Dual Transformer with Multi-Grained Assembly for Fine-Grained Visual Classification

Ruyi Ji, Jiaying Li, Libo Zhang*, Jing Liu, Yanjun Wu

***Abstract*—Fine-grained visual classification requires distinguishing sub-categories within the same super-category, which suffers from small inter-class and large intra-class variances. This paper aims to improve the FGVC task towards better performance, for which we deliver a novel dual Transformer framework (coined Dual-TR) with multi-grained assembly. The Dual-TR is well-designed to encode fine-grained objects by two parallel hierarchies, which is amenable to capturing the subtle yet discriminative cues via the self-attention mechanism in ViT. Specifically, we perform orthogonal multi-grained assembly within the Transformer structure for a more robust representation, *i.e.*, intra-layer and inter-layer assembly. The former aims to explore the informative feature in various self-attention heads within the Transformer layer. The latter pays attention to the token assembly across Transformer layers. Meanwhile, we introduce the constraint of center loss to pull intra-class samples' compactness and push that of inter-class samples. Extensive experiments show that Dual-TR performs on par with the state-of-the-art methods on four public benchmarks, including CUB-200-2011, NABirds, iNaturalist2017, and Stanford Dogs. The comprehensive ablation studies further demonstrate the effectiveness of architectural design choices.**

***Index Terms*—Transformer, multi-grained assembly, fine-grained visual classification**

I. Introduction

Fine-grained visual classification (FGVC) is tasked with distinguishing sub-categories within the same super-category, for example, different species of birds and dogs. As upstream foundational research, FGVC has facilitated a set of visual understanding tasks such as fine-grained action recognition [1], person reidentification [2], and human parsing [3]–[5], *etc*. With the advance of deep learning techniques [6]–[10], recent years have witnessed remarkable progress in the FGVC domain. However, it is still challenged by minor inter-class and significant intra-class variances due to deformation, occlusion, and illumination, see Fig. 1 (a).

In the early stage, considerable efforts have been made to design a desirable localization module under the supervision of object- or part-level annotations [11]–[13]. However, such a paradigm heavily depends on manual annotations of objects and parts, which are labor-intensive and time-consuming, even requiring expertise. Developments over this research line therefore gradually shift towards the attention mechanism to locate the distinct parts only with image-level labels, which dramatically reduces the requirement of annotation efforts. For example, Zheng *et al.* [14] leverage the attention mechanism to group different feature channels into various visual patterns, which are subsequently projected to category prediction. Zhang *et al.* [15] adopt multi-granularity sub-networks to learn global and local features for better performance jointly. For capturing the mutual features, Zhang *et al.* [16] develop a co-attention module to measure channel-wise feature similarities between paired samples within the same class. Despite their decent performance, the methods mentioned above often struggle to locate subtle yet discriminative parts and are far from accurate classification results in most cases.

Recently, Transformer architecture has gained momentum from natural language processing to computer vision. Notably, a series of variants derived from Vision Transformer (ViT) have touched or even outperformed those based on convolutional neural network (CNN) in a wide range of visual understanding tasks, such as image classification [17], object detection [18], semantic segmentation [19], and object tracking [20]. Moreover, as evidenced in Fig. 1 (b), a similar trend is emerging as such in the FGVC task. For instance, a more recent method [21] proposes to interact with features between multi-level patch representations to encode locally informative features, significantly improving the classification accuracy. Unfortunately, despite many attempts to reveal the potential of Transformer architecture, it still falls short in the feature representation of local regions. Contemporary works like [21], [22] select the informative tokens of Transformer layers as the input of the last layers, their success motivates us to review and rethink the usage of multiple head self-attention mechanism in vision transformer, going one step further towards more flexible and discriminative feature representation.

To this end, we propose a novel dual Transformer (dubbed Dual-TR) in this paper to fully unleash the potential of such an architecture for FGVC task. Overall, the Dual-TR inherits the powerful ability of feature representation from ViT, which characterizes correlations from the global perspective and satisfies the requirement to explore discriminative visual clues in the subtle regions. Under the guidance of self-attention mechanism in Transformer, the local hierarchy of the parallel architecture adaptively takes a semantic part view from the raw image as input. The essence of our method lies within the hierarchy of Transformer, where we perform

*Corresponding author: Libo Zhang

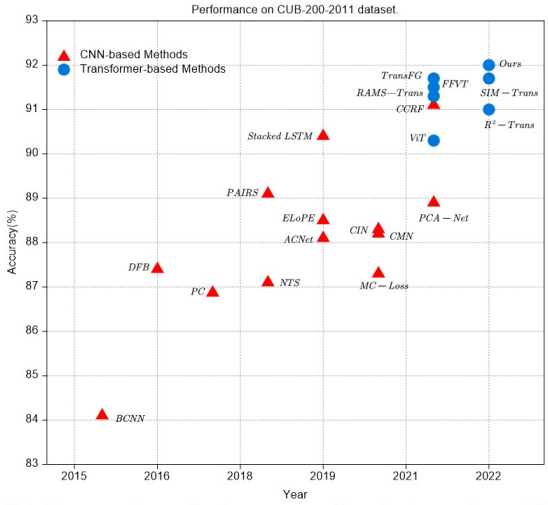
Ruyi Ji is with the State Key Laboratory of Computer Science, Institute of Software Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 101400, China (e-mail: ruyi2017@iscas.ac.cn).

Jiaying Li is with Beijing Information Science And Technology University, Beijing, China (2019020212@bistu.edu.cn)

Libo Zhang and Yanjun Wu are with the State Key Laboratory of Computer Science, Institute of Software Chinese Academy of Sciences, Beijing 100190, China (e-mail: libo@iscas.ac.cn).

Jing Liu is with School of Artificial Intelligence, University of Chinese Academy of Sciences and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. (email: jliu@nlpr.ia.ac.cn)

(a) Exemplars of fine-gained visual categorization



(b) Comparions between methods based on CNN architecture and those based on Transformer Design

Fig. 1. (a) FGVC remains challenging due to the following two factors: ❶ high intra-class variances: the birds belonging to the same category usually present significantly different appearances, such as illumination variations (the first column), clutter background (the second column), occlusion (the third column) and view-point changes (the fourth column); ❷ low inter-class variances: the birds in different columns belong to different categories, but share similar appearance in the same rows. (b) Comparison between CNN-based methods and those on top of Transformer framework on CUB-200-2011 dataset.

orthogonal multi-grained assembly, *i.e.*, intra-layer and inter-layer assembly. The former aims to explore the informative feature in various attention heads within the Transformer layer, while the latter emphasizes token assembly across the Transformer layers. Meanwhile, we exchange the class tokens of the dual hierarchies to ensure representation consistency. In this way, different hierarchies in the parallel architecture focus on fine-grained targets at different scales. And the subtle yet discriminative clues are explored via orthogonal multi-grained assembly within the Transformer hierarchy. Furthermore, we introduce the center loss to enhance the compactness of intra-class features within the same sub-category.

We conduct extensive experiments on four public benchmark datasets (including CUB-200-2011, NABirds, iNaturalist2017, and Stanford Dogs) to verify the effectiveness of the proposed method. Experimental results demonstrate that our method consistently performs favorably against the state-of-the-art approaches. In short, our contributions can be summarized in the following three folds:

- We propose a well-designed dual Transformer framework for the FGVC task. In our design, different hierarchies in the parallel architecture focus on the fine-grained target at different scales, and orthogonal multi-grained assembly within the Transformer hierarchy is performed to explore the subtle yet discriminative clues.
- To the best of our knowledge, we are the first to assemble the informative token features from the perspective of multi-head under the guidance of self-attention mechanism in the Transformer, which is proven to enjoy subtle yet discriminative information for the FGVC task.
- In quantitative and qualitative experiments, we demonstrate that the proposed Dual-TR achieves competitive performances compared with the state-of-the-art methods in four highly competitive benchmarks, including CUB-200-2011, NABirds, iNaturalist2017, and Stanford Dogs.

