02
03
04
45%

Received: 4 October 2022 | Revised: 18 January 2023 | Accepted: 30 January 2023

DOI: 10.1049/cvi2.12181

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Loop and distillation: Attention weights fusion transformer for fine-grained representation

Sun Fayou[1] | Hea Choon Ngo[2] | Zuqiang Meng[1] | Yong Wee Sek[2]

[1]Guangxi University, Nanning, Guangxi, China

[2]Universiti Teknikal Malaysia Melaka, Durian Tunggal, Melaka, Malaysia

Correspondence
Sun Fayou and Zuqiang Meng
Email: 314565679@qq.com and 171313540@qq.com

Funding information
National Natural Science Foundation of China, Grant/Award Number: 62266004

Abstract
Learning subtle discriminative feature representation plays a significant role in Fine-Grained Visual Categorisation (FGVC). The vision transformer (ViT) achieves promising performance in the traditional image classification filed due to its multi-head self-attention mechanism. Unfortunately, ViT cannot effectively capture critical feature regions for FGVC due to only focusing on classification token and adopting the strategy of one-time image input. Besides, the advantage of attention weights fusion is not applied to ViT. To promote the performance of capturing vital regions for FGVC, the authors propose a novel model named RDTrans, which proposes discriminative region with top priority in a recurrent learning way. Specifically, proposed vital regions at each scale will be cropped and amplified as the next input parameters to finally locate the most discriminative region. Furthermore, a distillation learning method is employed to provide better supervision for elevating the generalisation ability. Concurrently, RDTrans can be easily trained end-to-end in a weakly-supervised learning way. Extensive experiments demonstrate that RDTrans yields state-of-the-art performance on four widely used fine-grained benchmarks, including CUB-200-2011, Stanford Cars, Stanford Dogs, and iNat2017.

KEYWORDS
computer vision, fine-grained image recognition, image processing

1 | INTRODUCTION

Fine-grained visual categorisation (FGVC) plays a vital role in the field of image recognition, which aims to distinguish subclasses within a general category, such as subcategories of birds [1, 2], dogs [3] etc. Meanwhile, it is a challenging field due to inappreciable inter-class variations. With the progress of the research methods [4–7], the performance of FGVC achieves a giant leap in recent years. Lots of research verified that distinguishing easily confused images rely on efficient location discrimination regions and feature learning [8, 9].

Inspired from the performance of discriminative regions in FGVC tasks, the accurate region proposing shows pivotal value [10, 11]. In the past few years, convolutional neural network (CNN) learned high-level semantic features from the shallower to the deeper, which reveals formidable advantages in computer vision [12, 13]. Fu et al. [8] proposed RA-CNN which utilises a multi-scale network to multi-step refine unique local region. Liu et al. [10] proposed FDL which adopts efficient supervision for discriminative part proposals and region-based feature learning. However, CNN builds long-distance dependencies among features in a limited way, which is difficult to effectively focus on the global receptive field. Concurrently, CNN cannot dynamically adapt to rich variety of input due to the fixed weights.

Recently, researchers creatively apply transformer to computer vision tasks [14]. Compared with CNN, the self-attention mechanism in transformer is not affected by local interactions, which can both achieve long-distance dependencies and perform parallel computing. Carion et al. [15] proposed DETR which combines a simple CNN with a transformer to generate the final detection set. Dosovitskiy et al. [16] proposed ViT which is a novel image classification model based on self-attention mechanism completely and the first research of transformer substituting for CNN. Zhang et al. [17] proposed AFTrans which utilises adaptive attention multi-scale fusion transformer for FGVC. Nonetheless, on the one hand the raw attention weights of transformer are used in a simple combination way, on the other hand the proposed transformer approaches rely on inputs at once to obtain discriminative regions and do not employ the advantages and tricks of CNN to improve the performance of transformer in FGVC.

To address the above challenges, we proposed a novel model RDTrans which can further boost the performance of ViT via lots of improvements according to the traits of FGVC. To be specific, we find that benefiting from knowledge transferring for distillation learning and recurrent discriminative region proposing, the performance of the critical region proposing obtains a steady progress [18–20]. Note that region proposing with a recurrent manner takes the input from the image region selected by attention weight fusion. Firstly, the attention weights in each transformer layer are grouped according to the index of attention heads and then fused by matrix product within the group, which outputs attention weight maps. Secondly, the attention weight fusion block consists of channel attention module and feature refinement module to further focus on significant information. Thirdly, we utilise channel grouping and maximum connected region search methods to determine the currently selected discriminative region. Fourthly, we adopt a recurrent manner to locate the discriminative region from coarse to fine. Moreover, we propose a distillation learning approach to transfer the object-based knowledge by CNN to the region proposing subnetwork. Intuitively, the object-based feature learning can bring about relatively reliable label distribution knowledge. To compensate for the shortcoming of ViT region proposing, we design object-based feature learning to supervise region proposing.

Since the finer-scale network utilises the recurrent manner, RDTrans can gradually focus on the most discriminative regions from coarse to fine. Concurrently, the distillation learning can reinforce the generalisation ability of the network. Thus, RDTrans benefits from the co-action of the recurrent learning and distillation learning. To future apply the superiority of ensemble learning, our model utilises fusing attention weights to enhance feature representation. The RDTrans outperforms existing vision transformer networks (e.g. ViT [16], Deit [21] etc.) on the ImageNet [22] as shown in Figure 1. Our contributions are summarised as follows:

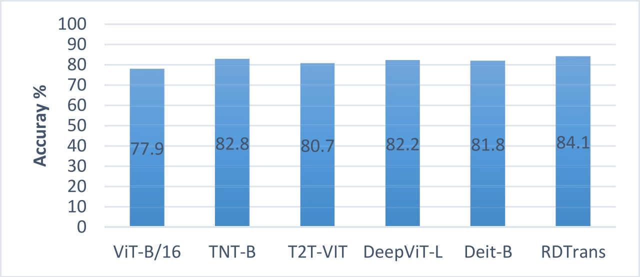

FIGURE 1 The comparison of top-1 accuracy of SOTA methods on ImageNet.

