Contents lists available at ScienceDirect

AI Open

journal homepage: www.keaipublishing.com/en/journals/ai-open

Associating multiple vision transformer layers for fine-grained image representation

Fayou Sun [a,*], Hea Choon Ngo [b], Yong Wee Sek [b], Zuqiang Meng [c]

[a] Fuzhou Institute of Technology, Fuzhou, Fujian, China
[b] Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[c] Guangxi University, Nanning, 530004, Guangxi, China

ARTICLE INFO

Index Terms:
Increasing vision transformer layers
Weight attention fusion
Feature detection enhancement

ABSTRACT

- Accurate discriminative region proposal has an important effect for fine-grained image recognition. The vision transformer (ViT) brings about a striking effect in computer vision due to its innate multi-head self-attention mechanism. However, the attention maps are gradually similar after certain layers, and since ViT used a classification token to achieve classification, it is unable to effectively select discriminative image patches for fine-grained image classification. To accurately detect discriminative regions, we propose a novel network AMTrans, which efficiently increases layers to learn diverse features and utilizes integrated raw attention maps to capture more salient features. Specifically, we employ DeepViT as backbone to solve the attention collapse issue. Then, we fuse each head attention weight within each layer to produce an attention weight map. After that, we alternatively use recurrent residual refinement blocks to promote salient feature and then utilize the semantic grouping method to propose the discriminative feature region. A lot of experiments prove that AMTrans acquires the SOTA performance on four widely used fine-grained datasets under the same settings, involving Stanford-Cars, Stanford-Dogs, CUB-200-2011, and ImageNet.

1. Introduction

Detecting discriminative regions are critical for fine-grained image recognition, which are challenging tasks due to the subtle yet vital feature learning. As the progress of neural network methods, the performance of fine-grained image recognition tasks achieves great upswing (Woo et al., 2018) (Lin et al., 2015) (Sun and Hea Choon Ngo & Yong Wee Sek, 2022) (Chakraborty et al., 2022) (Jarina et al., 2021) (Olugboja et al., 2021). Currently, weakly-supervision methods with only image-level label are popular approaches (Fayou et al., 2023) (Sun et al., 2022a) (Han et al., 2021) (Zheng et al., 2020) (Jiang et al., 2021). There are two types of network backbones (i.e., CNN-based and ViT-based). The networks of ViT-based are easy to train, lower in complexity and more accurate in capturing subtle discriminant features, which make ViT more valuable in practice.

The models of CNN-based include two categories, i.e., localization and feature-coding approaches. Relatively, localization methods are more interpretable and easier to understand. The former usually trains a discriminative region proposal network and reuse regions to achieve classification. RA-CNN (Fu et al., 2017) proposed recurrent attention CNN to recurrently learn attention maps in three scales. MA-CNN (Zheng et al., 2017) employed channel grouping approach to generate multiple consistency feature vectors by end-to-end training. However, the attention numbers are hyper-parameters, which limit the productivity and flexibility of network. Liu (Liu et al., 2020) et al. proposed filtration and distillation learning network(FDL),which adopted the knowledge distillation method to recurrently detect critical regions. Zheng (Zheng et al., 2019) et al. proposed TASN, which utilized learning trilinear attention sampling network and a feature distiller module to strengthen discriminative regions. Unfortunately, these two networks are difficult to be trained, expanded and have high complexity. The latter methods rely on deep feature representations to achieve better performance for fine-grained image recognition. Yu (Yu et al., 2018) et al. proposed HBP method to do cross-layer bilinear pooling, which verified that the low-level features can compensate for the lack of an object structure feature in high-level semantics. Zheng (Heliang et al., 2019) et al. proposed general block DTB, which used the channel grouping method and group bilinear. Because DTB block keeps consistent feature dimensions between input and output, thus CNN may integrate it into any layer as long as necessary. However, with the increase of network depth, the

[*] Corresponding author.
E-mail address: 314565679@qq.com (F. Sun).

https://doi.org/10.1016/j.aiopen.2023.09.001
Received 23 June 2023; Received in revised form 19 August 2023; Accepted 24 September 2023
Available online 25 September 2023

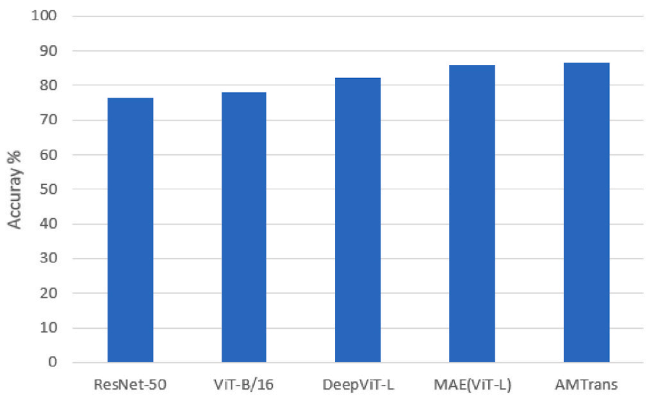

Fig. 1. The accuracy comparison of SOTA networks.

networks are heavy and difficult to explain how to obtain the subtle salient regions.