The rest of this paper is organized as follows. In section II, we briefly review some advanced techniques relevant to our proposed method. In section III, we describe the overall framework of Dual-TR in detail. Then in section IV, extensive experiments and comprehensive ablation studies are performed to validate the effectiveness of the proposed method. The conclusion is drawn in section VI.

II. Related Work

This section briefly reviews two research directions closely related to our method, *i.e.*, fine-grained visual classification and Transformer.

A. Fine-Grained Visual Classification

Due to the labor-intensive annotation process and the limited expertise, a weakly-supervised paradigm with image-level annotation has become a main workhorse for the FGVC task. Roughly speaking, the existing works can be divided into methods that locate the discriminative parts and those dedicated to learning high-order information.

The salient response of feature activation maps underpins the first research strand. These methods attempt to locate the informative parts by such visual cues. Typically, NTS-Net [23] utilizes a teacher network to supervise the informative part generation via a navigator network in a self-supervised style. Exchnet [24] focuses on the fine-grained hashing topic and generates compact binary codes for fine-grained images to alleviate the issues of slow query speed and highly redundant storage cost. Rather than the most informative part solely, Hanselmann *et al.* [25] adopt the top k discriminative parts to strengthen the modeling capacity of neural networks. Du *et al.* [26] iteratively optimize image patches at different

scales to capture multi-granularity visual clues. Inspired by Feature Pyramid Network (FPN), Ding *et al.* [27] introduce a bottom-up attention pathway and combine regions of interest (ROI) to locate informative regions. MMAL-Net [28] locates the salient object based on activation maps and excavates different parts by a sliding window mechanism. These models achieve stunning performance by picking discriminative areas. However, they disregard rich correlation information between discriminative parts and often suffer from the low-quality part location, limiting further improvement towards accurate results.

As for the second research line, bilinear pooling is a widely-used approach to encode second-order features [16], [29]–[31]. For example, prior art [29] is one representative of this research line, which leverages bilinear pooling to collect high-order statistical information, and reveals the power of bilinear pooling in combination with deep learning techniques for the FGVC task. Nevertheless, these methods fail to focus on subtle yet distinguishable regions explicitly. Moreover, it is quite hard to verify whether high-order representation can pay enough attention to discriminative clues within confused sub-categories.

It is well known that granularity scheme [1], [2] contributes a lot to various visual understanding tasks. Borrowing this inspiration, the Dual-TR casts salient responses of the attention mechanism as the semantic part view from the raw image and feeds it into the local hierarchy of parallel architecture to explicitly abstract multi-scale features for a more stable and robust representation.

B. Transformer

Since ViT [17] popularizes the Transformer architecture in the computer vision community, this up-and-coming star has shown astounding talent in representation learning and become dominant in a wide variety of visual understanding tasks, such as image classification [17], object detection [18], semantic segmentation [19], and image reconstruction [32]. Although the pure Transformer shows impressive performance, it suffers much from local modeling. To alleviate this issue, two paradigms are proposed in the literature. One intuitive way is to introduce CNN into architecture design to capture correlations in the local region. For example, Zhou *et al.* [33] integrate the self-attention mechanism with CNN for the sake of inductive bias characteristics of the convolution kernel. The alternative remedy is to modify the Transformer framework from the perspective of the patch token. Numerous works combine information interaction and global representation learning into a unified pipeline to characterize local features. Approach [34] exemplifies this research line, which develops a Tokens-to-Token module ahead of Transformer to enable each token to carry neighboring token information. More recently, a set of variants [22], [35], [36] built upon ViT successfully reveal its effectiveness for the FGVC task. Specifically, TransFG [22] captures discriminative patch tokens as well as rich relationships between each other. FFVT [35] interacts learned fine-grained features with multi-level patch representations under the guidance of class tokens. Hu *et al.* [36] utilize the attention mechanism to explore semantic parts as inputs for other branches, which guides the model to focus on informative areas at the cost of prohibitive computation burden and memory consumption. Although existing approaches have reported decent classification accuracy, the performance still needs improvement in the challenging FGVC task. In light of Transformer architecture, we explicitly design a dual Transformer framework to construct global and local encoding jointly. Unlike the methods above, the self-attention weight derived by Transformer motivates us to explore orthogonal assembly within the Transformer hierarchy and renders our Dual-TR framework the basis.

Notably, here we emphasize the difference between the proposed method and architectures [36], [37]. Even though sharing the spirit of dual Transformer structure, the proposed method considerably differs from RAMS-Trans [36] in the following aspects. Firstly, we only turn to the strength of attention weight in the first layer of the encoder to generate the semantic view for the local hierarchy, while RAMS-Trans depends heavily on the attention weights of all layers. Hence, compared to RAMS-Trans, our method enjoys an efficient forward computation process. Secondly, beyond the dual transformer structure, our method presents orthogonal multi-grained assembly within the Transformer hierarchy. By contrast, RAMS-Trans directly adopts the Transformer structure off-the-shelf to finalize the prediction without considering any fine-grained adaption for the FGVC task. Although decent performance is reported on the common classification datasets (for example, ImageNet), CrossViT [37] hardly generalizes to such a fine-grained domain. The proposed method differs from CrossViT in its inherent motivation. Firstly, we dynamically generate the input of the local hierarchy conditioned on the semantic part view. Instead, CrossViT projects a raw image into smaller patch sizes, resulting in the overlong batch sequence. Secondly, CrossViT focuses on multi-scale feature fusion strategies, while our method emphasizes exploring and discovering local information in ViT. These primary motivations determine the difference between our method and CrossViT significantly.

III. METHOD

Observing that the attention weights incidental to the Transformer structure reflect the correlation with the extent to which the patches contain the discriminative information, we argue that an effective way to emphasize the informative tokens and depress the irrelevant tokens can benefit a more discriminative representation for the FGVC task. To this end, we propose orthogonal multi-grained assembly within the Transformer hierarchy for effective token proposal. To be specific, we design intra-layer and inter-layer assembly within the Transformer hierarchy. The former aims to explore the informative features in various self-attention heads within the Transformer layer; the latter pays attention to token assembly across Transformer layers. We emphasize that the semantic view generation is computed on the fly, and our method only relies on the attention weights derived from the ViT structure. With these properties, our method enjoys the ability to provide