- To more accurately capture salient feature, we employ attention weights fusion method to increase the sensitivity to informative features and suppress less useful ones.
- To capture the most discriminative region, we adopt a recurrent manner to gradually focus on the most salient discriminative regions.
- To reinforce the generalisation ability of the network, we adopt a distillation learning approach to achieve the improvement of performance.
- To our best knowledge, RDTrans outperforms transformer-based network for FGVC tasks.

2 | RELATED WORK

In this section, we review previous methods that are most closely related to this study, including discriminative region proposal, distillation learning, and transformer combines with CNN.

2.1 | Discriminative region proposal

Discriminative region proposing methods have been given more attention due to the decisive role of obtaining discriminative features. Lots of approaches have been proposed by detecting the corresponding discriminative regions. CNN played an important role in the fgvc field and a lot of studies have made contributions in dealing with discriminative regional proposal [23, 24]. Zheng et al. [9] proposed MA-CNN which is one-scale network to recommend multiple discriminative regions by channel grouping layer. He et al. [11] proposed TASN which transfers the distilled learned fine-grained knowledge from hundreds of region proposing to a simple CNN. Recently, transformer has been applied to computer vision tasks and achieved satisfactory results [25–27]. ViT does not rely on CNN, which cannot meet FGVC tasks well due to each token which has equal role in ViT. He et al. [28] proposed TransFG for FGVC, which cannot utilise all attention weights in ViT, such that it only chooses some tokens with important contribution. Conde et al. [29] proposed a multi-stage ViT framework for FGVC, while it simply uses the attention regions generated by ViT, and continuously refines discriminant region through serial multiple ViT networks, which significantly improves the complexity of model. Zhang et al. [17] proposed an adaptive multi-scale transformer model AFtrans for FGVC, while it only corrects the discriminant region once, which has limited improvement on the performance of FGVC. However, we utilise a ViT network to refine the discriminant region in a recurrent way to reduce the complexity, and complement the transformer with the fused attention weight to further enhance the feature representation. Meanwhile, we apply the advantage of distillation learning to improve the generalisation ability of the model.

2.2 | Distillation learning

Knowledge distilling transfers knowledge from a high performance network into a smaller, distilled network in a teacher-student manner [30]. Liu et al. [10] proposed the FDL model, which transfers the knowledge from object to part regions as 'teacher' and 'student'. On the contrary, Zheng et al. [11] proposed TASN with knowledge distillation, which transfers fine-grained knowledge into object-based feature learning. As far as we know, both FDL and TASN achieve SOTA for FGVC at that time. Touvron et al. [21] proposed DeiT which adopts attention to distil transformer. Interestingly, DeiT demonstrate that using convolution network as teacher network for distillation is better than using transformer network as teacher. Whereas distillation learning can evidently strengthen the feature learning, we employ distillation learning to look for the devil in the details for FGVC.

2.3 | Transformer combines with CNN

The locality of CNN can enrich the feature diversity of transformer, which is a benefit for optimising the problem of over-smoothing of transformer features. The first method is to utilise the local characteristics of CNN in transformer to improve the network representation ability. Esser et al. [25] built VQGAN which combines CNN with transformer to yield high-resolution images. He et al. [28] proposed TransFG which just keeps vital tokens as the inputs of the final transformer layer. Swin transformer [31] explicitly interacts locally by limiting attention to local window. CeiT [32] adds local feature learning to FFN module to establish local relationship. The second method is to combine transformer and CNN to form a new network structure, for example, CeiT [32] and ViTc [33] add convolution blocks to the front of transformer to enhance the extraction of deep local information. Currently, the popular method is to fuse CNN features and transformer features. FFVT [34] aggregates the local information of multi-level tokens for classification. AFTrans [17] reinforces the location of discriminative regions with the help of channel attention mechanism. In this case, we carry on the magic of the fusion of CNN and transformer to further promote efficiency and effectiveness of RDTrans.

3 | METHOD

In this section, we will introduce the RDTrans network, which contains three modules (i.e. attention weight map, vital region proposing, and distillation learning). We utilise three modules to focus on the relative importance of the raw attention weights.

An overview of RDTrans is shown in Figure 2. Note that ViT is the backbone of the RDTrans which combines with the multiple attention weights fusion blocks for salient feature representation in a recurrent way and the distillation learning sub-network transfers knowledge to backbone to enhance discriminative region proposing.

From Figure 2, it can be observed that we group the transformer layers according to head for generating grouped head feature map (i.e. Head 1#, Head k# etc). Each grouped head feature map generates a feature map ($d \in R^{1\times H\times W}$, where H=W= the number of patches) by matrix product, and they are concatenated to output feature weight maps. Concurrently,