Recently, some studies innovatively introduced transformer into computer vision tasks, which creates a new era for CV. Visual models of transformer-based developed rapidly from 2019 and there are many achievements worth recommending (e.g., BERT (Devlin et al., 2019), DETR (Carion et al., 2020), iGPT (Chen et al., 2020a), etc.). Dosovitskiy (Alexey Dosovitskiy et al., 2020) et al. presented vision transformer (ViT), which was the 1st to use transformer to solve computer vision tasks. However, it only employs a classification token to detect categories, which is inappropriate for fine-grained representation. He (Ju et al., 2021) et al. presented TransFG, which developed a region selection method to propose a discriminative region, but it cannot generate multiscale fine-grained classification features. To resolve this problem, Zhang (Zhang et al., 2021) et al. presented AFTrans, which adaptively selects relatively sensitive patches for optimizing regions proposal.Wang (Wang et al., 2021) et al. proposed FFVT, which adopted the feature fusion ViT to select the best significant tokens within each encoder layer as the inputs of the last layer. However, all of above studies are limited by the depth of transformer encoder layers, thus they only fuse narrow features.

To solve these problems, our research proposes a novel model AMTrans, which employs re-attention to replace multi-head self-attention mechanism to raise the depth of transformer encoder layers. Then, we utilize the feature fusion method to enhance the salient feature map. To be specific, we use DeepViT (Zhou et al., 2021) to increase the number of layers. Concurrently, this research fuses all attention weights within every transformer encoder layer and then integrates the shallow level features and deep level features as the input of recurrent residual refinement blocks(RRBs). Subsequently, the salient feature map will be output from RRBs, which is the input of channel attention module that will propose the most vital region of the input image. Finally, the proposed region will be the input of our model again to achieve classification. The AMTrans outperforms existing networks on ImageNet (Jia et al., 2009) as shown in Fig. 1. The contributions of this research are as follows:

- To capture more diverse features, we increase the depth of layers and fuse the attention weights within each transformer encoder layer.
- To saliently enhance the discriminative region proposal, we utilize recurrent residual refinement blocks to improve salient feature detection and then utilize semantic grouping method to select the excellent discriminative region in the input image.
- To be our best knowledge, this research is the 1st to successfully use increasing the depth of layers for fusing more attention weight.

2. Related work

This section discusses currently approaches closely this research, including attention weight fusion, selecting discriminative region.

Attention Weight Fusion: There has been a growing interest in exploring attention mechanisms in vision transformers, which have shown significant performance in computer vision tasks (e.g., image classification, object detection, semantic segmentation, etc.).

One path in this field has focused on the attention weight fusion from multiple levels in the transformer encoder layers. These methods aim to capture both local and global information by combining the attention weights of different transformer blocks or attention heads. Zhang (Nam et al., 2017) et al. proposed a dual-attention network that combines attention weights from both image and text modalities to perform multimodal reasoning and matching. This model used a “cross-modal attention fusion” method to combine the attention weights from different modalities. Liu (Liu et al., 2021) et al. proposed a hierarchical transformer architecture that fuses attention weights across different scales. This method introduced a new “Shifted Window” way that allows the model to capture multiscale features efficiently and proposed a “Swish” fusion method to combine the attention weights of different levels in the transformer. Chen (Chen et al., 2021) et al. proposed a cross-attention multiscale vision transformer for image classification tasks. The model introduced a cross-level fusion approach that aggregates the attention weights from different levels of the network to enhance the feature representation. In addition to these methods, several other works have explored the use of fusion attention weights in ViT, including the Hierarchical attention Vision Transformer (Hu et al., 2023) (HAVT), the Trans2Seg (Xie et al., 2021), etc. These methods demonstrate improved performance on various benchmarks, highlighting the effectiveness of fusion attention weights in ViT.

In view of this, as fusing attention weights is an effective method, this research fuses the attention weights of each head based on the characteristics of fine-grained image classification to achieve the promised results.

Selecting Discriminative Region: Fine-grained image classification is a challenging task that requires identifying subtle differences between similar objects within a fine-grained category. One way to improve the performance of fine-grained classification is to focus on discriminative regions within an image.

Woo (Woo et al., 2018) et al. proposed CBAM, which uses both channel attention and spatial attention. It has been used for fine-grained image classification tasks, where it learns to attend to discriminative regions while suppressing irrelevant regions. Xu (Xu et al., 2022) et al. proposed ADDS, which generated region proposals at multiple scales and then combined them to identify the most discriminative regions in an image. The model first applies a convolutional network to the input image to generate feature maps at different scales. These feature maps are then used to generate region proposals, which are combined to identify the most discriminative regions in the image. Zhong (Zhong et al., 2021) et al. proposed STAN, which used a spatial transformer module to identify discriminative regions in an image. The model first applies a localization network to the input image, which generates a set of transformation parameters that are used to warp the input image. The warped image is then fed into a classification network, which is trained to classify the image based on the warped features. Zhang (Zhang et al., 2019) et al. proposed an attention guided network that uses a self-attention mechanism to select discriminative regions. The model learns to attend to the most informative regions of an image and then aggregates the feature vectors of these regions for classification.

