevel hierarchical representation. And we decrease the sampling length gradually to generate a high-resolution feature image.

However, there is a encoding failure problem of feature position in the traditional ViT network by the separate class token. We will maximally pool the remaining sequences in each transformer block to solve this problem, which contain more distinguishing information. The pooling vector is then turned into an additional positional encoding. We will use the positional encoding instead of class token to predict the output.

The traditional ViT networks used the normalization layer [9] to accelerate the speed of training. And it can avoid the disappearing or exploding of gradients in ViT. However, the normalization layer could increase the risk of over-fitting due to the bias and gain [10]. To address this issue, we will reduce the feature breakage problem due to normalization by changing the normalization layer to train model, which can improve the expressive performance of proposed HAT model.

To sum up, our contributions are three-fold as follows:

(1) We propose a new framework HAT with a non-normalization hierarchical attention transformer for image recognition. HAT realizes the serial conversion between multi-head and threshold attentions under a progressive pyramid structure. It enhances the feature learning ability and reduce the computational cost.

(2) We compare the different attention efficiency of the hierarchical visual transformer. We also discuss the influence f the layer normalization in the hierarchical visual Transformer. We quantitatively analyze the reasons why the proposed HAT is beneficial form the removing the normalization layer.

(3) The proposed HAT achieves the state-of-the-art performance of image classification and segmentation on the ImageNet and CIFAR100 datasets.

II. RELATED WORKS

Visual Transformers use the hierarchical structure to deal with long feature sequences effectively [11]. For example, The HVT [12] used a hierarchical stacking approach to construct a hierarchical Transformer. And it used the progressive pooling instead of a single class token to obtain more location information. It gradually merges the visual tokens to improve the accuracy and training efficiency with the depth of the network increasing.

On the other hand, the global and local attention derive from the soft and hard attention, which are widely used in machine vision. The soft attention focuses on regions and channels with a high determinism. Whereas the hard attention emphasizes more on stochastic prediction process and dynamic changes. In contrast, the global attention focuses on the global hidden layer state. which will calculate a weighted average of all input feature sequences and then fed them into the neural network to generate the vectors. The local attention selects a local data in the input sequence, which only considers a subset. Therefore, the introduction of attention in ViT has become the most common practice. For example, the RegionViT [13] proposed a region-to-local attention with a pyramidal structure, which received global information in all tokens by paying attention to local token.

In order to solve the problems that the larger of feature map size in shallow networks and the higher of computational complexity Vit, academics have proposed many innovative methods, such as the spatially separable self-attention [14] and the local-global feature interaction method [15]. However, as the network layers deepen, the attention graphs of each layer will gradually become similar or even identical [16]. To balance the relationship between local and global information attention, Yang et al. [17] decomposed self-attention into Local and Context term. The two terms calculate the activation by observing itself and other respectively, and then extract the weights from the softmax layer. To study the effect of global information on Local Transformer, some scholars have proposed multi-resolution overlapping attention (MOA) module [18], inductive bias soft transfer [19], gate position self-attention (GPSA) [20] and focal self-attention [21].

Although the above related works effectively improved the performance of ViT in image classification and recognition, there are still the problems of large computational overhead, global and local features interacting with each other, and insufficient coverage of network layers by self-attention. In order to solve the above problems, our propose a hierarchical attention transfer network (HAT) based on the different roles of the attention at different stages. And we introduce the threshold attention (TSA) and the multi-head attention (MHSA) after pooling at each layer in HAT, respectively. We use the MHSA to construct the global relationships of sequence blocks and the TSA to model the localization of convolution layer. And the HAT will regulate the degree of attention to location and content information through learnable control parameters. In this paper, we also fuse the normalization layer in the HAT to further improve the performance of computer vision.

III. HIERARCHICAL ATTENTION TRANSFER NETWORK (HAT)

In this section, we will first define the hierarchical attention transfer mechanism and the hierarchical normalization in the HAT. And then, we will describe the structure of HAT in detail.

A. HIERARCHICAL ATTENTION TRANSFER MECHANISM

The attention mechanism in ViT is used to extract dependencies between parts within in a feature sequence by combining different behaviors, which is based on a trainable associative