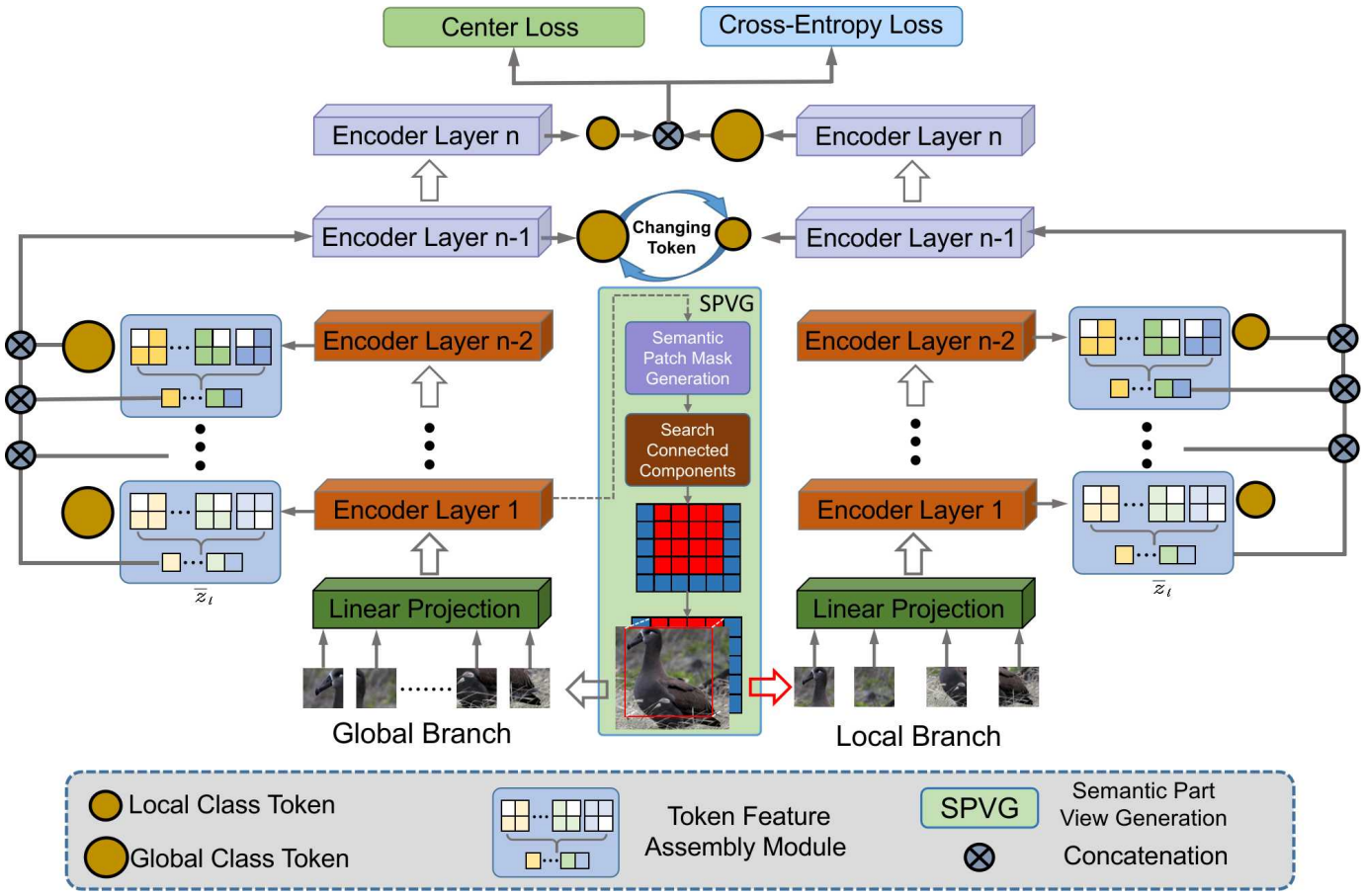


Fig. 2. The overview of our Dual-TR architecutre. The Dual-TR is well-designed to encode objects at different scales in two parallel hierarchies. Best visualization in color.

multi-grained informative features and efficient computation. Since the proposed method is built on top of the ViT, we briefly recall such a framework and then elaborately describe the Dual-TR design in this section.

A. Vision Transformer (ViT)

Overall, the architecture of ViT is similar to its counterpart in natural language processing. We briefly recall the tokenization, position embedding, and encoder in ViT, which are significantly involved in our research.

ViT first projects an image $\mathcal{I}$ with resolution $H \times W$ into a sequence of patches $P = \{p_1, p_2, \ldots, p_L\}$, where each patch token $p_i \in \mathbb{R}^{S^2 \times C}$ and $L = \frac{H}{S} \times \frac{W}{S}$, S refers to the spatial size of patch, C is the channel number. Notably, we can split an image into different number of small patches by changing the size of each patch. Then, a learnable linear function f_Θ with parameters Θ is adopted to generate embedded features $f_\Theta(P) \in \mathbb{R}^{(L+1)\times D}$, which includes the class token as well. Following the common practice, a learnable position embedding vector is injected into tokens to retain positional information, which is formally defined as follows:

$$Z_0 = f_\Theta(P) + E_P \tag{1}$$

where $E_P \in \mathbb{R}^{(L+1)\times D}$ provides the position information for patch tokens.

After that, a well-designed encoder upon linearly projected patch tokens is utilized to model long-range dependencies between patch tokens. Typically, the multi-head self-attention (MSA) mechanism and multi-layer perceptron (MLP) make up the encoder's core in ViT architecture, in which the computation can be formally given as follows:

$$\begin{aligned} Z_l^{'} &= \mathrm{MSA}(\mathrm{LN}(Z_{l-1})) + Z_{l-1} \\ Z_l &= \mathrm{MLP}(\mathrm{LN}(Z_l^{'})) + Z_l^{'} \end{aligned} \tag{2}$$

where Z_{l-1} and Z_l respectively represent the features before and after computation in the l-th layer, $\mathrm{LN}(\cdot)$ denotes the layer normalization operation, MSA interacts information between tokens, and MLP provides the non-linear transformation for each token.

B. Dual-TR Framework

To compensate for the drawback of ViT that suffers from modeling visual cues in subtle regions, we design a dual Transformer framework that jointly considers global long-range dependencies and local detail discrepancies. In this subsection, we describe semantic part view generation, then elaborate on the pipeline of orthogonal assembly within the Transformer hierarchy, followed by the introduction of center loss, and end with the loss functions adopted in our method. Fig. 2 illustrates the overall architecture of Dual-TR.

1) Semantic Part View Generation: For the input of local hierarchy in our parallel architecture, rather than randomly cropping content from raw image, we adaptively pick semantic image patches based on the multi-head self-attention mechanism to generate a semantic part view from the raw image. Concretely, the self-attention weight matrix series in the l-th level is defined as $\mathcal{A} = \{A^1, A^2, \ldots, A^N\}$, where N represents the number of self-attention heads. Therein, the self-attention weight matrix A^i is computed as follows:

$$A^i = \text{softmax}(\frac{QK^T}{D^{1/2}}) = \{a_0^i, a_1^i, \cdots, a_L^i\} \tag{3}$$

where L stands for the length of token sequence and Q, K represent query and key vectors respectively. $a_j^i \in \mathbb{R}^{L+1}$ denotes the similarities between the j-th token and the others in i-th head. Particularly, a_0^i measures the similarities between class token and other patch tokens. For the integrity and diversity of token information, we mean the values of a_0^i $(i = 1, 2, \ldots, N)$ across all self-attention heads, see Equation (4):

$$\mathcal{A}' = \frac{1}{N}\sum_{i=1}^{N} a_0^i \tag{4}$$

The rationale behind the above operation is that the class token contains the essential features responsible for final classification in the ViT design. Next, we reshape $\mathcal{A}'$ to $\mathcal{A}''$ with resolution $L^{1/2} \times L^{1/2}$ (here we only use the similarities between class token and patch tokens). Conditioning on $\mathcal{A}''$ and empirical hyper-parameter ε, we obtain semantic patch mask $\mathcal{M}$ as follows:

$$\mathcal{M}_{(i,j)} = \begin{cases} 1 & \text{if } \mathcal{A}''_{(i,j)} \geq \varepsilon * max(\mathcal{A}''), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

After that, we utilize the **Algorithm 1** to search the largest connected region $\hat{\mathcal{M}}$ from the $\mathcal{M}$. A semantic part view is required by $\odot$ (crop operation) on the original image.

$$\mathcal{V} = \mathcal{I} \odot \hat{\mathcal{M}} \tag{6}$$

For a better trade-off between efficient computation and memory consumption, the local hierarchy of Dual-TR is fed with a scaled semantic part view with resolution $\frac{H}{2} \times \frac{W}{2}$. We experimentally observe that as the Transformer layer goes deep, the patch representations tend to be over-smoothing, which is not friendly to the diversity of features. Therefore, we empirically rely on the first layer to generate the semantic part view in our experiments. The ablation study section will present more discussion about layer selection.

In the following, we describe how to perform the assembly from the intra-layer and inter-layer perspectives within the Transformer hierarchy.