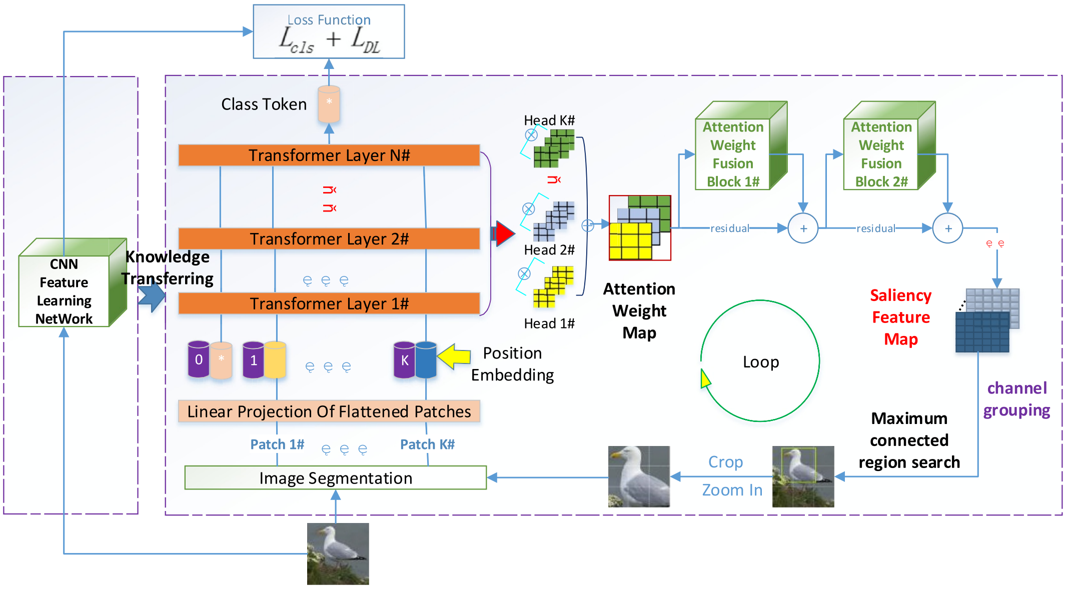

FIGURE 2 The architecture of RDTrans. Images are split into patches of the same size and sent into ViT. We group the transformer layers by head to generate an attention weight map. Subsequently, CNN is utilised to reinforce the feature representation and detect the most discriminative region at the moment over image. With a distillation learning method, RDTrans continuously refines the discriminative region in a recurrent way.

RDTrans consists of multiply attention weights fusion blocks, which finally outputs the current saliency feature map. Then, we utilise channel grouping method to cluster neighbouring locations for producing part attentions and we apply maximum connected region search method for proposing discriminative region. Finally, we crop and zoom in this region to the given size as the input of ViT.

3.1 | Attention weight map

To obtain more comprehensive and rich features, ViT utilises multi-head self-attention mechanism and each head focus on different region feature. ViT splits an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a Transformer encoder layer. The attention weight of each head is described as follows:

$$\mathrm{W} = soft\ max\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{1}$$

where Q, K, V are Query, Key, and Value vectors respectively. d_k is the dimension of Key and the dimension of W is $d_w \in R^{P\times P}$ (P is the number of patches).

In view of this, the attention weights of ?-th transformer encoder layer are shown below:

$$w_L = \left[w_L^1, w_L^2, \dots, w_L^k\right] \tag{2}$$

where k is the number of the self-attention heads. Thus, the attention weight map in RDTrans is as follows:

$$M_A(\text{Attention Weight Map}) = \left[\left(\prod_{h=1}^{h=n} W_1^h\right), \left(\prod_{h=1}^{h=n} W_2^h\right), \dots, \left(\prod_{h=1}^{h=n} W_k^h\right)\right] \tag{3}$$

where n is the number of transformer encoder layer, Π is matrix product, attention weight map with $M_A \in R^{k\times(P\times P)}$. The detailed process description is shown in Figure 3.

3.2 | Vital region proposing

Locality is a typical feature of CNN. In contrast, the learning process of transformer focuses on the interaction of global information. Thus, CNN can compensate for the deficiency of transformer (e.g. Ceit [32], Vitc [33], MobileViT [31] etc.). The detailed architecture of attention weight fusion block in Figure 2 is shown in Figure 4.

AWFB consists of channel attention (CA) and feature refinement module (FR). CA is composed of Max Pooling, Max Pooling, and a MLP with two layers sharing weights. The output of CA can be denoted as $M_{(c)}$:

$$M_{(c)} = M_A * sigmoid(MLP(AvgPool(M_A)) + MLP(MaxPool(M_A))) \tag{4}$$

where the dimension of $M_{(c)}$ is consistent with that of M_A.

FR is a residual network composed of two 3 * 3 Conv layer and one 1 * 1 Conv layer, which further improves the network representation ability. The output of AWFB is denoted Y.

$$\mathrm{Y} = f_{1\times1}\left(f_{3\times3}\left(f_{3\times3}\left(M_{(c)}\right)\right)\right) + M_A \tag{5}$$