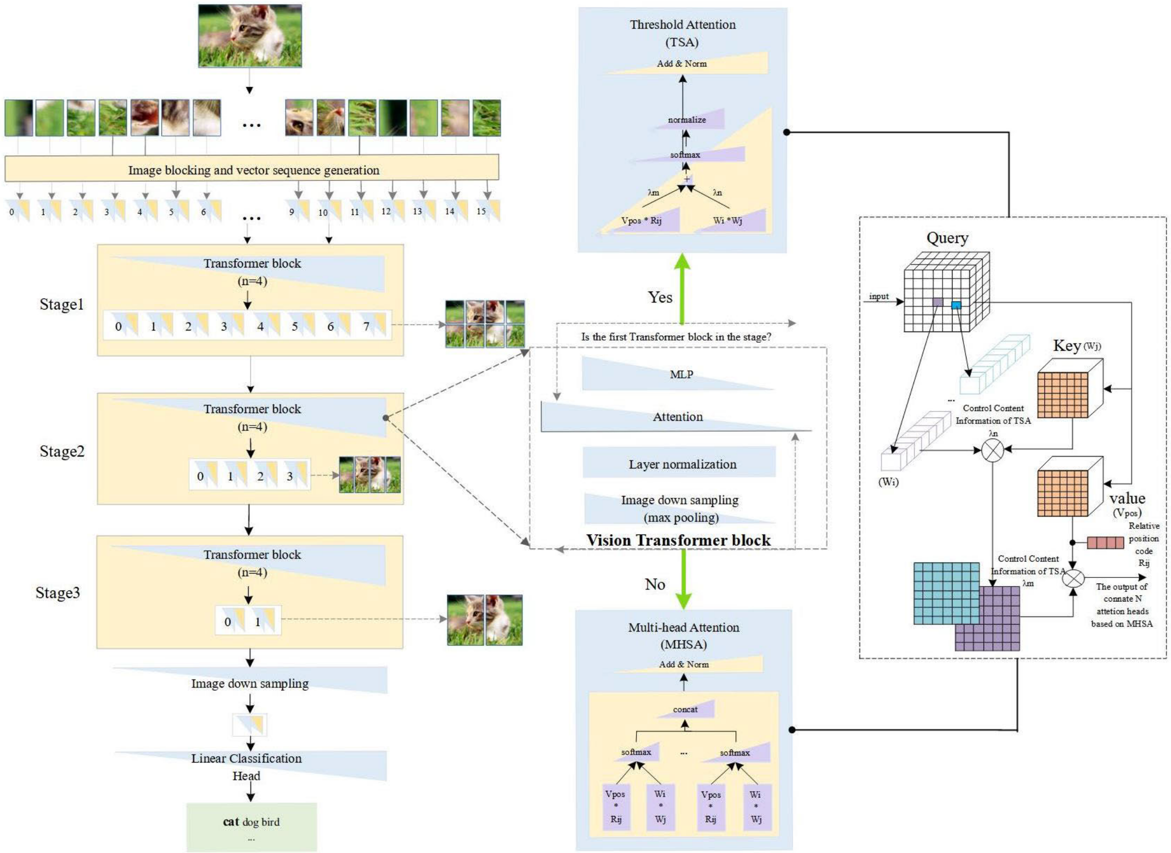


FIGURE 1. The framework of hierarchical attention transformer network.

memory for vectors. There are two typical mechanisms of attention in ViT, the multi-head attention (MHSA) and the threshold attention (TSA) mechanism.

1) THE MULTI-HEAD ATTENTION MECHANISM (MHSA)

ViT uses the MHSA to focus on multiple regions by merging individual attention heads. And MHSA adaptively decides the weight of each input item by inter-playing whitin input term.

2) THE THRESHOLD ATTENTION MECHANISM (TSA)

TSA can be initialized like a convolutional layer and sum the content and location terms after softmax.

Since the relative importance of the two attention is controlled by the learnable threshold $\lambda \in (n, m)$ in each attention head, we define a novel hierarchical attention transfer mechanism for balancing between focusing on local features and recovering global features. And the mechanism finally normalizes the sum of the resulting matrices to make the data feature distribution more stable. The definition of the hierarchical attention transfer mechanism and the normalization layer in the HAT are as followings.

3) THE HIERARCHICAL ATTENTION TRANSFER MECHANISM (HAT)

We assume that the input image X has been initially divided into N blocks. And there are $I-th$ stage in our HAT. At each stage ST_i, We first use the TSA to calculate the threshold attention weight $TSA(X)$ at the first block $b1$. And Then we calculate the multi-head attention weight $MHSA(X)$ in another $b(n-1)$ blocks. We use a linear transformation to calculate the query matrix $Q(W_i, W_j)$, the key representation K, and the value matrix V in each attention mechanism from the feature vector X. So the hierarchical attention weight $HAT(X)$ is defined as following.

$$HAT(X) = ST_i^{b1}(TSA(X)) + ST_i^{b(n-1)}(MHSA(X)) \tag{1}$$

where the $TSA(X)$ and $MHSA(X)$ are described as the equations (2) and (3).

$$\begin{aligned} TSA(X) = norm[&\lambda_n(softmax(Q * K^T)) \\ &+ \lambda_m(softmax(V_{pos}^T * R_{i,j}))]XM \end{aligned} \tag{2}$$

where the M is the weight matrix, V_{pos} is a trainable embedding vector, and R_{ij} is a fixed relative position encoding which is only related to the distance between pixels i and j. The λ_n and λ_m are the upper and lower limits of the threshold λ, respectively. We sum the attention weights with λ_n by calculating the pairwise similarity between query matrix Q and key representation K, and with λ_m by calculating the pairwise similarity between V_{pos}^T and $R_{i,j}$. Finally, the sum of the attention weights is normalized and multiplied by X and

M to obtain the $TSA(X)$.

$$MHSA(X) = concat((softmax(\frac{QK^T}{\sqrt{D}})V)_{1\ldots n})XM \tag{3}$$

where we first calculate the attention weight between Q and K for each head by *softmax*. d is the vector dimension, and $\sqrt{D}$ is the scaling of the attention weights. Then, we connect the attention weights of each head and multiple by X and M to get the $MHSA(X)$.