2) Assembly of Intra-Layer in the Transformer Hierarchy: To explore the intra-layer discriminative cues, we design the Token Feature Assembly Module (TFAM) to capture discriminative token features from the self-attention heads. Taking the l^{th} layer of ViT architecture for example, the information carried by a token patch z_j is a D-dimension latent feature which is typically composed of N-head latent features, as defined in Equation (7).

$$z_j = \prod_{i=1}^{N} z_j^i \tag{7}$$

Algorithm 1 Search Connected Components via the semantic patch mask $\mathcal{M}$

Require: A patch mask $\mathcal{M}$;

```
Pick a patch p as the starting point;
while True do
    Leverage a flood-fill algorithm to label all the patches
    in the connected region that covers the patch p;
    if All the patches traverse then
        Break;
    end if
    Search for the next unprocessed patch as p;
end while
return Connectivity of the connected regions, and the
according region size
```

where $z_j^i \in \mathbb{R}^{\frac{D}{N}}$ denotes the j-th token feature in the i-th self-attention head, $j \in (1, 2, \cdots, L)$ and $\prod$ denotes the concatenation operation.

As claimed in work [38], the multi-head self-attention mechanism enables tokens to characterize discriminative information in various feature spaces, which inspires us to capture the discriminative discrepancies from different self-attention heads. The class token generally carries the most discriminative information. Based on this observation, we pick patch tokens that are closely relevant to the according class token in each self-attention head and assemble them along the channel dimension by the concatenation operation. In other words, picked features are the representatives corresponding to each self-attention head. The process can be formally written as follows:

$$\hat{z} = \prod_{i=1}^{N} z^i_{\underbrace{\arg\max}_{\{1,2,\cdots,L\}}(a_0^i)} \tag{8}$$

where a_0^i is defined in the Equation (3), referring to the similarities between the class token and other tokens in the i-th self-attention head, $z^i_{\lceil\rfloor}$ is the picked feature from the token with index $\underbrace{\arg\max}_{\{1,2,\cdots,L\}}(a_0^i)$ in the i-th self-attention head, and $\prod$ denotes the concatenation operation.

3) Assembly of Inter-Layer in the Transformer Hierarchy: After the investigation within the Transformer layer, we turn to the inter-layer assembly across the Transformer layers. In specific, we assemble the picked token feature $\hat{z}_l$ $(l = 1, 2, \ldots, n-2)$ with $(n-2)$-th class token and feed them as the input for the penultimate layer of encode,

$$Z_{n-1} = \{x_{n-2}, \hat{z}_1, \cdots, \hat{z}_{n-2}\} \tag{9}$$

where $\hat{z}_1, \cdots, \hat{z}_{n-2}$ represent picked token features from the first $n-2$ layers. x_{n-2} refers to the class token of the $(n-2)$-th layer. Experimentally, we observe that such assembly between the class token and picked patch tokens can significantly benefit the classification performance.

Even though our dual Transformer structure allows the network to focus on the semantic parts within the Transformer hierarchy, the information interaction between the parallel hierarchies needs to be better investigated, which is essential

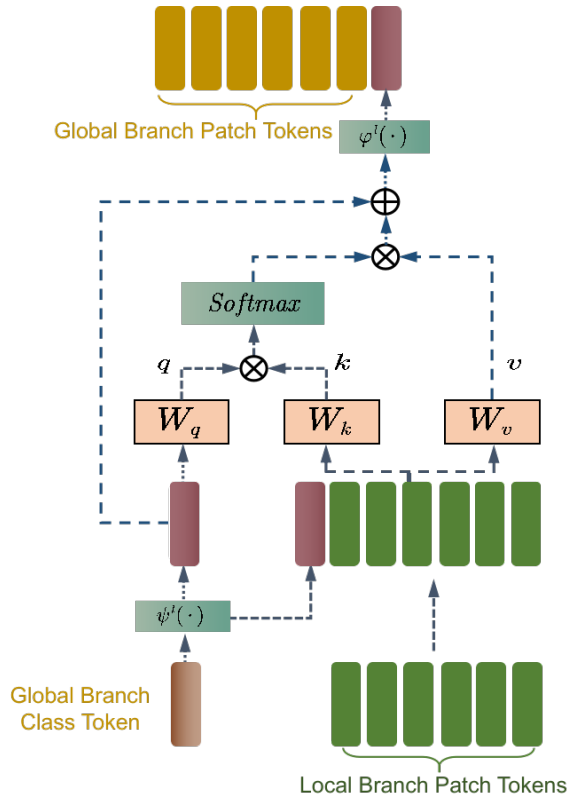


Fig. 3. Alignment design for global branch. The class token of the global branch serves as a query token to interact with the patch tokens from the local branch through attention mechanism. $\psi^l(\cdot)$ and $\varphi^l(\cdot)$ are the projection functions to align feature dimensions. The local branch follows the same procedure but exchange class token and patch tokens from another branch.

for optimizing the parallel hierarchies towards accurate results consistently. Inspired by the success of methods [6], [24], [37], we exchange the class tokens between the parallel branches in the penultimate layer, which explicitly ensures semantic consistency across the parallel hierarchies. Fig. 3 demonstrates the overall flow-chart of alignment between class token and patch tokens from the other branch. Specifically, the class token of the global branch acts as a query token to interact with the patch tokens from the local branch through the attention mechanism. Wherein, $\psi^l(\cdot)$ and $\varphi^l(\cdot)$ are projections to align dimensions. Notably, the local branch follows the same procedure but exchange class token and patch tokens from another branch. As illustrated in Fig. 3, we concatenate the class token with patch tokens as follows,

$$Z_{n-1}^{l'} = [\psi^l(x_{n-2}^l), \hat{z}_1^s, \hat{z}_2^s, \cdots, \hat{z}_{n-2}^s] \tag{10}$$

where $\psi(\cdot)$ denotes the linear projection function for dimension alignment. We perform cross-attention between class token and patch tokens. The class token is the only query since the patch tokens' information is merged into the class token. The above procedure can be mathematically formulated as follows

$$\begin{gathered} \boldsymbol{q} = \psi^l(x_{n-2}^l)\boldsymbol{W_q}, \quad \boldsymbol{k} = Z_{n-1}^{l'}\boldsymbol{W_k}, \quad \boldsymbol{v} = Z_{n-1}^{l'}\boldsymbol{W_v} \\ \omega = \text{softmax}(\boldsymbol{q}\boldsymbol{k}^T/\sqrt{C/h}) \\ M_{head}(Z_{n-2}^{l'}) = \omega\boldsymbol{v} \end{gathered} \tag{11}$$

where $\boldsymbol{W_q}, \boldsymbol{W_k}, \boldsymbol{W_v} \in \mathbb{R}^{C\times(C/h)}$ are learnable parameters, C and h denote the feature dimension and number of multiple heads. Note that we only use the class token in the query. Moreover, like self-attention, we also use multiple heads mechanism. Specifically, the output $Z_{n-1}^{l''}$ of an alignment module based on the given class token with layer normalization (LN) and skip connection is defined as follows.

$$\begin{gathered} x_{n-2}^{l'} = \psi^l(x_{n-2}^l) + M_{head}(\text{LN}([\psi^l(x_{n-2}^l), \hat{z}_1^s, \hat{z}_2^s, \cdots, \hat{z}_{n-2}^s])) \\ Z_{n-1}^{l''} = [\varphi^l(x_{n-2}^{l'}), \hat{z}_1^l, \hat{z}_2^l, \cdots, \hat{z}_{n-2}^l] \end{gathered} \tag{12}$$

where $\psi(\cdot)$ and $\varphi(\cdot)$ are used to squeeze feature and project feature back for dimension alignment.

After the computation of the last layer, we concatenate the class tokens of parallel hierarchies to finalize the category prediction.