As is known to all, selecting discriminative regions is an important aspect of fine-grained image classification. Different methods have been used to identify these regions, including attention mechanisms, part-based methods, localization techniques, and fine-grained object retrieval. Due to the use of deepViT as backbone in this research, R^3Net and semantic grouping modules were added to the network to

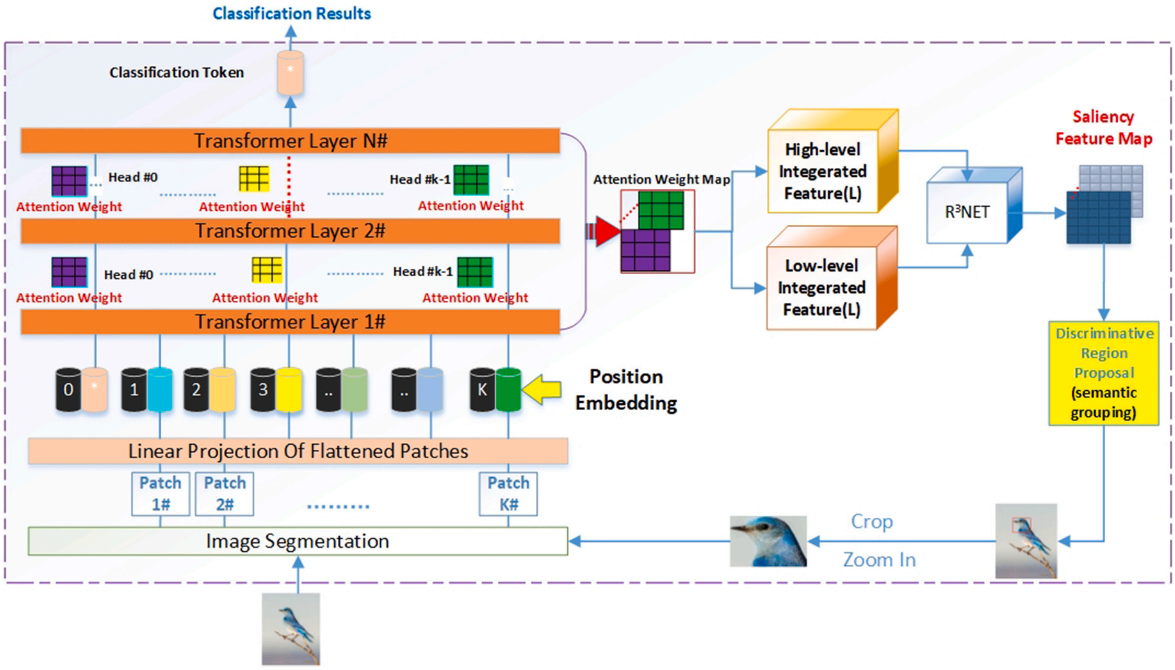

Fig. 2. The architecture of AMTrans.

implement “attention in attention”.

3. Method

This section introduces our model AMTrans, which consists of three parts (i.e., fusion attention weight, salient feature detection and discriminative region proposal).

An overview of AMTrans is shown in Fig. 2. The backbone of our model is DeepViT (Zhou et al., 2021), which focus on fusing attention weight within each transformer encoder layer and then utilizes R^{3}NET (Deng et al., 2018) to reinforce salient feature for the critical region proposal.

In Fig. 2, the image is divided into blocks of same size, which are the input of DeepViT. This study utilizes the Hadamard product to fuse the multihead attention weights of all layers according to the head and then generates attention weight map by concatenation them. Subsequently, we employ R^{3}NET to achieve salient feature enhancement and then propose a discriminative region. Finally, we cut and enlarge the selected region on the input image, which is the input of DeepViT. From Fig. 2, it can be seen that the dimension of attention weigh map is $D \in R^{B \times L \times K \times P \times P}$ (B=batch size, L= the number of transformer layers, K = the quantity of head, P =the count of patches). After that, we split the feature map into shallow level features (L) (i.e., from layer 1# to layer 16#) and deep level features (H) (i.e., from layer 17# to layer 32#) as the input of R^{3}NET, which generates saliency feature map (the input and output dimension of R^{3}NET is without changing). To the end, we utilize a CNN to propose critical region.

3.1. Attention weight fusion

To avoid attention collapse issue (i.e., as the transformer goes deeper, the attention maps gradually become similar and even much the same after certain layers.), we utilize a simple yet effective method DeepViT (Zhou et al., 2021) to stack more transformer encoder layers to increase diversity of attention weights with negligible cost. DeepViT (Zhou et al., 2021) adopted re-attention to replace self-attention mechanism. Specifically, the output of re-attention is:

$$Z(Q, K, V) = \sigma\left(\theta^T \times \delta\left(\frac{QK^T}{\sqrt{d_k}}\right)\right) \tag{1}$$

where σ is normalization, δ is softmax, $\theta \in R^{H \times W}$ is a learnable transformation matrix, Q is Query tensor, K is Key tensor, V is Value tensor, and d_k is the dimension of Key tensor.