4) THE LAYER NORMALIZATION (LN)

We use the 'LN' to count the values of all dimensions and channels, means μ and standard deviation σ for each sample $\mathrm{x} = \{\mathrm{x}_1 \ldots \mathrm{x}_n\}$. And we use the b and g in the current layer to denote the deviation and gain obtained from the "LN" vector, respectively. The b and g is used to ensure that the normalization operation does not deviate from the previously information in the current layer. We define the feature z_i, which is transmitted between different layers in HAT, as the equations (4), (5) and (6).

$$Z_0 = \left[x_c; x_p^1E; x_p^2E; \ldots x_p^n\right] + E_{pos} \tag{4}$$

$$LN(Z_{i-1}) = g_{i-1} \odot (\frac{x_{i-1} - \mu_{i-1}}{\sigma_{i-1}}) + b_{i-1} \tag{5}$$

$$Z_i = A(LN(Z_{i-1})) \tag{6}$$

where A denotes the different attention mechanism MHSA or TSA. The z_i is the features extracted from the $i-th$ layer of the Transformer encoder. After blocking and position encoding of image in our HAT, we feed the features $\mathrm{x} = \{\mathrm{x}_1 \ldots \mathrm{x}_n\}$ into the Transformer encoder for analysis. We use the linearly projecting E for the block x_p^i to get the embedded vectorx_p^iE, the class token x_c and the embedded location E_{pos}. HAT get the Z_i in the i_th layer by transmission the image feature between each stage.

B. HIERARCHICAL ATTENTION TRANSFORMER NETWORK

The proposed hierarchical attention transformer network is described as FIGURE 1, which constructs a hierarchical representation by adding a maximum pooling layer. The maximum pooling layer is used to gradually reduce the length of the feature sequences. The hierarchical representation is used to reduce the redundancy of full length patch feature sequences.

Give an input image $H \times W \times C$, where H, W and C denote the height, width, and channel number of the input image, respectively. Similar to the ViT [6], HAT processes the data first by slicing the input image into 16 × 16 non-overlapping patches, and the size of each patch is 14 × 14 pixel. Then, we use the convolution to embed these patches into vectors, which the channel number is 64. HAT spread each patch into a $1D$ vector. After linearly transforming and encoding each vector, the sequence of these patches is propagated into 12 structural blocks with same dimension. The HAT is then divided into M stage, each containing multiple Transformer blocks. The local attention module is introduced at the beginning of the first stage, and HAT recover the perception of the global features at the rest of the stages.

TABLE 1. The feature size and parameters of the HAT network.

Input Type	Sequential Number	Output Size	The parameters
Conv2d	1	[1,384,14,14]	295,296
PatchEmbed	2	[1,196,384]	0
Dropout	3	[1,196,384]	0
LayerNorm	4	[1,196,384]	768
Linear	5	[1,196,768]	295,680
TSA	11	[1,196,384]	0
MLP	19	[1,196,384]	0
MaxPool1d	21	[1,384,97]	0
Block	22	[1,97,384]	0
LayerNorm	23	[1,97,384]	768
Linear	24	[1,97,1152]	443,520
MHSA	28	[1,97,384]	0
MLP	36	[1,97,384]	0
MaxPool1d	38	[1,384,97]	0
Block	209	[1,11,384]	0
Linear	210	[1,200]	77,000

Since ViT get the sequence position information by concatenating an additional learnable vector for classification with the input vector. In this paper, a new positional encoding replaces the position embedding class labeling. HAT uses a $1D$ hierarchical maximum pooling, which the kernel size is k and the step size is s, to gradually reduce the sequence length and the computational cost. We predict and train with non-layer normalization to improve the model performance. Then our network is activated by GELU and calculate the loss by soft target cross entropy loss function. Finally, we use the MLP header for classification.

we use the $\varphi(Block(n, d))$ to calculate the FLOPs cost of TSA and MHSA, where n is the number of labels in the sequence and d is the dimension of each label. The PLOPs of TSA include the projections of $Q \setminus K \setminus V$ matrices $3nd^2$, the computation of the attention map $(\lambda_n + \lambda_m)n^2d$, the self-attention operation n^2d, and the computation of output linearly projecting nd^2, respectively. Similarly, the FLOPs of MHSA include the projections of $Q \setminus K \setminus V$ matrices $3nd^2$, the computation of the attention map n^2d, the self-attention operation n^2d, and the computation of output linearly projecting nd^2, respectively. The FLOPs of MLP consist mainly of two fully connected layers, and each computation is $4nd^2$. We can control the size of block computation by adjusting the threshold in the attention switching, thus reducing the cost of computation. So, the FLOPs of our HAT is described as equation (7).

$$\varphi(Block(n, d)) = \begin{cases} \varphi(TSA(n, d)) + \varphi(MLP(n, d)) \\ = 12nd^2 + (1 + \lambda_n + \lambda_m)n^2d \\ \varphi(MHSA(n, d)) + \varphi(MLP(n, d)) \\ = 12nd^2 + 2n^2d \end{cases} \tag{7}$$

IV. EXPERIMENTS

A. EXPERIMENT SETTINGS

We train and test the proposed HAT network with PyTorch in Linux. The size of input image is 224 × 224 × 3, and the size of patch is 16 × 16. The batch size is 128 in the training.