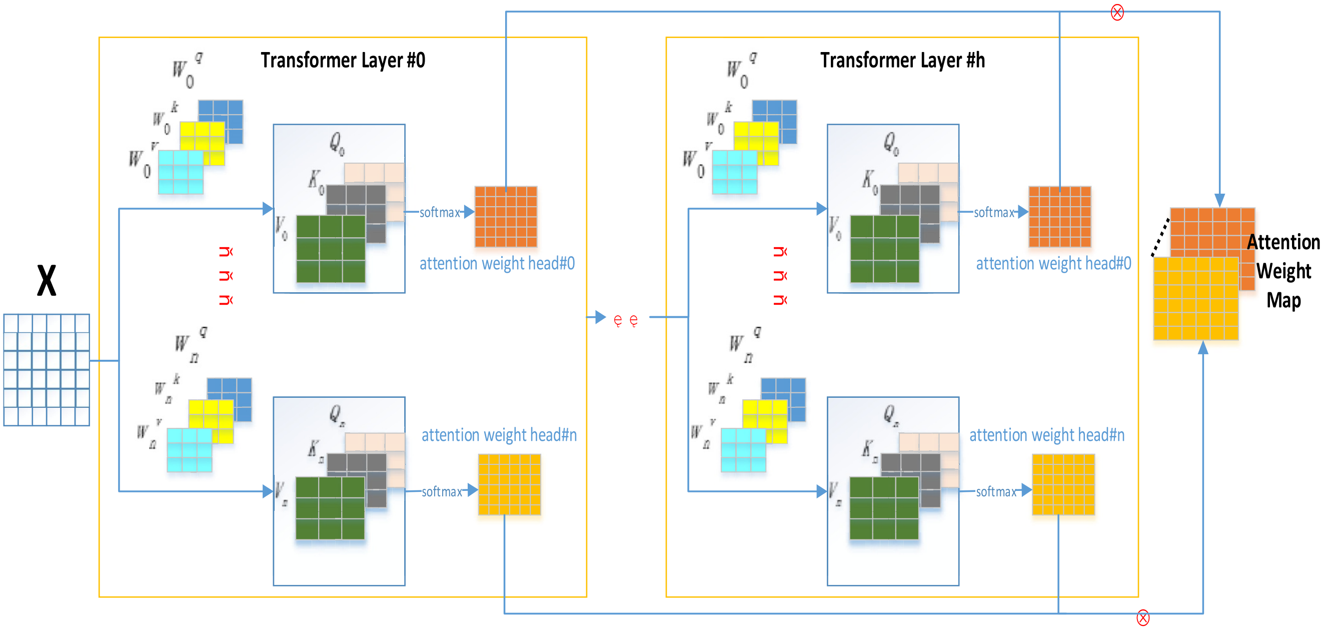

FIGURE 3 The detailed process of attention weigh map in RDTrans.

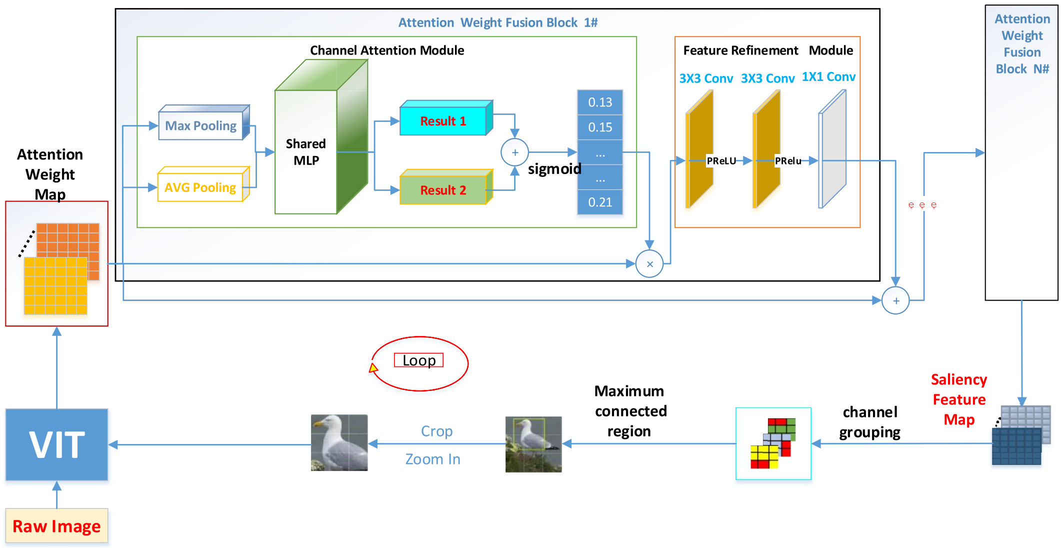

FIGURE 4 An overview of discriminative region proposing. An Attention Weight Fusion Block (AWFB) includes channel attention module and feature refinement module, which are stacked to produce saliency feature map. We adopt some tricks (i.e. channel grouping and maximum connected region methods) to assert the discriminative region. Note that the input of Attention Weight Fusion Block is M_A, and all of them share weights. We adopt matrix addition in the residual connection for Attention Weight Fusion Block.

where the dimension of Y is the same as M_A. We utilise multiple AWFB to obtain saliency feature map.

As far as we know, previous work [24] verified that convolutional channels in high-level layers tend to have responses to specific semantic patterns. Thus, we can divide convolutional channels into several groups by their semantic information. We adopt cosine similarity to achieve channel grouping.

$$M_\lambda = F_\lambda\left(\sum\left(cos\left(F_i, F_j\right)\right)\right) \tag{6}$$

where i and j are different channels, λ is the threshold of the number of grouping which is a hyperparameter with default values of 3. We regard each group as a connected region. Then, we employ the maximum connected region search algorithm to select the largest connected component from M_λ for locating, cropping, and zooming in this region over the raw image. In Figure 4, as ViT ignores lots of significant details for FGVC, we adopt a recurrent way to gradually focus on the most accurate discriminative region for achieving efficient object recognition.

In this work, we zoom in the cropped image region and its size is consistent with the raw image for training network effectively. At the same time, the number of loop is set to 3, and we give the reason in the corresponding experiments.

3.3 | Distillation learning

To improve the generalisation ability of RDTrans, we employ knowledge distillation method which transfers the learned details from CNN to ViT-based region proposing network in a teacher-student manner.

Specifically, the teacher-student outputs are denoted $Z_t(teacher), Z_s(student)$ respectively. Then, we convert Z_t, Z_s into a soft probability distribution over classes. Taking Z_s for example:

$$q_s^{(i)} = \frac{exp\left(\frac{Z_s^{(i)}}{T}\right)}{\sum_j exp\left(\frac{Z_s^{(j)}}{T}\right)} \tag{7}$$

where T is a temperature parameter to produce a soft probability distribution over classes. Hence, we get the soft target cross entropy for the distillation learning as:

$$L_{soft}(q_t, q_s) = -\sum_{i=1}^{N} q_t^{(i)} log^{q_s^{(i)}} \tag{8}$$

where N is the number of classes. Thus, the loss function of the RDTrans is as follows:

$$L_f = \alpha L_{soft} + \beta L_{VIT} \tag{9}$$

where L_{VIT} is the loss function for ViT. Note that α and β are hyperparameter with default values of 1, unless otherwise specified.