In this paper, we use pre-trained DeepViT-32B (Zhou et al., 2021)

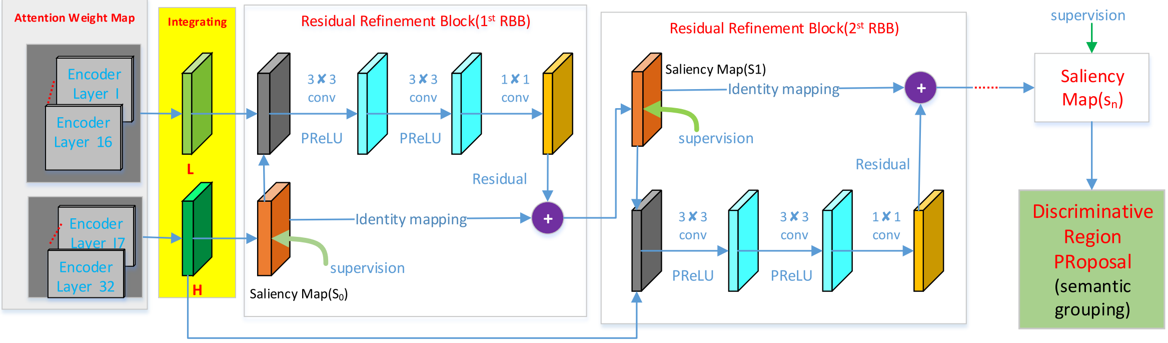

Fig. 3. The structure schematic of R^{3}NET.

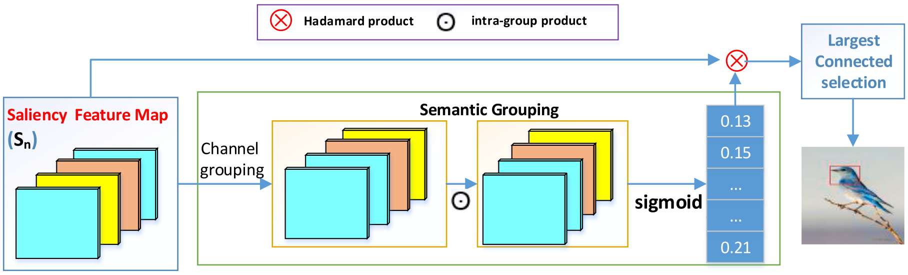

Fig. 4. The overview of discriminative region proposal.

network, which is composed of 32 layers. To be our best knowledge, every head represents a different region over the image within each layer. Thus, we adopt element-wise product to fuse all attention weights in every encoder layer according to K multi-heads to reinforce effective attention features. The attention weight map of the m-th head in each layer is as follows:

$$H^m = \alpha_m^{p\times p} \tag{2}$$

where P is the count of patches. Then we can do Hadamard product on each head of all layers grouped by head. Thus, the final attention weight map of the m-th head is as follows:

$$W^m = {}_1^N\prod H^m \tag{3}$$

where $\prod$ is hadamard product, N =the count of layers and $W^m \in R^{B\times P\times P}$ (B= batch size and P is the count of patches).

Thus, the final fused attention weight map is as follows:

$$F = \mathrm{concat}\left(W^0, .., W^{K-1}\right) \tag{4}$$

where ″concat″ is concatenation operation, K is the number of head and $F \in R^{\mathrm{B}\times\mathrm{K}\times\mathrm{P}\times\mathrm{P}}$.

3.2. *Feature detection enhancement*

Detecting the subtle feature is the soul of fine-grained image representation, but it is a difficult task. To resolve this matter, we take the R^{3}NET (Deng et al., 2018) to enhance salient feature. The overview of R^{3}NET is shown in Fig. 3.

In Fig. 3, the original saliency map (S_0) is H, which is many times optimized by some residual refinement block (RRB). At the same time, from Fig. 3, it can be observed that R^{3}NET utilize integrated shallow level features to capturing more saliency details, which compensate for the weakness that deep level features only rely on rich semantic features. As far as we know, the RRB can accurately propose salient feature regions on the input image. An RRB is defined as:

$$R_k = \partial(\varphi(S_{k-1}, M))$$

$$S_k = S_{k-1} + R_k \tag{5}$$

where ∂ is CNN, "φ" is concatenation operation, S_{k-1} is the predicted saliency map of the (k-1)-th step, and the feature map M is alternatively set as an integrated shallow level feature or integrated deep level feature. In this work, we use three RBBs.

In our research, because the count of encoder layers of DeepViT is 32, we set the range of shallow level layers are {1–15} and the deep level layers are {16–32}.

3.3. *Discriminative region proposal*

Capturing the critical and subtle region is the core of fine-grained task. We utilize output of R^{3}NET(S_n) as input feature for selecting the discriminative region. The process is shown in Fig. 4.

Table 1
Comparison performance on CUB-200-2011,Stanford-Dogs, Stanford-Cars.