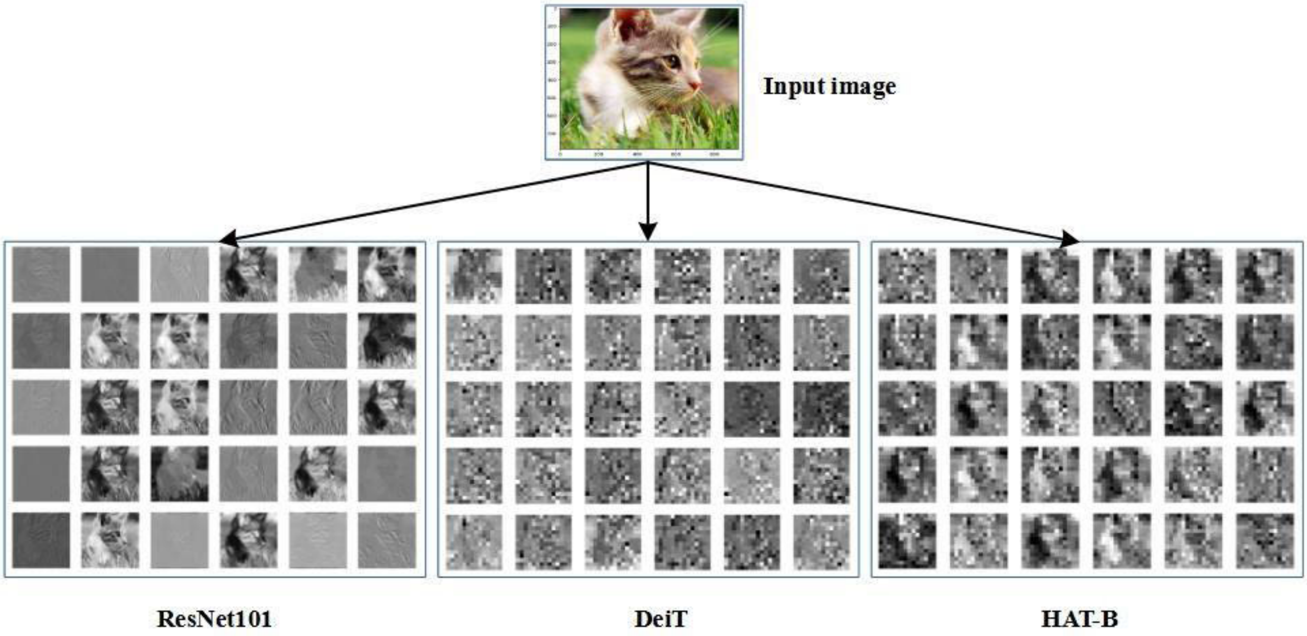


FIGURE 2. The visualization comparison of linear projection features of HAT-B with ResNet101 and DeiT on the ImageNet-1K dataset.

The weight decay coefficient is 0.025. We use the AdamW optimizer [22] to perform back propagation to optimize the network. The feature size and parameters of the HAT network are shown in TABLE 1.

1) DATASETS

In this paper, we evaluate our HAT network on the ImageNet-1K dataset [23]. And we also evaluate the classification performance on the CIFAR-100 [24] and ImageNet200 [25] datasets. The ImageNet-1K dataset is a sub-set of ImageNet, which is conducted on 1k categories. Each category of ImageNet-1K consists of 1300 images for training and 50 images for testing. And the ImageNet200 dataset contains 100000 images of 200 classes images. Each class has 500 training images, 50 validation images and 50 test images. The CIFAR-100 dataset consists of 60000 images. The 100 classes in the CIFAR-100 are grouped into 20 superclasses, which each class has 600 images. There are 500 training images and 100 testing images per class. We use the ADE20K dataset [26] with 150 targets to validate semantic segmentation for the proposed HAT. The ADE20K dataset contains more than 20K scene-centric images with 150 semantic categories.

2) EVALUATION METRICS

We evaluate the performance of our HAT network based on Top-1 and Top-5 accuracy in this paper. We evaluate the pixel accuracy using IoU of overlapping to joint areas between predicted segmentation and labels. And we use the FLOPs(G) to measure the computational cost and the number of Params(M) to evaluate model size.

TABLE 2. The semantic segmentation performances of HAT-A and HVT on the ADE20K validation dataset (*Iteration* = 40*K*).

class	HVT		HAT-B	
	IoU (%)	Acc (%)	IoU (%)	Acc (%)
wall	50.94	**79.85**	52.52	77.90
building	63.88	87.06	64.85	89.69
sky	88.41	94.34	88.87	94.84
floor	49.61	76.61	52.03	78.27
ceiling	59.37	74.52	61.50	82.05
road	58.52	82.31	59.59	82.44
windowpane	36.44	57.68	38.44	59.54
grass	53.64	70.92	59.16	78.75
Average	57.60	77.91	**59.62**	**80.44**

3) BASELINES

We evaluate our proposed method HAT with ResNet50&101 [27], DeiT [19], Swin Transformer [8], HVT [13] and ConViT [20]. In order to conduct the ablation study, we evaluate the hierarchical network with single-head attention, denoted as "HAT-base". Based on HAT-base, we add the threshold attention TSA and the multiple attention MHSA to different blocks after pooling, denoted as "HAT-T" and "HAT-M", respectively. We introduce the hierarchical attention transformation mechanism on HAT-base, denoted as "HAT-TM". Finally, we remove the normalization layer on the HAT-TM, denoted as "HAT". On the other hand, we set different number of attention head to test the performance of network. The number of attention head are 9 and 6, which the network is noted as "HAT-A" and "HAT-B", respectively.

B. FEATURE VISUALIZATION COMPARISON

We first compare the feature vectors of HAT-B with ResNet101 and DeiT on the ImageNet-1K dataset. We show