4 | EXPERIMENTS

In this section, we evaluate and analyse the performance of RDTrans on four widely used fine-grained benchmarks. The url of our codes is https://github.com/dlearing/RDTrans.git.

4.1 | Experiments setup

Datasets. To verify the performance of RDTrans, we carry out experiments on four datasets, including CUB-200-2011, Stanford Cars, Stanford Dogs, and iNat 2017. The detailed statistics for example, quantity, category numbers, and data splits are summarised in Table 1.

Implementation. In our experiments, the input images are resized 224*224 for fair comparison, which are split patches of size 16*16. Meanwhile, the step size of sliding window is set to be 12. Then we load weights from the official ViT-B_16 model pre-trained on ImageNet [22]. We utilise SGD optimiser with a momentum 0.8 and weight decay 0. The batch size is set to 16. We employ cosine annealing to adjust the learning rate which is initialized as 0.001 for four benchmarks. RDTrans is trained on four GPU (i.e. GeForce RTX 2070 8GB) with Pytorch as our code-base. Note that the backbone of the 'teacher' subnetwork is ResNet50 [37].

TABLE 1 Detailed statistics of the four datasets used in this paper.

Dataset	Total	Category	Train	Test
CUB-200-2011 [1]	11788	200	5994	5794
Stanford cars [35]	16185	196	8144	8041
Stanford dogs [3]	20580	120	12000	8580
iNat 2017 [36]	859000	5089	579184	95986

4.2 | Performance comparison

To prove the performance of RDTrans, we compared it with other SOAT models on above mentioned benchmarks. From Table 2, we found that RDTrans obtains SOAT performance on CUB-200-2011 [1], Stanford Cars [11], and Stanford Dogs [3].

Compared with the best result CLNET50 [39] so far, RDTrans achieves a further 0.9% improvement and reaches 93.3% on CUB-200-2011. Although ViT-based models obtain good performance on CUB, our RDTrans gets 1.7% performance gain compared to FFVT [34] and outperforms all CNN-based and ViT-based methods. CNN-based methods (e.g. RA-CNN [8] etc.) dependents on the relationship among local regional features to obtain discriminative regions, which are difficult to improve the performance. Later, methods with distillation learning (e.g. FDL [10], TASN [11] etc.) utilise 'teacher' sub-network to refine the discriminative region proposing several times. The appearance of ViT activates the potential of transformer in FGVC. Thus ViT-based methods achieve much better performance than the CNN-based methods. However, compared to the ViT-based methods, CLNET [39] employ long-distance feature dependency to take the lead in the FGVC again. In view of this, RDTrans combines the advantages of CNN and transformer to further improve performance for FGVC.

The 4th column of Table 2 shows the results on Stanford Cars. We observe that RDTrans first outperforms CNN-based methods with 0.4% improvement. We believe that this benchmark has less image noise than others, which leads to easily

TABLE 2 Comparison results on CUB-200-2011, Stanford Cars, Stanford Dogs.

		ACC. (%)		
Method	Backbone	CUB-200-2011	Stanford cars	Stanford dogs
RA-CNN [8]	VGG19	85.2	92.5	87.3
MA-CNN [9]	VGG19	86.5	92.8	-
TASN [11]	VGG19	86.1	92.4	-
FDL [10]	VGG19	86.84	91.52	84.9
NTS-Net [38]	Resnet-50	87.5	93.3	
DBTNet [24]	Resnet-50	87.5	94.1	
TASN [11]	Resnet-50	87.9	93.8	-
CLNET50 [39]	Resnet-50	92.4	96.7	-
FDL [10]	Resnet-50	-	-	85
DBTNet [24]	Resnet-101	88.1	94.5	-
FDL [10]	DenseNet161	89.09	94.02	84.46
StackedLSTM [40]	GoogleNet	90.4	-	-
ViT [16]	ViT-B_16	90.2	93.5	91.2
TransFG [28]	ViT-B_16	90.9	94.1	90.4
AFTrans [17]	ViT-B_16	91.5	95.0	91.6
FFVT [34]	ViT-B_16	91.6	-	91.5
RDTrans	ViT-B_16	93.3	97.1	93.6

obtain discriminative regions for classifying sub-categories. Even so, RDTrans achieves a 2.1% improvement compared to AFTrans [17] in terms of accuracy. We analysis that the reason is fusion CNN and locating the most discriminative region in a recurrent way.

Similarly, due to the subtle inter-class differences among certain species, Stanford Dogs [3] is a challenging benchmark. The 5th column of Table 2 shows that ViT-based methods outperforms CNN-based by a large margin. However, our RDTrans still outperform all of methods, which gets 2.0% gain compared to AFTrans and reaches 93.6% with its discriminative region proposing.

We show evaluation results on iNat2017 in Table 3. The iNat2017 [36] is a large-scale dataset with complicate background and high computational complexity. Luckily, ViT outperforms RA-CNN [8] by 3.9%, which demonstrates that ViT is born for large-scale benchmark. Based on ViT, RDTrans gets 1.8% gain compared to AFTrans. Concurrently, we know that the GFLOPs of our model outperform DeiT-B by 6.1%. Although the total of params achieve 0.2% compared to DeiT-B, its influence is limited. We believe that it is due to the adoption of attention weight fusion blocks.