Method	Backbone	Acc. (%)		
		CUB-200-2011	Stanford-Cars	Stanford-Dogs
GP-256 (Wei et al., 2018)	VGG16	85.8	–	83.1
ResNet50 (He et al., 2016)	ResNet50	84.5	91.7	82.7
TASN (Zheng et al., 2019)	VGG19	86.1	92.4	–
FDL (Liu et al., 2020)	VGG19	86.84	91.52	84.9
DCL (Sun et al., 2022b)	Resnet50	87.4	94.5	–
S3N (Ding et al., 2019)	Resnet50	88.5	94.7	87.1
ViT (Alexey Dosovitskiy et al., 2020)	ViT_B/16	90.2	93.5	91.2
TransFG (Ju et al., 2021)	ViT_B/16	90.9	94.1	90.4
RAMS-Trans (Hu et al., 2021)	ViT_B/16	91.5	–	90.7
HAVT (Hu et al., 2023)	ViT_B/16	91.8	–	91.0
AMTrans	DeepViT-32B	93.1	96.8	92.7

In Fig. 4, we take the semantic grouping (SG) to obtain the relative weight parameters of the regions and then use Hadamard product with S_n to reinforce what to pay attention to. Finally, we use the largest connected selection method to select the best discriminative region. It can be observed that SG consists of channel grouping and intra-group strengthen. The outputs of SG can be denoted as:

$$R = \vartheta(S_n)$$

$$G = \omega(R)$$

$$C_o = \epsilon\ (G) \tag{6}$$

where ϑ is channel grouping method (fastcluster (Müllner, 2013)), ω is matrix product in each intra-group, and ϵ is sigmoid. The output of SG denoted C_o. Hence, the refined feature can be denoted as:

$$T = S_n \otimes C_o \tag{7}$$

where S_n and C_o conduct Hadamard product and the dimension of T is same as S_n.

Finally, this study utilizes the largest connected region selecting method to capture the excellent discriminative region from T to cut this region on the input image. Finally, this region is amplified as the input of the DeepViT.

4. Experiments

This section reflects the advantage of AMTrans on widely used fine-grained datasets, that is, CUB-200-2011 (Berg et al., 2014),

Table 2
Comparison of SOTA methods on ImageNet1K.

Method	Backbone	ImageNet		
		Top1-Acc. (%)	#param (M)	GFLOPs
ResNet152 (He et al., 2016)	ResNet/152	75.3	60.2	11.3
EfficientNet (Tan and Le, 2019)	EfficientNet-B7	84.4	66	–
ViT (Alexey Dosovitskiy et al., 2020)	ViT-L/16	76.53	307	4.6
iGPT (Chen et al., 2020b)	iGPT-L	69	1362	41
Moco V3 (Xinlei et al., 2021)	ResNet-50	73.9	81	4.1
DeiT-B (Touvron et al., 2021)	ViT-B/16	84.4	86	17.6
MAE (He et al., 2022)	ViT-L/16	85.9	–	–
AMTrans	ViT-B/16	86.6	321	3.9

Table 3
Comparison of Top1 accuracy on ImageNet1K using different classification networks.

Method	ImageNet
	Top1-Acc. (%)
SG+MLP	85.1
SG+DeepViT	86.6

Table 4
Ablation experiment on different backbone.

Backbone	Acc. (%)
DeepViT_16B	91.4
DeepViT_24B	92.3
ViT_32B	91.6
DeepViT_32B	93.1
DeepViT_44B	93.4

Table 5
Ablation experiment on R^{3}NET, discriminative region proposal.

Method	Acc. (%)
baseline	90.7
+R^{3}Net	91.8
+ semantic grouping (SG)	92.3
+R^{3}Net+SG	93.1

Stanford-Dogs (Khosla et al., 2011), Stanford-Cars (Krause et al., 2013), and ImageNet (Jia et al., 2009). The code of AMTrans is https://github.com/dlearing/AMTrans.git.

Next, we initialize the parameters for the experiment. For fair comparison, the image size is 224 × 224 and every patch size is 16 × 16. Batch size adopts a generic value 256.Then we adopt the well-trained DeepViT-32B network on ImageNet1k (Jia et al., 2009) and the well-trained R^{3}NET network on MSRA10K. Concurrently, we employ SGD optimizer and fixed learning rate 0.0002. AMTrans is trained on 2 T V100 GPUs with Pytorch as our code-base. Specifically, AMTrans also is pre-trained on ImageNet1k (Jia et al., 2009).

4.1. Performance comparison

To prove the performance of network, AMTrans compares with current SOTA approaches on three benchmarks. From Table 1, it can be observed that AMTrans gets good results and surpasses CNN-based methods and ViT-based methods.

Compared with the good ViT-based model HAVT, AMTrans brings further 1.3% gain and reaches 93.1% on CUB-200-2011. Typically, we can stack the depth of CNN and integrate some tricks in CNN (e.g.,TASN (Zheng et al., 2019), DCL (Sun et al., 2022b), etc.) to boost the performance for fine-grained image representation, but the gain is not obvious. The vision transformer methods (e.g., TransFG (Ju et al., 2021), HAVT (Hu et al., 2023), etc.) rely on self-attention mechanism to represent a great potential for feature representation. In view of this, we adopt DeepViT as backbone and CNN as necessary and profitable compensation to achieve fine-grained image recognition.

From the 4th column of Table 1, we know that AMTrans achieves 2.1% improvement than S3N (Ding et al., 2019) and surpasses all CNN-based methods. We argue that all methods can obtain better results on this dataset due to less background noise over images. In terms of accuracy, our model brings 2.7% gains compared to TransFG (Ju et al., 2021).We believe that feature fusion and salient region proposal are the main reasons.