4) Loss Function: To further alleviate the issue of large intra-class and small inter-class variances, we introduce the center loss to guide the optimization procedure. Concretely, we first define class centers as $\mathcal{C} \in \mathbb{R}^{2D\times S}$ ($2D$ equals the dimension of the concatenation of two class tokens, S is the number of sub-categories) and initialize it following a uniform distribution. And then, we dynamically renew each sub-category center by the update rule below, which effectively characterizes the intra-class variations.

$$\boldsymbol{c}_j' \leftarrow \boldsymbol{c}_j + \tau\Delta\boldsymbol{c}_j \tag{13}$$

where the sub-category center $\boldsymbol{c}_j \in \mathbb{R}^{2D}$. $\boldsymbol{c}_j$ and $\boldsymbol{c}_j'$ denote the class centers of the deep feature before and after iterations, respectively. τ denotes the update rate for centers to avoid large perturbations. To achieve this goal, we borrow a weighted update mechanism from work [39] to adjust class centers. To be specific, the update equation of $\Delta\boldsymbol{c}_j$ is mathematically expressed by

$$\Delta\boldsymbol{c}_j = \frac{\sum_{i=1}^{\Omega}\delta(y_i = j)(\boldsymbol{c}_j - [x_c, x_c']_i)}{1 + \sum_{i=1}^{\Omega}\delta(y_i = j)} \tag{14}$$

where x_c and x_c' are the class tokens from the parallel hierarchies belonging to the y_i-th category. $\delta(condition) = 1$ if the condition is satisfied, and $\delta(condition) = 0$ if not. Ω denotes the total number of training samples in each mini-batch, and the weight parameter β_i is designed as the maximum predicted probability of sample x_i:

$$\beta_i = \max \mathcal{P}(x_i) \tag{15}$$

After that, we apply the ℓ_2 loss to explicitly ensure the consistency between $[x_c, x_c']_i$ and $\boldsymbol{c}_{y_i}'$, which is mathematically formulated as follows:

$$\mathcal{L}_{ctr} = \frac{1}{2}\sum_{i=1}^{\Omega} \parallel [x_c, x_c']_i - \boldsymbol{c}_{y_i}' \parallel_2^2 \tag{16}$$

Based on all the above considerations together, at the training stage, the total loss of the Dual-TR is expressed as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{ctr} \tag{17}$$

where $\mathcal{L}_{cls}$ means the cross-entropy loss function and α is the hyper-parameters to balance the above two loss items, which is empirically set 0.1 in our experiments.

IV. Experiments

In this section, we first describe the experiment settings in Subsection IV-A. Then, the introduction of the involved datasets is illustrated in Subsection IV-B. Finally, quantitative and qualitative experimental results are reported in Subsection IV-C.

A. Implementation Details

We perform experiments on four public benchmarks, including CUB-200-2011 [40], NABirds [41], Stanford Dogs [42], and iNaturalist2017 [43] to validate the effectiveness of the Dual-TR. Following the common protocol in the FGVC task, we resize the input image to 448×448 and adopt the official ViT-B_16 model as the backbone, which is pre-trained on ILSVRC CLSLOC [44] dataset. In all experiments, the Stochastic Gradient Descent (SGD) algorithm, with a momentum of 0.9, a total of $10,000$ steps, batch size of 16, is utilized to optimize the Dual-TR in an end-to-end manner. The initial learning rate is 0.03 for CUB-200-2011, Stanford Dogs, and NABirds datasets, and 0.01 for the iNaturalist2017 dataset, and scheduled by cosine annealing strategy. All experiments are conducted on four Tesla V100 GPUs and implemented with the PyTorch [45] deep learning framework.

TABLE I
STATISTICS OF BENCHMARK DATASETS FOR FINE-GRAINED VISUAL CLASSIFICATION.

Dataset	Object	Category	Training	Testing
iNaturalist2017 [43]	Nature	5,089	579,184	95,986
NABirds [41]	Bird	555	23,929	24,633
CUB-200-2011 [40]	Bird	200	5994	5794
Stanford Dogs [42]	Dog	196	8144	8041

B. Fine-grained visual classification Datasets

We report experiments on four widely used benchmark datasets, *i.e.*, iNaturalist2017, NABirds, CUB-200-2011, and Stanford Dogs, and rank all methods based on the top-1 (Acc.@1) evaluation metric. Statistics of datasets and their *train/test* splits are summarized in Table I. It is worth noticing that we rely on the image-level annotation without any extra information (*e.g.*, part annotations, object bounding boxes, and web prior knowledge of categories). Thus, we do not compare with the methods which rely on these annotations.

1) iNaturalist2017: As seen in Fig. 4 (a), iNaturalist2017 dataset contains cross-species images, and the biased distribution is evident between $5,089$ categories. Empirically, the image numbers in the *train* and *test* set are $579,184$ and $95,986$, respectively.

(a) The iNaturalist2017 dataset.

(b) The CUB-200-2011 dataset.

(c) The NABirds dataset.

(d) The Stanford Dogs dataset.

Fig. 4. Examples of the involved benchmarks.

2) CUB-200-2011: The CUB-200-2011 dataset is a well-known bird species dataset, which is competitive and commonly used for fine-grained image classification, see examples in Fig. 4 (b) for illustration. This dataset collects $11,788$ images of 200 different bird subcategories, which consists of $5,994$ images for training and $5,794$ images for testing. Each subcategory has roughly 30 *train* and *test* images. It has three-level annotations, including image-level subcategory labels, object bounding boxes, and part landmarks.

3) NABirds: The NABirds dataset is another widely used fine-grained classification dataset, which covers more categories and has a larger volume than CUB-200-2011 dataset, with $23,929$ *train* and $24,633$ *test* images for 555 categories. Fig. 4 (c) exhibits the samples from the NABirds dataset. In particular, this dataset covers 400 species and occasionally arranges classes for male and female birds.

TABLE II
THE QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON CUB-200-2011.

Method	Backbone	Acc.@1 (%)
RA-CNN [48]	VGG-19	85.3
MA-CNN [14]	VGG-19	86.5
MaxEnt [49]	DenseNet-161	86.6
DVAN [50]	VGG-16	87.1
CIN [51]	ResNet-50	87.5
API-Net [31]	ResNet-50	87.7
SnapMix [52]	ResNet-50	87.7
ACNet [53]	ResNet-50	88.1
AP-CNN [54]	ResNet-50	88.4
FDL [55]	ResNet-50	88.6
PCA-Net [56]	ResNet-101	88.9
TBMSL [28]	ResNet-50	89.6
Stacked LSTM [46]	GoogleNet	90.4
ViT [17]	ViT_B_16	90.3
EV [57]	ViT_B_16	91.0
RAMS [36]	ViT_B_16	91.3
TPSKG [58]	ViT_B_16	91.3
AFTrans [59]	ViT_B_16	91.5
FFVT [35]	ViT_B_16	91.6
R^2-Trans [60]	ViT_B_16	91.5
TransFG [22]	ViT_B_16	91.7
SIM-Trans [47]	ViT_B_16	91.8
Ours	ViT_B_16	**92.0**

4) tanford Dogs: There are $20,000$ images in the Stanford Dogs datasets for 120 categories, of which $8,580$ are in the *test* set and $12,000$ in the *train* set, coupled with image-level and object-level annotations, see the samples in Fig. 4 (d).

C. Comparison With the State-of-the-Arts

1) CUB-200-2011: Table II shows that Dual-TR achieves superior performance over the state-of-the-art methods. Among those CNN-based methods, Stacked LSTM [46] achieves the best performance. However, it relies heavily on object detection and instance segmentation to capture discriminative information, whose complexity makes it incapable of further improvements. It can be seen from Table II that when ViT has applied along, the classification accuracy reaches up to 90.3% of performance. TransFG [22] further unleashes the advantage of the ViT structure for the FGVC task, which selects patch tokens via attention weights matrix and ranks the third place. Different from TransFG, we explicitly adopt the multi-grained assembly within the Transformer hierarchy to guide the network to focus on discrepancies in the local region and surpass it by 0.3% top-1 accuracy. SIM-Trans [47] is proposed recently for FGVC, based on the Transformer structure as well. By comparison, the Dual-TR exceeds SIM-Trans by an absolute gain of 0.2% of performance, validating the advantage of the proposed method.