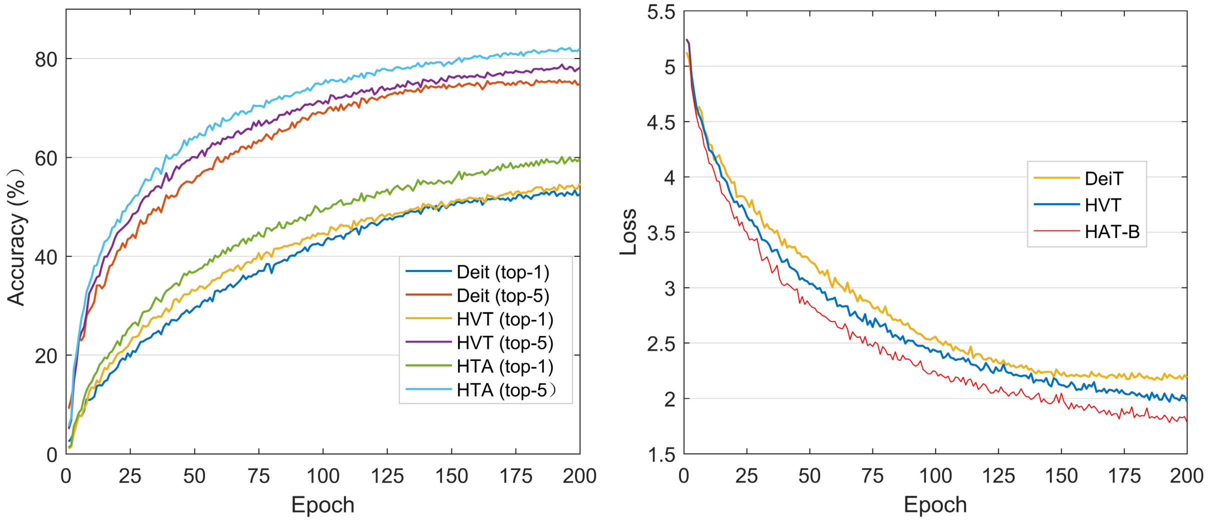


FIGURE 3. Accuracy and loss comparison of HAT-B with HVT and DeiT networks on the ImageNet200 datasets (*Epoch* = 200).

TABLE 3. The Classification performance comparison between HAT-B and other networks on ImageNet200 and CIFAR100 datasets (*Epoch* = 200).

Network	Embedding Dim	Patch Size	Heads	Blocks	FLOPs (G)	Params (M)	ImageNet200		CIFAR100	
							Top-1 Acc.(%)	Top-5 Acc.(%)	Top-1 Acc.(%)	Top-5 Acc.(%)
DeiT	768	16	6	12	16.86	85.95	52.51	74.81	65.29	88.52
HVT	768	16	6	12	2.29	21.67	54.74	78.14	63.10	88.40
HAT-B	768	16	6	12	**2.29**	**22.19**	**59.17**	**81.76**	**69.40**	**91.60**

the feature visualization comparison with those networks as FIGURE 2. As can be seen from the visualization comparison results, the features of HAT are more diverse than DeiT, and it contains richer local information. The feature mapping output after convolutional layer of ResNet tends to retain more edge information for discrimination specifically. In contrast, the image resolution reduces from 32 × 32 to 14 × 14 by linear projection layer in DeiT. So DeiT could learn more structural information than ResNet. On this basis, the proposed HAT down-samples the hidden sequences through a pooling layer, which can construct clearer structural information even at a shallow level.

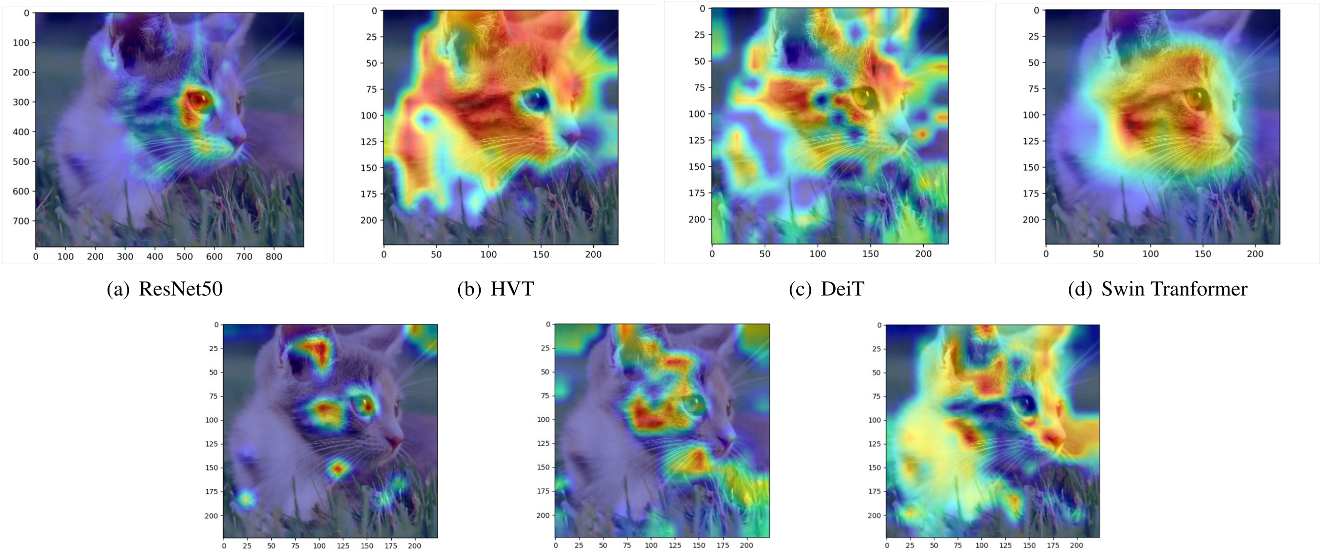
C. THE ABLATION EXPERIMENTS

In order to verify the validity of the proposed HAT, we conduct the ablation experiments with different modules in addition to the visual comparison with the classical networks.