From the 5th column of Table 1, we can know that vision transformer methods surpass CNN -based models.We analysis that the reason is the hard-to-find inter-class diversities between certain objects on Stanford-Dogs (Khosla et al., 2011). Hence, it proves the advantage of vision transforme. However, AMTrans still gets the best performance and reaches 92.7%,which brings 1.7% gains compared to HAVT (Hu et al., 2023).

In order to better compare the performance of our models with similar volumes, we conducted comparative experiments on other quantitative metrics (i.e., parameters and GFLOPS). From Table 2, it can be observed that AMTrans obtains 1.3% gains than self-supervised model MAE, which proves that weakly supervised methods still have advantages. At the same time, the Params and FLPOs metrics have decreased. We analyze that it is due to the increase in the depth of ViT (DeepViT) and the addition of CNN modules (i.e., R^{3}Net and semantic grouping).

In Table 3, SG is semantic grouping module. It can be observed that DeepViT can bring 1.1% improvement. This result further proves the innate advantage of the multi-head self-attention mechanism in ViT for computer vision tasks.

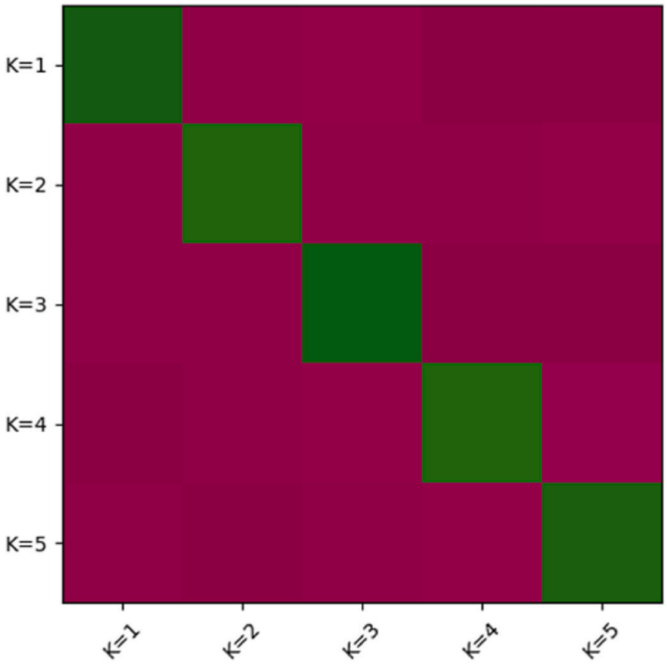

Fig. 5. A confusion matrix of inter-group pairwise interactions. Green and red represent large and small values, respectively.

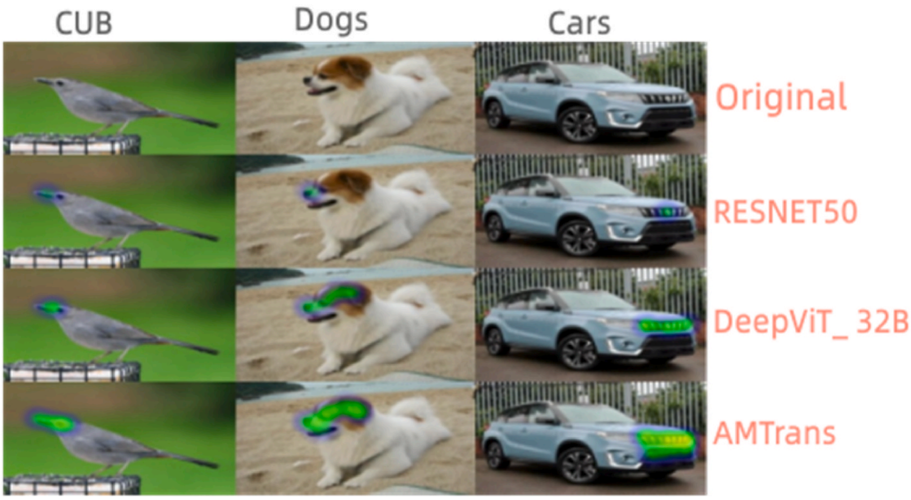

Fig. 6. Visualization of discriminative region proposal on three benchmarks.

4.2. Ablation studies

This section shows the influence of every part in AMTrans on ablation studies. We conduct all studies on CUB-200-2011 (Berg et al., 2014) and other datasets have the same results as well.

Our model adopts DeepViT (Zhou et al., 2021), which significantly promotes the diversity of features. The benefit it brings is shown in Table 4.

From Table 4, we observe that it is beneficial for improving performance by stacking the transformer encoder layers. We believe that the multi-head re-attention mechanism can increase discriminative feature diversity. Specifically, DeepViT_32B brings 1.5% improvement compared to ViT_32B. However, if we increase the transformer encoder layer to 44, the gain of accuracy is only 0.3% and the time complexity will increase. Thus, this research employs 32 layers.