2) NABirds: As listed in Table III, there is a similar trend to the CUB-200-2011 dataset, where the ViT-based methods beat those based on CNN architecture significantly. We can observe that the proposed Dual-TR outperforms the CNN-based best-performed method MGE-CNN [15] by absolute 2.7% top-1 accuracy and achieves a performance gain of absolute 1.4% compared to the ViT. Differing from TransFG [22], we significantly improve 0.5% as well. These results validate the effectiveness of orthogonal assembly within the Transformer hierarchy and the parallel structure design choice to a certain extent.

TABLE III
THE QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON NABIRDS.

Method	Backbone	Acc.@1 (%)
MaxEnt [49]	DenseNet-161	83.0
API-Net [31]	DenseNet-161	88.1
CS-Parts [61]	ResNet-50	88.5
MGE-CNN [15]	ResNet-101	88.6
ViT [17]	ViT_B_16	89.9
TPSKG [58]	ViT_B_16	90.1
R^2-Trans [60]	ViT_B_16	90.2
TransFG [22]	ViT_B_16	90.8
Ours	ViT_B_16	**91.3**

TABLE IV
THE QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON INATURALIST2017.

Method	Backbone	Acc.@1 (%)
R50 [62]	ResNet-152	59.0
SSN [63]	ResNet-101	65.2
IARG [64]	ResNet-101	66.8
Inception-v4 [65]	ResNet-101	67.3
TASN [66]	ResNet-101	68.2
RAMS [36]	ViT_B_16	68.5
ViT [17]	ViT_B_16	68.7
AFTrans [59]	ViT_B_16	68.9
SIM-Trans [47]	ViT_B_16	69.9
TransFG [22]	ViT_B_16	**71.7**
Ours	ViT_B_16	71.5

3) iNaturalist2017: The comparison between the Dual-TR and the state-of-the-art methods on the iNaturalist2017 dataset is summarized in Table IV. One can see that ViT surpasses ResNet-50 [62] by absolute 9.7% improvement on Acc.@1, significantly revealing the advantage of the Transformer structure. Further, the Dual-TR gains an additional improvement of 2.8% over ViT and is on par with the best-performed method. Notably, compared with RAMS [36], which also adopts the double-branch structure to capture local differences, we achieve a performance improvement of 3.0% consistently.

4) Stanford Dogs: The classification performance on the Stanford Dogs dataset is reported in Table V. We can see that Dual-TR rivals the best-performed method EV [57], *i.e.*, 93.2% *vs.* 93.2%. Even though TPSKG [58] builds an additional knowledge set to store category information and learn comprehensive representations, it still is inferior to ours

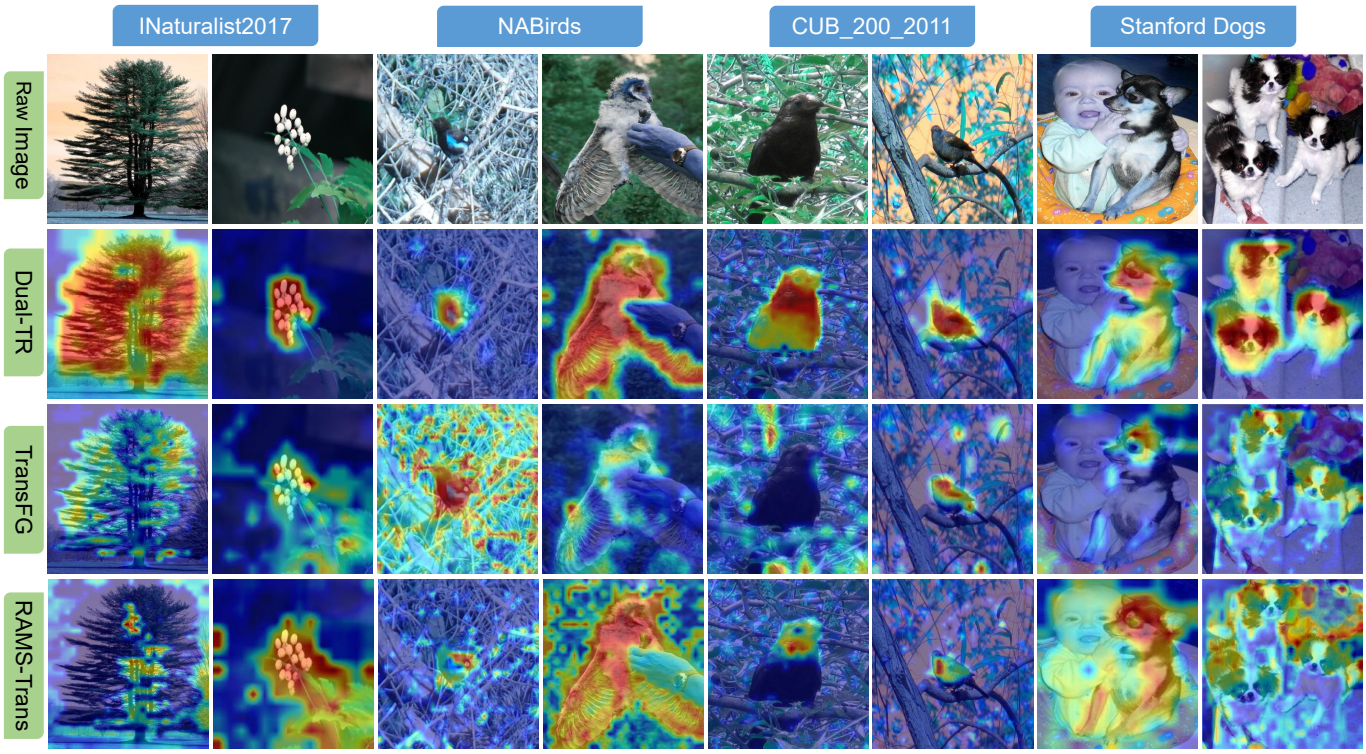


Fig. 5. Visualization of the results in our Dual-TR method. Each row presents response heatmaps of samples in iNaturalist2017, NABirds, CUB-200-2011, Stanford Dogs.

TABLE V
THE QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON STANFORD DOGS.

Method	Backbone	Acc.@1 (%)
FDL [55]	DenseNet-161	84.9
RA-CNN [48]	VGG-19	87.3
DB [67]	ResNet-50	87.7
SEF [68]	ResNet-50	88.8
API-Net [31]	ResNet-101	90.3
AFTrans [59]	ViT_B_16	91.6
ViT [17]	ViT_B_16	91.7
TransFG [22]	ViT_B_16	92.3
RAMS [36]	ViT_B_16	92.4
TPSKG [58]	ViT_B_16	92.5
R^2-Trans [60]	ViT_B_16	92.8
EV [57]	ViT_B_16	**93.2**
Ours	ViT_B_16	**93.2**

(-0.7%). We attribute these consistent improvements to the multi-grained assembly of token features, which allows the model to focus on discriminative regions.

5) Qualitative Visualizations: To validate the effectiveness intuitively, we compare the attention maps between Dual-TR and existing notable methods on four public datasets without cherry-picking. As depicted in Fig. 5, we can see that Dual-TR shows obvious advantages over existing methods and performs well in locating the salient object and concentrates on the subtle yet distinctive parts such as the wings and tail of a bird (for example, parts in the 5-th and 6-th columns). We believe that these visualizations present further insights into decisive factors for accurate performance and verify the efficacy of our architecture design choice to a certain extent.

V. ABLATION STUDIES

To explore our design choices of the Dual-TR framework, we conduct in-depth ablation studies to analyze how critical components or hyper-parameters in our method affect performance. We perform all ablation experiments on the CUB-200-2011 dataset, and experiment setups are the same as the description in Subsection IV-A.