We first compare the proposed HAT-B, which tje number of attention head is 6, with HVT on the DE20K Validation dataset. The IoU and accuracy results are show in TABLE 2. A We can see that the IoU of "glass" Wall in HVT is higher 1.95% than HAT-B. Comparatively, HAT-B has higher IoU and accuracy than HVT in all other categories. So, we can determine that the performance of HAT-B is better than HVT.

The comparative experiments are conducted for the classification performance of HAT-B and other networks following. TABLE 3 shows the classification results of each network on the ImageNet200 and CIFAR100 datasets with the same parameter settings. Compared to the DeiT network, HAT-B improves the Top-1 and Top-5 accuracies on the ImageNet200 dataset by 6.66% and 6.95%, respectively. Top-1 and Top-5 accuracies of HAT-B also improved by 4.43% and 3.62% relative to HVT, respectively. HAT-B similarly achieves the best accuracy on the CIFAR100 dataset. At the same time, The HAT-B is also optimal for both FLOPs and Params on the two datasets. To further analyze the performance of the proposed HAT-B, we compare the accuracy and loss during training on the ImageNet200 dataset. The comparison results are shown in FIGURE 3. As the number of training rounds Epoch increases, the Top-1 and Top-5 accuracies of HAT-B are significantly highest than the corresponding DeiT and HVT. On the other hand, the loss of HAT-B decreased significantly with the increase of Epoch in contrast. In general, the accuracy of HAT-B is improved and the loss is significantly reduced.

The reduction of computational complexity makes HAT highly scalable in terms of network width, depth and tile size. When the network parameters are increased, the computational cost of HAT is still lower than that of the traditional model. In order to confirm the above conclusion, we compare the proposed HAT-A and HAT-B with Swin Transformer and ResNet50. On the one hand, the comparative experiment is to confirm that the proposed HAT is more reasonable in network parameters and cost than other transformer networks. On the other hand, it is also to verify that the accuracy of the proposed HAT is higher when it has an approximate size and cost with a traditional conventional



(a) ResNet50 (b) HVT (c) DeiT (d) Swin Tranformer

(e) The 1-th stage in HAT (f) The 2-th stage in HAT (g) The 3-th stage in HAT

FIGURE 4. Attention-aware heatmaps comparisons between our HAT and other different networks on the ImageNet-1K dataset.

TABLE 4. Comparison of classification performance, network parameters and cost of HAT and other transformer network and traditional conventional network on the ImageNet200 dataset (*Epoch* = 50).

Model	FLOPs (G)	Params (M)	Top-1 Accuracy / Epoch (%)				
			10	20	30	40	50
Swin Transformer	15.14	86.95	11.04	19.26	24.60	30.49	35.70
ResNet50	4.12	23.92	12.95	23.02	30.67	34.93	39.13
HAT-B	**1.34**	**21.95**	**17.06**	**27.14**	33.13	37.00	40.96
HAT-A	1.56	22.00	16.47	26.66	**33.22**	**37.51**	**41.55**

TABLE 5. The classification performance comparison of HAT-B under different number of stages on the CIFAR100 dataset (*epoch* = 150).

Number of stages	FLOPs(G)	Params(M)	Top-1 Acc.(%)	Top-5 Acc.(%)
1	2.29	22.15	64.93	89.62
2	1.86	22.04	66.69	90.28
3	1.56	21.96	67.74	**91.19**
4	**1.34**	**21.91**	**68.71**	91.13

TABLE 6. Comparison of classification performance with different attentions in HAT-B on the ImageNet200 dataset (*Epoch* = 300).

Model	Top-1 Acc. (%)	Top-5 Acc.(%)
HAT-B-T	54.56	77.21
HAT-B-M	57.38	79.44
HAT-B-TM	**62.24**	**82.88**

network. The comparative results are shown as TABLE 4. As the number of attention heads increases, the cost and parameters of the HAT gradually increase from the TABLE 4. The proposed HAT-A and HAT-B require significantly less floating-point computations (FLOPs) compared to Swin Transformer, and the Top-1 accuracies of the top 50 selected Epochs are all improved. On the other hand, HAT-A and HAT-B have similar parameter sizes to the classical CNN model ResNet50, but the computation is reduced by62.14% and 67.48%, respectively. At the same time, the classification accuracies of both HAT-A and HAT-B are higher than that of ResNet50.

The number of stages is also an important factor that affects the performance of the proposed HAT. So, we validate the performance of HAT-B under different of stages on the CIFAR100 dataset. The experimental results are shown as TABLE 5. The comparison results show that the accuracy of the HAT-B gradually increases and the computational cost decreases as the number of stages increases. For example, the accuracy of HAT-B at the 4-th stage is increased by 3.78%, and PLOPs is reduced by 0.95*G* compared to the first stage.

D. COMPARATIVE EXPERIMENTS ON HIERARCHICAL ATTENTION

In order to resolve the problem that the features of the ViT model will gradually decrease with the construction of the hierarchical network, our HAT first gradually adjusts the number of layers and performs down-sampling at the pooling layer. The first Transformer block in each stage of HAT implements adaptive features and correlated local features using a threshold attention mechanism. The other blocks within the same stage use the multi-head attention mechanism