In Table 5, the DeepViT_32B is baseline. From Tables 5 and it can be observed that R^3Net can bring 1.1% gain. Thus, fusing all levels of features can be beneficial and reinforce the information of the region of interest. Concurrently, it can be observed that SG can bring 1.6% gain. We analysis that SG uses the channel grouping and intra-group strengthen, which can focus on discriminative informative features and suppress less useful ones. Hence, our model associates DeepViT with R^3NET and SG.

In the semantic grouping module, this research uses fastcluster (Müllner, 2013). At the same time, it is found by experiments that the performance is better when the number of groups is set to k=5.

From Fig. 5, it can be observed that the semantic grouping module clusters similar features and there are already significant feature differences between inter-group. This result can also prove that semantic grouping can achieve “attention in attention”.

4.3. Visualization experiments

We randomly select an image from each dataset and then do the visualization experiment. The result is as shown in Fig. 6. To verify the strength of AMTrans, we conduct a comparative experiment.

From Fig. 6, it can be observed that ViT-based methods can capture better discriminative parts with subtle critical features than CNN-based methods. Meanwhile, AMTrans obtains the most discriminative region as shown in the 5th row.

5. Conclusion

This research puts forward a novel model AMTrans, which achieves SOTA performance on widely used datasets. To resolve the attention collapse problem, this research uses the DeepViT. Thus, AMTrans can increase the depth of transformer encoder layers to obtain more diverse features and then fuse the attention weight within each layer to reinforce feature representation. Concurrently, this research utilizes multiple recurrent residual refinement blocks to prompt the discriminative features and suppress noise features. Finally, we adopt the semantic grouping method to capture what to pay attention to select the discriminative regions. At the same time, AMTrans obtains the promised results on four fine-grained benchmarks: CUB-200-2011, Stanford-Dogs, Stanford-Cars and ImageNet. In the future, we will conduct data fusion (e.g., internet data, videos, text, etc.) and self-supervised methods to obtain the progress of performance for fine-grained image representation.

Data availability

The datasets used in this study can be downloaded from.

- https://github.com/topics/cub200-2011
- https://github.com/sigopt/stanford-car-classification
- http://vision.stanford.edu/aditya86/ImageNetDogs/main.html

Author contributions

Investigation: Sun Fayou, Zuqiang Meng.
Methodology: Sun Fayou, Hea Choon Ngo.
Software: Sun Fayou, Yong Wee Sek.
Writing–original draft: Sun Fayou, Hea Choon Ngo, Yong Wee Sek, Zuqiang Meng.
Writing–review & editing: Zuqiang Meng.

Conflict of interest statement

The authors do not have any possible conflicts of interest.

Acknowledgement

This work is supported by Universiti Teknikal Malaysia Melaka (UTeM), Guangxi University, and National Natural Science Foundation of China (No. 62266004).

References

Alexey Dosovitskiy, Beyer, Lucas, Alexander, Kolesnikov, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, Uszkoreit, Jakob, Houlsby, Neil, 2020. An image is worth 16x16 words: transformers for image recognition at scale, 2 (3), 6.

Berg, Thomas, Liu, Jiongxin, Lee, Seung Woo, Alexander, Michelle L., Jacobs, David W., Belhumeur, Peter N., 2014. Birdsnap: large-scale fine-grained visual categorization of birds. In: CVPR, pp. 2011–2018, 1.

Carion, N., Massa, F., Synnaeve, G., et al., 2020. End-to-end Object Detection with transformers[C]//European Conference on Computer Vision. Springer, Cham, pp. 213–229.

Chakraborty, S., Amrita, Choudhury, T., Sille, R., Dutta, C., Dewangan, B.K., 2022. Multi-view deep cnn for automated target recognition and classification of synthetic aperture radar image. J. Adv. Inf. Technol. (5), 13.

Chen, M., Radford, A., Child, R., et al., 2020a. Generative pretraining from pixels[C]// International conference on machine learning. PMLR 1691–1703.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I., 2020b. Generative pretraining from pixels. In: International Conference on Machine Learning. PMLR, pp. 1691–1703.

Chen, C.F.R., Fan, Q., Panda, R., 2021. Crossvit: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 357–366.

Deng, Zijun, et al., 2018. R^{3}Net: recurrent residual refinement network for saliency detection. In: International Joint Conference on Artificial Intelligence, pp. 684–690.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., Jiao, J., 2019. Selective sparse sampling for fine-grained image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6599–6608.

Fayou, S., Ngo, H.C., Meng, Z., Sek, Y.W., 2023. Loop and distillation: attention weights fusion transformer for fine-grained representation. IET Computer Vision 17 (4), 473–482.

Fu, J., Zheng, H., Mei, T., 2017. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR), Jul., pp. 4438–4446

Han, K., Xiao, A., Wu, E., et al., 2021. Transformer in transformer[J] arXiv preprint arXiv:2103.00112.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009.

Heliang, Zheng, et al., 2019. Learning deep bilinear transformation for fine-grained image representation. Neural Information Processing Systems 32, 4279–4288.

Hu, Y., Jin, X., Zhang, Y., Hong, H., Zhang, J., He, Y., Xue, H., 2021. Rams-trans: recurrent attention multi-scale transformer for fine-grained image recognition. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4239–4248.