A. The dual Transformer architecture design

We construct a series of variants to prove the effectiveness of dual architecture in our design. As shown in Table VI, "#1 ViT" denotes that we directly transfer the ViT framework to the FGVC domain. "#2 Dual-ViT" implies a dual ViT structure. "#3 TR" demonstrates a single hierarchy variant of the proposed method. "#4 Dual-TR" means the full version of our method. From Table VI, it is observed that Dual-TR achieves the best performance among all variants. The comparisons "#1 ViT" *vs.* "#2 Dual-ViT" and "#3 TR" *vs.* "#4 Dual-TR" consistently demonstrate that dual structure is beneficial to representation learning, which is in line with the actual situation where dual hierarchies can explicitly capture informative visual cues at different scales, which are the

complement of each other. Notably, the performance comparison between "#3 TR" and "#2 Dual-ViT", *i.e.*, 91.7% *vs.* 90.7%, demonstrates that the performance improvement of our method comes mainly from the design within the Transformer hierarchy.

TABLE VI
ABLATION STUDY ON DUAL ARCHITECTURE DESIGN CHOICES.

#	Variants	Acc.@1 (%)
1	ViT	90.3
2	Dual-ViT	90.7
3	TR	91.7
4	Dual-TR	92.0

TABLE VII
ABLATION STUDY ON DUAL TRANSFORMER ARCHITECTURE AND THE SEMANTIC PART VIEW GENERATION.

#	Variants	Acc.@1 (%)
1	Baseline (ViT)	90.3
2	+Assembly within Transformer Layer	91.0
3	+Assembly across Transformer Layer	91.5
4	Dual-TR+Center Crop	91.6
5	Dual-TR+Random Crop	91.7
6	Dual-TR+SPVG	92.0

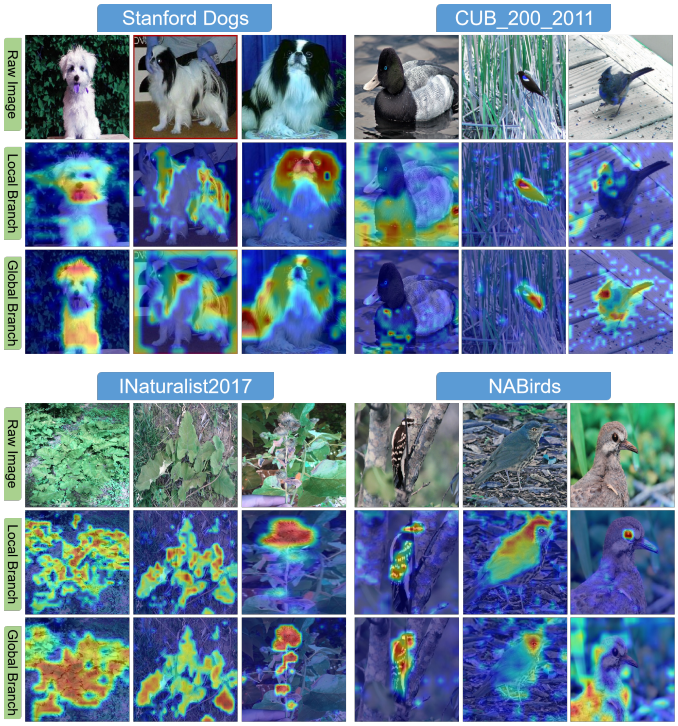


Fig. 6. An illustration of learning discriminative details by Dual-TR. The first row is the raw images, the second row shows the heatmaps yielded by local hierarchy, and the third row indicates the heatmaps generated by global hierarchy.

B. The Token Feature Assembly Module

Moreover, each token feature distribution may differ in various self-attention heads. For instance, in the i-th self-attention head, a patch token is closely relevant to the class token but may not be in the j-th self-attention head. Thus, in each self-attention head, we separately search for a patch token the most relevant to the class token. We ablate each assembly in TFAM and show their contributions in Table VII. We show that the assembly within the Transformer layer dramatically increases the accuracy from 90.3% to 91.0%, which convincingly proves that such a schema is a promising direction for this visual domain. Moreover, the assembly across the Transformer layer further boosts performance from 91.0% to 91.5%. These improvements validate the effectiveness of crucial components in TFAM.

C. The effect of adaptive semantic view generation

We ablate various design choices to analyze the effect of the semantic part view in our design. As listed in Table VII, "#4 Dual-TR+Random Crop" indicates that the local Transformer hierarchy is directly fed with the content randomly cropped from the raw image. "#5 Dual-TR+Center Crop" suggests that the local Transformer hierarchy is directly fed with the center content fixedly cropped from the raw image. "#6 Dual-TR+SPVG" means that the local Transformer hierarchy takes semantic part view generation as input. The experimental results are summarized in Table VII. It is concluded that "#6 Dual-TR+SPVG" achieves the best performance among different variants. We ascribe such improvements to the flexible part view generation, which turn out to be beneficial for the FGVC task. Moreover, we present the visualizations of parallel hierarchies in Fig. 6, where we show that the visualizations from the local Transformer hierarchy are complementary to the counterpart of the global hierarchy, validating the effectiveness of our design choice.

TABLE VIII
ABLATION STUDY ON ECT AND CL. ECT REFERS TO EXCHANGING CLASS TOKENS, AND CL IS DEFINED AS CENTER LOSS.

#	ECT	CL	Acc.@1 (%)
1	✗	✗	91.4
2	✓	✗	91.8
3	✗	✓	91.7
4	✓	✓	92.0

D. The effectiveness of exchanging class tokens

According to Table VIII, exchanging class tokens can contribute to a significant improvement of 0.4% in classification accuracy. Besides, it demonstrates that exchanging class tokens provides a simple yet effective way to consistently ensure the parallel hierarchies towards better performance, enabling the Dual-TR to learn a more distinctive representation and achieve better performance.

E. The affect of center loss

We also summarize the performance comparison between baseline ("#1") and a variant ("#3") with center loss in

Table VIII. When coupled with the center loss, it can be observed that there is a noticeable improvement in classification performance. The center loss module produces loss penalties based on feature distances between inputs and their category centers, leading to compact feature representation in the training process. Therefore, we believe that the introduction of center loss is well-suited to learning intra-class mutual features within the same sub-category and effectively alleviate the issue of small inter-class variance.

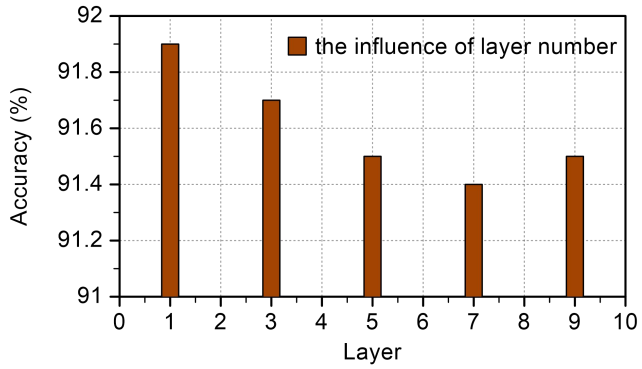


Fig. 7. Influence of Dual-TR using the attention weights of different layers to generate semantic part view on CUB-200-2011 dataset.

F. The influence of layers Transformer architecture

As Fig. 7 demonstrates, when more layers of attention weights are involved, there is a dramatic drop in classification performance. We argue that too many layers of the Transformer architecture applied may lead to a dilemma in which the foreground and background features are entangled, and the semantic part view contains much irrelevant information, impairing the classification accuracy. Thus, for better performance, we choose the attention weights of the first layer to generate a semantic part view empirically.

TABLE IX
ABLATION STUDY ON THE EFFECT OF ε ON CUB-200-2011 DATASET.

Values of ε	0.1	0.3	0.5	0.7	0.9
Acc.@1	91.5	92.0	91.7	91.5	91.2

G. The effectiveness of threshold ε

To determine a desirable value of the empirical threshold value ε, we conduct experiments with the different thresholds and summarize results in Table IX. Experiments in the table show that the classification accuracy significantly fluctuates along with the value ε varying from 0.1 to 0.9. Hence, the value of ε is empirically set to 0.3 in our experimental settings by default for better classification performance.