4.3 | Ablation studies

We conduct ablation studies to show the effect of variants in RDTrans architecture on FGVC results. All ablation studies are done on CUB [1] while other datasets have the same phenomenon as well.

Impact of Attention Weight Fusion Block. RDTrans has multiple attention weight fusion blocks, which significantly enhance fine-grained feature representation.

From Table 4, it can be observed that attention weight fusion block has a significant improvement in performance for FGVC. We argue that these blocks focus on the relative importance of the raw attention weights. Specifically, if we add one block, the model can outperform ViT by 1.3%. Concurrently, the computational complexity is directly proportional to the number of blocks, but the improvement of accuracy is limited. Thus, we add three blocks in RDTrans.

From Table 5, both channel attention module and feature refinement module are beneficial to the improvement of performance. They achieve 0.7% and 0.4% gain compared to ViT respectively. We think that these two blocks further fuse attention weights, which is able to selectively emphasise informative features and suppress less useful ones.

Impact of Channel Grouping. In order to produce the discriminative region, we have to select the maximum connected component of the channel with the largest peak value. Besides, we have to cluster the similar channels and then uses method one to obtain the coordinates of the discriminative region.

TABLE 4 Ablation experiment on different number of attention weight fusion blocks.

Quantity	ACC. (%)
ViT	90.2
One block	91.5
Two blocks	92.6
Three blocks	93.3
Four blocks	93.6
Five blocks	93.8

TABLE 5 Ablation experiment on the framework of attention weight fusion block.

Methos	ACC. (%)
ViT	90.2
+Channel attention module	90.9
+Feature refinement module	90.6
+Channel attention Module + Feature refinement Module (one block)	91.5

TABLE 3 Comparison of SOTA methods on iNat 2017, ImageNet.

		iNat 2017	ImageNet	
Method	Backbone	ACC. (%)	#param ($\times 10^6$)	GFLOPs
ResNet152 [37]	ResNet-152	59	-	-
SSN [41]	ResNet-101	65.2	-	-
TASN [11]	ResNet-101	68.2	-	-
IncResNet [42]	IncResNet-101	67.3	-	-
ViT [16]	ViT-B_16	68	86	743
TransFG&PSM [28]	ViT-B_16	67.4	-	-
AFTrans [17]	ViT-B_16	68.7	-	-
DeiT-B [21]	ViT-B_16	-	86	17.6
DeepViT-L [43]	ViT-B_32	-	55	12.5
RDTrans	ViT-B_16	70.5	86.2	23.7

From Table 6, it conveys that channel grouping block can improve by 0.9% gain. We argue that similar feature regions are put together by channel grouping, which is beneficial to yield the discriminative region.

Impact of Distillation Learning. To show the advantages of the distillation learning, RDTrans boosts the performance significantly as shown in Table 6.

In Table 7, using ResNet50 as the backbone of distillation learning sub-network. The distillation learning brings 1.9% accuracy gains, which verifies that the knowledge learned in entire object by CNN is beneficial for the discriminative region proposing.

Impact of Recurrent Learning. To select the best number of loop, we conduct comparative experiments as shown in Figure 5.

From Figure 5, it can be observed that recurrent method can significantly improve performance. Similarly, if the number of loops is set 3, it can achieve 2.4% gains compared to ViT. Meanwhile, the performance improvement is limited while the value greater than 3, but the complexity of RDTrans will increase significantly. Hence, we configure the value of this parameter as 3. When we use RDTrans in different situations, if the increase in the accuracy of the model is less than 0.5, this is the best value of loop.

TABLE 6 Ablation experiment on channel grouping constraint.

Grouping	ACC. (%)
No Grouping (ViT + three blocks)	92.7
+Grouping (ViT + three blocks)	93.3

TABLE 7 Ablation experiment on different components.

Methods	ACC. (%)
Resnet50	83.5
ViT	90.2
RDTrans (No distillation learning)	91.4
RDTrans (+Distillation learning)	93.3

4.4 | Visualisation analysis

One image is randomly selected from each dataset. We conduct the visualisation experiment and the result is shown in Figure 6. To demonstrate the excellent performance of RDTrans, we conduct a comparative experiment.

Specifically, the ViT-based approaches can select multiple discriminative parts of the object. Furthermore, our RDTrans can capture the most discriminative regions and enhance the feature representation as shown in the 4th row. Furthermore, we use Attention Rollout to compute maps of the attention from the output token to the input space in the 5th row.

5 | CONCLUSION

In this work, we propose a novel network RDTrans and achieve SOAT performance on four benchmarks. We fuse attention weight grouped by head to reinforce the attention of different regions. Subsequently, we adopt three attention weight fusion blocks to obtain salient feature map. Afterwards, we utilise a channel grouping to produce part attentions from a group of channels whose peak responses appear in neighbouring locations. In addition, we adopt a distillation learning method to transfer the learned knowledge by CNN from object to regions proposing. Finally, we gradually refine the discriminative regions in a recurrent way. With the promising results achieved by RDTrans, we believe that the ViT and CNN are legend for FGVC. We need to seek the behind story of legend (e.g. combining multiple maximum connected components) in the future.