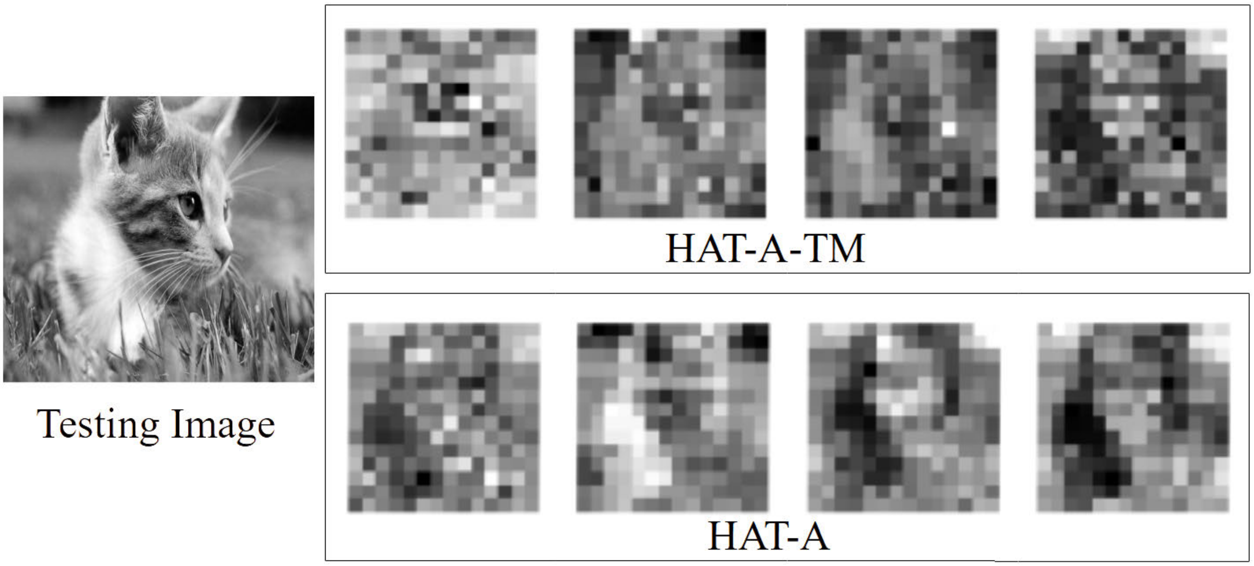


FIGURE 5. Feature visualization of HAT-A under the conditions with and without LN.

to obtain global information, transforming the local to global range of classification predictions.

We first use the Grad-CAM [28], [29], [30] method to localize regions of the image, which are critical for classification, by averaging gradient-weighted activation. FIGURE 4 is the attention-aware comparisons between our HAT and other different networks on the ImageNet-1K dataset. The ResNet50 only localizes a small area of the test image. HVT is focusing on the global features of the test image. Although DetiT focuses on the global features of the test image, the distribution of the global features is scattered and does not describe the category information of the test image well. Swin Transformer's attention map is more focused on the main features of the category of the test image compared to other transformer networks. To better describe the attention map changing belong with different transformer stages in our HAT, we draw the attention-aware heat-maps of the 1-th, 2-th and 3-th stages as FIGURE 4(e), (f) and (g). The attention maps of HAT varies with the hierarchical stages. As the hierarchy progresses in HAT, the scope of attention to features gradually expands from the local to the global. In other words, our HAT enables a transformation of the scope of attention from local to global.

On the ImageNet200 dataset, we also conduct the comparative Experiments in HAT-B between hierarchical attention and threshold attention, multi-head attention. The classification accuracy comparisons are described as TABLE 6, where the "HAT-B-T" means HAT-B with threshold attention, and so on for the rest. From the comparison results, compared with the single attention mechanism, the attention switching mechanism proposed in this paper effectively changes the attention range and can significantly improve the classification performance.

We further verify the effect of different attention head number on the classification performance for the proposed HAT. We compare the FLOPs, Params, Top-1 accuracy and Top-5 accuracy on the ImageNet200 dataset, when the number of attention heads is 4, 8, and 12, respectively. The experimental results are shown as TABLE 7. The FLOPs and Parmas increased with the number of attention head in HAT, while Top-1 classification accuracy gradually increases. TABLE 7 shows that increasing the number of attention heads improves the HAT efficiency, and HAT obtains better performance with increasing network width.

TABLE 7. Top-1 and Top-5 accuracy of HAT with different attention heads on the ImageNet200 dataset (*Epoch* = 300).

Heads	FLOPs (G)	Params (M)	Top-1 Acc.(%)	Top-5 Acc.(%)
4	**0.704**	**9.95**	60.66	**83.02**
8	2.74	38.77	62.23	82.36
12	6.10	86.47	**62.48**	81.85

E. NORMALIZATION VERIFICATION EXPERIMENTS

To explore the effect of layer normalization (LN) on the transformer network, we validate our HAT-A (the attention heads is 9) versus HVT and ConViT on the ImageNet200 and CIFAR100 datasets, respectively. The comparison results are shown in Table 8. In our experiments, we first remove the LN of HVT and ConViT, denoted as HVT-Non LN and ConViT-Non LN. And our HAT-A, which is added the LN, is denoted as HAT-A-TM. Table 8 shows that for HVT and HAT with hierarchical structure, if the LN in the network is removed, the classification accuracy of them will be significantly improved. For example, HVT raises Top-1 accuracy by 0.5% and 1.73% on the two datasets, and HAT raises it by 0.76% and 1%, respectively. On the other hand, the ConViT, which does not have a hierarchical structure, is improved the performance by the layer normalization. This is because there are same 12 blocks in ConViT, and

TABLE 8. Experiments of normalization operation on HAT-A and other Networks on the ImageNet200 and CIFAR100 datasets (*Epoch* = 200).