Hu, X., Zhu, S., Peng, T., 2023. Hierarchical attention vision transformer for fine-grained visual classification. J. Vis. Commun. Image Represent., 103755

Jarina, R.A., Abas, P.E., Silva, L., 2021. A simulated water type dataset (swtd) based on jerlov water types for underwater image quality analysis. J. Adv. Inf. Technol. (4), 12.

Jia, Deng, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, Fei-Fei, Li, 2009. Imagenet: a large-scale hierarchical image database. In: CVPR, vol. 1, p. 7, 5, 6.

Jiang, Y., Chang, S., Wang, Z., 2021. Transgan: Two Transformers Can Make One Strong gan[J] arXiv preprint arXiv:2102.07074.

Ju, He, Chen, Jie-Neng, Liu, Shuai, Adam, Kortylewski, Cheng, Yang, Bai, Yutong, Wang, Changhu, Yuille, Alan, 2021. Transfg: A Transformer Architecture for Fine-Grained Recognition arXiv preprint arXiv:2103.07976.

Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.-F., 2011. Novel dataset for fine-grained image categorization. In: ICCV Workshop.

Krause, Jonathan, Stark, Michael, Jia, Deng, Fei-Fei, Li, 2013. 3D object representations for fine-grained categorization. ICCV Workshop 1, 5.

Lin, T.-Y., RoyChowdhury, A., Maji, S., 2015. "Bilinear CNN models for finegrained visual recognition,". Proc. IEEE Int. Conf. Comput. Vis. (ICCV) 1449–1457.

Liu, C., Xie, H., Zha, Z.-J., Ma, L., Yu, L., Zhang, Y., 2020. Filtration and distillation: enhancing region attention for fine-grained visual categorization. Proc. AAAI Conf. Artif. Intell. 34, 11555–11562.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al., 2021. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Müllner, D., 2013. Fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. J. Stat. Software 53 (9), 1–18.

Nam, H., Ha, J.W., Kim, J., 2017. Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 299–307.

Olugboja, A., Wang, Z., Sun, Y., 2021. Parallel convolutional neural networks for object detection. J. Adv. Inf. Technol. (4), 12.

Sun, Fayou, Hea Choon Ngo & Yong Wee Sek, 2022. Combining multi-feature regions for fine-grained image recognition. Int. J. Image Graph. Signal Process. (1).

Sun, f., Ngo, H.C., Sek, Y.W., 2022a. Adopting attention and cross-layer features for fine-grained representation. IEEE Access 10, 82376–82383.

Sun, K., Yao, T., Chen, S., Ding, S., Li, J., Ji, R., 2022b. Dual contrastive learning for general face forgery detection. Proc. AAAI Conf. Artif. Intell. 36 (No. 2), 2316–2324.

Tan, M., Le, Q., 2019. Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR, pp. 10347–10357.

Wang, Jun, et al., 2021. Feature Fusion Vision Transformer Fine-Grained Visual Categorization.

Wei, X., Zhang, Y., Gong, Y., Zhang, J., Zheng, N., 2018. Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 355–370.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. "CBAM: convolutional block attention module,". In: Proc. Eur. Conf. Comput. Vis. (ECCV), Sep., pp. 3–19.

Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., Luo, P., 2021. Segmenting Transparent Object in the Wild with Transformer arXiv preprint arXiv:2101.08461.

Xinlei, C., Saining, X., Kaiming, H., 2021. An Empirical Study of Training Self-Supervised Visual Transformers, p. 8 arXiv preprint arXiv:2104.02057.

Xu, S., Li, Y., Hsiao, J., Ho, C., Qi, Z., 2022. A Dual Modality Approach for (Zero-Shot) Multi-Label Classification arXiv preprint arXiv:2208.09562.

Yu, Chaojian, et al., 2018. Hierarchical bilinear pooling for fine-grained visual recognition. European Conference on Computer Vision 11220, 595–610.

Zhang, S., Fu, H., Yan, Y., Zhang, Y., Wu, Q., Yang, M., et al., 2019. Attention guided network for retinal image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, I 22. Springer International Publishing, Shenzhen, China, pp. 797–805. October 13–17, 2019, Proceedings, Part.

Zhang, Y., et al., 2021. A Free Lunch from ViT: Adaptive Attention Multi-Scale Fusion Transformer for Fine-grained Visual Recognition.

Zheng, H., Fu, J., Mei, T., Luo, J., 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 5209–5217.

Zheng, H., Fu, J., Zha, Z.J., Luo, J., 2019. Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 5012–5021.

Zheng, S., Lu, J., Zhao, H., et al., 2020. Rethinking Semantic Segmentation from a Sequence-To-Sequence Perspective with Transformers[J] arXiv preprint arXiv:2012.15840.

Zhong, Z., Li, Y., Ma, L., Li, J., Zheng, W.S., 2021. Spectral–spatial transformer network for hyperspectral image classification: a factorized architecture search framework. IEEE Trans. Geosci. Rem. Sens. 60, 1–15.

Zhou, D., Kang, B., Jin, X., et al., 2021. Deepvit: towards Deeper Vision transformer[J] arXiv preprint arXiv:2103.11886.