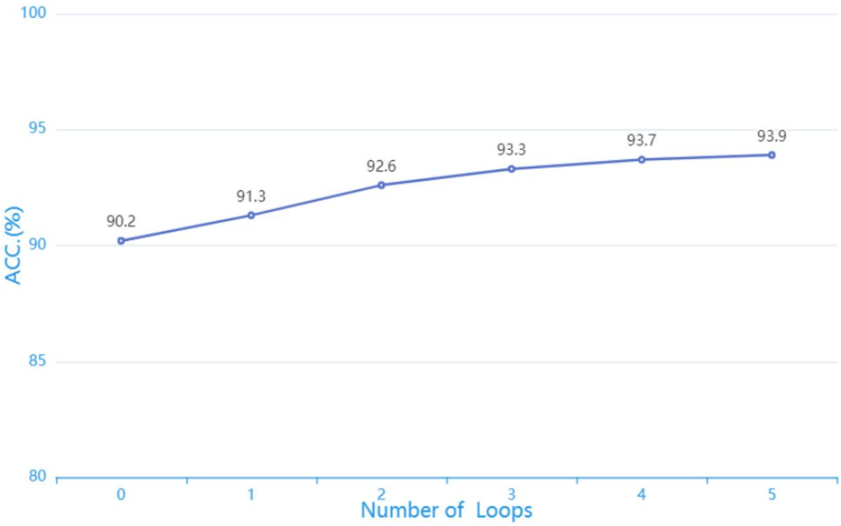

FIGURE 5 Influence of number of loops on accuracy.

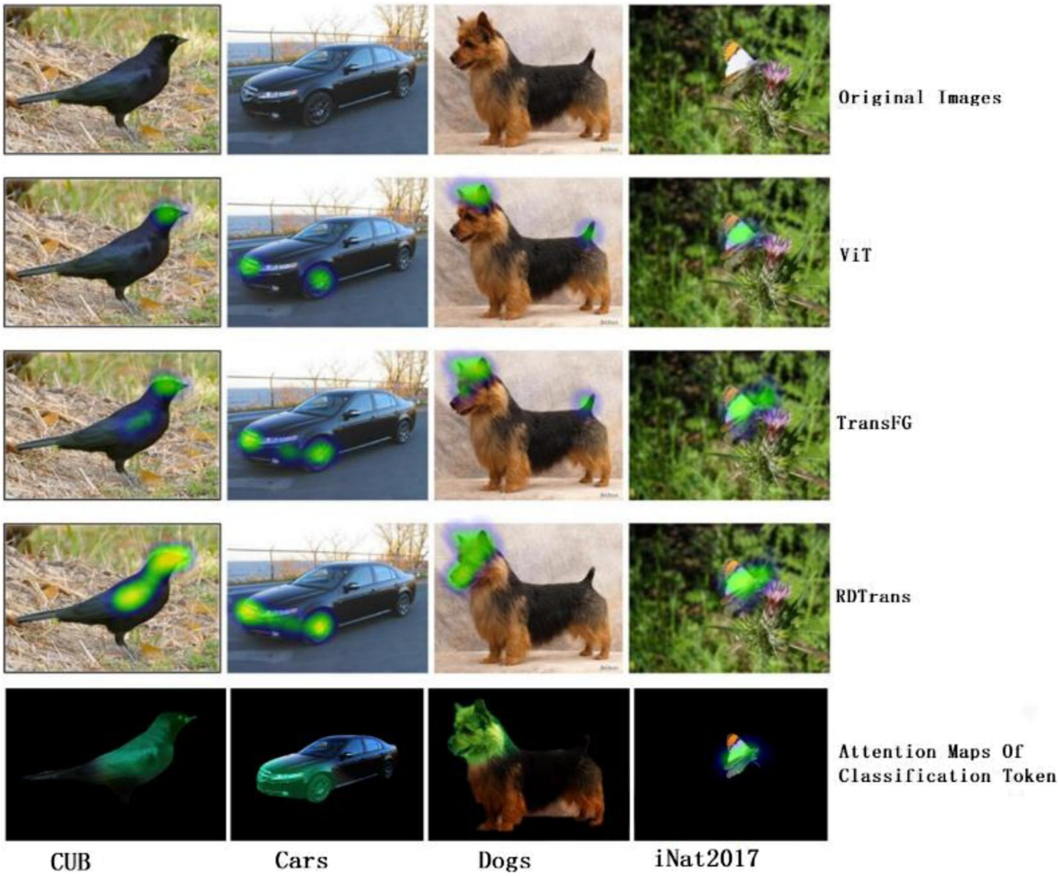

FIGURE 6 Visualisation of discriminative region proposing by the ViT-based methods on four benchmarks.

AUTHOR CONTRIBUTIONS

All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No. 62266004).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

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; https://tensorflow.google.cn/datasets/catalog/i_naturalist2017.

ORCID

Sun Fayou https://orcid.org/0000-0002-1590-9300

REFERENCES

1. Thomas, B., et al.: Birdsnap: large-scale fine-grained visual categorization of birds. Comput. Vis. Pattern Recogn., 2019–2026 (2014)
2. Welinder, P., et al.Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology. 1, 5 (2010)
3. Aditya, K., et al.: Novel dataset for fine-grained image categorization: Stanford dogs, Proc. In: CVPR Workshop on Fine-Grained Visual Categorization (FGVC), vol. 2 (2011)
4. Sun, F., Choon Ngo, H., Wee Sek, Y.: Combining multi-feature regions for fine-grained image recognition. Int. J. Image Graph. Signal Process. 1, 15–25 (2022). https://doi.org/10.5815/IJIGSP.2022.01.02
5. Sanghyun, W., et al.: CBAM: convolutional block attention module. In: European Conference on Computer Vision. vol. 11211, pp. 3–19 (2018)
6. Lin, T.Y., Roychowdhury, A., Maji, S.: Bilinear CNN Models for Fine-Grained Visual Recognition (2015)
7. Jie, H., Li, S., & Gang, S.: Squeeze-and-Excitation networks. Computer Vision and Pattern Recognition, abs/1709.01507 (2018)
8. Jianlong, F., Heliang, Z., Tao, M.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Computer Vision and Pattern Recognition, pp. 4476–4484 (2017)
9. Heliang, Z., et al.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: International Conference on Computer Vision. vol. 1, pp. 5219–5227 (2017)
10. Chuanbin, L., et al.: Filtration and distillation: enhancing region attention for fine-grained visual categorization. National Conference on Artificial Intelligence. vol. 34(07), pp. 11555–11562 (2020). https://doi.org/10.1609/aaai.v34i07.6822
11. Heliang, Z., et al.: Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-Grained Image Recognition, pp. 5012–5021 (2019). arXiv: Computer Vision and Pattern Recognition, abs/1903.06150
12. Yao, D., et al.: Selective sparse sampling for fine-grained image recognition. International Conference on Computer Vision 1, 6598–6607 (2019)
13. Ren, S., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Neural Information Processing Systems.