Model	ImageNet200		CIFAR100	
	Top-1 Acc. (%)	Top-5 Acc. (%)	Top-1 Acc. (%)	Top-5 Acc. (%)
HVT	56.66	79.54	61.66	87.30
HVT-Non LN	57.16	79.75	63.49	87.89
ConViT	55.22	79.68	64.84	89.80
ConViT-Non LN	54.32	79.53	64.69	89.28
HAT-A-TM	59.52	81.47	68.68	**91.41**
HAT-A	**60.28**	**81.97**	**69.68**	91.38

the embedding of each token of ConViT is generated by superimposing other embedding encoding blocks. compared to ConVit-Non LN, Convit's top-1 accuracy improves by 0.9% and 0.15%, respectively.

Consequently, in the hierarchical transformer network, after the feature data is not normalized, the process of finding the optimal solution for each stage becomes less smooth and the likelihood of correctly converging to the optimal level is higher. We also visualized the comparison of the features of HAT-A for the conditions with and without LN, as shown in FIGURE 5. As seen in FIGURE 5, the HAT-A clearly visualizes the features better than HAT-A-TM, and the corresponding contours of HAT-A are more clear.

V. CONCLUSION

In this paper, we proposed a non-normalization hierarchical attention transforming Transformer network (HAT) for computer vision. HAT is able to expand the scope of attention from local to global using serial transformation of threshold attention and multi-head attention. And HAT can effectively decrease the computational cost. Compared with the popular Vision Transformer network, the performance of HAT is significantly improved. In the follow-up work, we will further explore more effective methods in terms of attention mechanism and loss function. And We will continue to optimize the netwrok structure to improve the accuracy and stability for image recognition.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[2] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[3] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.

[4] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "ViTAEv2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," *Int. J. Comput. Vis.*, vol. 131, no. 5, pp. 1141–1162, May 2023.

[5] T. Yao, Y. Li, Y. Pan, Y. Wang, X.-P. Zhang, and T. Mei, "Dual vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 10870–10882, Sep. 2023.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[7] Z. Dai, G. Lai, Y. Yang, and Q. Le, "Funnel-transformer: Filtering out sequential redundancy for efficient language processing," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4271–4282.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[9] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 10524–10533.

[10] H. Zhang, Y. N. Dauphin, and T. Ma, "Fixup initialization: Residual learning without normalization," 2019, *arXiv:1901.09321*.

[11] P. Nawrot, S. Tworkowski, M. Tyrolski, L. Kaiser, Y. Wu, C. Szegedy, and H. Michalewski, "Hierarchical transformers are more efficient language models," in *Proc. Findings Assoc. Comput. Linguistics, NAACL*, 2022, pp. 1559–1571.

[12] Z. Pan, B. Zhuang, J. Liu, H. He, and J. Cai, "Scalable vision transformers with hierarchical pooling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 367–376.

[13] R. Chen, R. Panda, and Q. Fan, "RegionViT: Regional-to-local attention for vision transformers," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–18.

[14] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9355–9366.

[15] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 15908–15919.

[16] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[17] C. Yang, S. Qiao, A. Kortylewski, and A. Yuille, "Locally enhanced self-attention: Combining self-attention and convolution as local and context terms," 2021, *arXiv:2107.05637*.

[18] K. Patel, A. M. Bur, F. Li, and G. Wang, "Aggregating global features into local vision transformer," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 1141–1147.

[19] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[20] S. D'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving vision transformers with soft convolutional inductive biases," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2286–2296.

[21] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal self-attention for local–global interactions in vision transformers," 2021, *arXiv:2107.00641*.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[24] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," M.S. thesis, Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

[25] Y. Le and X. Yang, "Tiny ImageNet visual recognition challenge," Tech. Rep. CS 231N 7.7, 2015, vol. 3.

[26] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, Mar. 2019.

[27] 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.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[29] T. Chakraborty, U. Trehan, K. Mallat, and J.-L. Dugelay, "Generalizing adversarial explanations with grad-CAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 186–192.

[30] S. Zhu, Y. Zhang, and Y. Feng, "GW-Net: An efficient grad-CAM consistency neural network with weakening of random erasing features for semi-supervised person re-identification," *Image Vis. Comput.*, vol. 137, Sep. 2023, Art. no. 104790.

XUANYU ZHAO is currently pursuing the degree with the College of Intelligent Systems Science and Engineering, Hubei Minzu University, Enshi, China. Her research interests include computer vision and privacy protection.

TAO HU (Member, IEEE) received the B.S. degree in software engineering from the School of Computer, Wuhan University of Technology, Wuhan, in 2006, the M.S. degree in software engineering from the School of Software and Microelectronics, Peking University, Beijing, in 2009, and the Ph.D. degree in computer science from the School of Computer Science, Wuhan University, Wuhan, in 2020. He is currently an Associate Professor with the College of Intelligent Systems Science and Engineering, Hubei Minzu University, Enshi, China. His current research interests include deep learning, computer animation, and image processing.

CHUNXIA MAO is currently pursuing the degree with the College of Intelligent Systems Science and Engineering, Hubei Minzu University, Enshi, China. Her current research interests include computer vision and information hiding.

YE YUAN received the degree from the School of Electronic Information, Jingzhou Voctional College of Technology, in 2004. Since 2012, he has been engaged in the application of databases and big data analysis in auditing. He is currently the Director of the Computer Audit Center, Enshi Audit Office, Enshi, China.

JUN LI received the bachelor's degree in mathematics from Hubei University for Nationalities, in 1995, and the master's degree in computer application from the Huazhong University of Science and Technology, in 1999. He is currently a Professor with the College of Intelligent Systems Science and Engineering, Hubei Minzu University, Enshi, China. His research interests include computer graphics and image processing, virtual reality technology, and digital protection of intangible cultural heritage.