vol. 39(6), pp. 1137–1149 (2017). https://doi.org/10.1109/tpami.2016.2577031

14. Sachin, M., Mohammad, R.: MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. In: International Conference on Learning Representations (2022)
15. Carion, N., et al.: End-to-End object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020)
16. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Yuan, Z., et al.: A free lunch from ViT: adaptive attention multi-scale fusion Transformer for fine-grained visual recognition. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 3234–3238 (2022)
18. Karen, S., Andrew, Z.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
19. Krizhevsky, A., Sutskever, I., HintonGeoffrey, E.: ImageNet classification with deep convolutional neural networks. In: Neural Information Processing Systems 60(6), 84–90 (2017). https://doi.org/10.1145/3065386
20. Jinnian, Z., et al.: MiniViT: compressing vision transformers with weight multiplexing. In: Computer Vision and Pattern Recognition. vol. 1, pp. 12135–12144 (2022)
21. Hugo, T., et al.: Training data-efficient image transformers and distillation through attention. In: International Conference on Machine Learning. vol. 139, pp. 10347–10357 (2021)
22. Jia, D., et al.: ImageNet: a large-scale hierarchical image database.In: Computer Vision and Pattern Recognition, 248–255 (2009)
23. Chaojian, Y., et al.: Hierarchical bilinear pooling for fine-grained visual recognition. In: European Conference on Computer Vision. vol. 11220, pp. 595–610 (2018)
24. Zheng, H., et al.: Learning deep bilinear transformation for fine-grained image representation. In: Neural Information Processing Systems. vol. 32, pp. 4279–4288 (2019)
25. Patrick, E., Robin, R., Björn, O.: Taming transformers for high-resolution image synthesis. In: Computer Vision and Pattern Recognition, pp. 12873–12883 (2021)
26. Niki, P., et al.: Image transformer. In: International Conference on Machine Learning (2018). abs/1802.05751
27. Mark, C., et al.: Generative pretraining from pixels. In: International Conference on Machine Learning, pp. 1691–1703 (2020)
28. Ju, H., et al.: TransFG: a transformer architecture for fine-grained recognition. In: National Conference on Artificial Intelligence, pp. 852–860 (2022)
29. Marcos, V.C., Kerem, T.: Exploring Vision Transformers for Fine-grained Classification (2021). arXiv preprint arXiv, 2106.10587
30. Geoffrey, E.H., Oriol, V., Jeffrey, D.: Distilling the knowledge in a neural network. In: Computing Research Repository (2015). abs/1503.02531
31. Ze, L., et al.: Swin transformer - hierarchical vision transformer using shifted windows. In: International Conference on Computer Vision, pp. 9992–10002 (2021)
32. Kun, Y., et al.: Incorporating convolution designs into visual transformers. In: International Conference on Computer Vision, pp. 559–568 (2021)
33. Tete, X., et al.: Early convolutions help transformers see better. In: Neural Information Processing Systems, pp. 30392–30400 (2021)
34. Wang, J., et al.: Feature Fusion Vision Transformer Fine-Grained Visual Categorization (2021)
35. Tao, H., & Honggang, Q.: See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification. arXiv: Computer Vision and Pattern Recognition, abs/1901.09891 (2019)
36. Horn, G.V., et al.: The Inaturalist Species classification and Detection Dataset, vol. 5 (2018)
37. kaiming, h., et al.: Deep Residual Learning for Image Recognition, Computer Vision and Pattern Recognition, vol. 1, pp. 770–778 (2016). abs/1512.03385
38. Ze, Y., et al.: Learning to Navigate for Fine-Grained Classification. arXiv: Computer Vision and Pattern Recognition, vol. 11218, pp. 438–454 (2018)
39. Sun, f., Ngo, H.C., Sek, Y.W.: Adopting attention and cross-layer features for fine-grained representation. In: IEEE Access, vol. 10, pp. 82376-–82383 (2022). https://doi.org/10.1109/ACCESS.2022.3195907
40. Weifeng, G., Xiangru, L., Yizhou, Y.: Weakly Supervised Complementary Parts Models for Fine-Grained Image Classification from the Bottom up. arXiv: Computer Vision and Pattern Recognition, pp. 3034–3043 (2019). abs/1903.02827
41. Adrià, R., et al.: Learning to zoom: a saliency-based sampling layer for neural networks. In: European Conference on Computer Vision, vol. 11213, pp. 52–67 (2018)
42. Christian, S., Sergey, I., Vincent, V.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: National Conference on Artificial Intelligence, pp. 4278–4284 (2017). abs/1602.07261
43. Daquan, Z., et al.: DeepViT: Towards Deeper Vision Transformer (2021). arXiv preprint arXiv, 2103.11886

How to cite this article: Fayou, S., et al.: Loop and distillation: attention weights fusion transformer for fine-grained representation. IET Comput. Vis. 17(4), 473–482 (2023). https://doi.org/10.1049/cvi2.12181