VI. CONCLUSION

In this paper, we design a dual Transformer architecture (abbreviated as Dual-TR) to explicitly encode discriminative features from global and local perspectives for the FGVC task. Dual-TR is well-designed to encode fine-grained objects via parallel Transformer hierarchies, where we generate the adaptive semantic view for the local Transformer hierarchy. The essence within each hierarchy lies in orthogonal assembly based on attention weight derived by the ViT structure, *i.e.*, intra-layer and inter-layer assembly. The former explores informative token features in various self-attention heads within the Transformer layer, while the latter pays more attention to token assembly across the Transformer layers. Extensive experimental results on CUB-200-2011, NABirds, Stanford Dogs, and iNaturalist2017 demonstrate that Dual-TR achieves competitive or even better performances compared to recent SOTAs. Comprehensive ablation studies verify the effectiveness of the proposed method and evidence that such fine-grained feature assembly is suitable for this visual understanding task.

Dual-TR also encounters some challenges, including the trade-off between the computation cost and accuracy when designing the model architecture, and the results on occluded objects, which we leave for future work. Nevertheless, based on the above promising achievements, we believe this work can inspire the future research in this field to explore effective solutions for better performance.

REFERENCES

[1] X. Shu, B. Xu, L. Zhang, and J. Tang, "Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2022.

[2] L. Zhang, G. Du, F. Liu, H. Tu, and X. Shu, "Global-local multiple granularity learning for cross-modality visible-infrared person reidentification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2021.

[3] J. Zhao, J. Li, H. Liu, S. Yan, and J. Feng, "Fine-grained multi-human parsing," *Int. J. Comput. Vis.*, vol. 128, no. 8, pp. 2185–2203, 2020.

[4] J. Li, J. Zhao, C. Lang, Y. Li, Y. Wei, G. Guo, T. Sim, S. Yan, and J. Feng, "Multi-human parsing with a graph-based generative adversarial model," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 17, no. 1, pp. 29:1–29:21, 2021.

[5] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and A new benchmark for multi-human parsing," in *ACM Multimedia*, 2018, pp. 792–800.

[6] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," in *NeurIPS*, 2020.

[7] Y. Lu, C. Yuan, X. Li, Z. Lai, D. Zhang, and L. Shen, "Structurally incoherent low-rank 2dlpp for image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 6, pp. 1701–1714, 2019.

[8] Y. Shan, X. Zhou, S. Liu, Y. Zhang, and K. Huang, "Siamfpn: A deep learning method for accurate and real-time maritime ship tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 315–325, 2021.

[9] N. Han, J. Wu, X. Fang, W. K. Wong, Y. Xu, J. Yang, and X. Li, "Double relaxed regression for image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 307–319, 2020.

[10] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1092–1108, 2020.

[11] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. M. Elgammal, and D. N. Metaxas, "SPDA-CNN: unifying semantic part detection and abstraction for fine-grained recognition," in *CVPR*, 2016, pp. 1143–1152.

[12] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: deep localization, alignment and classification for fine-grained recognition," in *CVPR*, 2015, pp. 1666–1674.

[13] X. He, Y. Peng, and J. Zhao, "Fast fine-grained image classification via weakly supervised discriminative localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1394–1407, 2019.

[14] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *ICCV*, 2017, pp. 5219–5227.

[15] L. Zhang, S. Huang, W. Liu, and D. Tao, "Learning a mixture of granularity-specific experts for fine-grained categorization," in *ICCV*, 2019, pp. 8330–8339.

[16] T. Zhang, D. Chang, Z. Ma, and J. Guo, "Progressive co-attention network for fine-grained visual classification," *arXiv preprint arXiv:2101.08527*, 2021.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, and T. Unterthiner, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020, pp. 213–229.

[19] B. Duke, A. Ahmed, C. Wolf, P. Aarabi, and G. W. Taylor, "SSTVOS: sparse spatiotemporal transformers for video object segmentation," *CoRR*, 2021.

[20] T. Meinhardt, A. Kirillov, L. Leal-Taixé, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," *CoRR*, 2021.

[21] J. Wang, X. Yu, and Y. Gao, "Feature fusion vision transformer fine-grained visual categorization," *arXiv preprint arXiv:2107.02341*, 2021.

[22] J. He, J. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, C. Wang, and A. L. Yuille, "Transfg: A transformer architecture for fine-grained recognition," *CoRR*, 2021.

[23] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *ECCV*, 2018, pp. 420–435.

[24] Q. Cui, Q. Jiang, X. Wei, W. Li, and O. Yoshie, "Exchnet: A unified hashing network for large-scale fine-grained image retrieval," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 12348, pp. 189–205.

[25] H. Hanselmann and H. Ney, "Elope: Fine-grained visual classification with efficient localization, pooling and embedding," in *CACV*, 2020, pp. 1247–1256.

[26] R. Du, D. Chang, A. K. Bhunia, J. Xie, Z. Ma, Y. Song, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *ECCV*, 2020, pp. 153–168.

[27] Y. Ding, Z. Ma, S. Wen, J. Xie, and D. Chang, "Ap-cnn: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification," *TIP*, vol. 30, pp. 2826–2836, 2021.

[28] F. Zhang, M. Li, G. Zhai, and Y. Liu, "Multi-branch and multi-scale attention learning for fine-grained visual categorization," in *MMM*, 2021, pp. 136–147.

[29] T. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *ICCV*, 2015, pp. 1449–1457.

[30] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *ECCV*, 2018, pp. 595–610.

[31] P. Zhuang, Y. Wang, and Y. Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *AAAI*, 2020, pp. 13 130–13 137.

[32] F. Tan, Y. Kong, Y. Fan, F. Liu, D. Zhou, H. Zhang, L. Chen, L. Gao, and Y. Qian, "Sdnet: mutil-branch for single image deraining using swin," *CoRR*, vol. abs/2105.15077, 2021.

[33] D. Zhou, Y. Shi, B. Kang, W. Yu, Z. Jiang, Y. Li, X. Jin, Q. Hou, and J. Feng, "Refiner: Refining self-attention for vision transformers," *CoRR*, 2021.

[34] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *CoRR*, 2021.

[35] J. Wang, X. Yu, and Y. Gao, "Feature fusion vision transformer for fine-grained visual categorization," *CoRR*, 2021.

[36] Y. Hu, X. Jin, Y. Zhang, H. Hong, J. Zhang, Y. He, and H. Xue, "Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition," in *MM*, 2021, pp. 4239–4248.

[37] C. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *ICCV*. IEEE, 2021, pp. 347–356.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[39] P. Du, Z. Sun, Y. Yao, and Z. Tang, "Exploiting category similarity-based distributed labeling for fine-grained visual classification," *IEEE Access*, vol. 8, pp. 186 679–186 690, 2020.

[40] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[41] G. V. Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. J. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *CVPR*, 2015, pp. 595–604.

[42] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, vol. 2, no. 1, 2011.

[43] G. V. Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. J. Belongie, "The inaturalist species classification and detection dataset," in *CVPR*, 2018, pp. 8769–8778.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, pp. 211–252, 2015.

[45] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[46] W. Ge, X. Lin, and Y. Yu, "Weakly supervised complementary parts models for fine-grained image classification from the bottom up," in *CVPR*, 2019, pp. 3034–3043.

[47] H. Sun, X. He, and Y. Peng, "Sim-trans: Structure information modeling transformer for fine-grained visual categorization," *CoRR*, vol. abs/2208.14607, 2022.

[48] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *CVPR*, 2017, pp. 4476–4484.

[49] A. Dubey, O. Gupta, R. Raskar, and N. Naik, "Maximum-entropy fine grained classification," in *NeurIPS*, 2018, pp. 635–645.

[50] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Trans. Multim.*, pp. 1245–1256, 2017.

[51] Y. Gao, X. Han, X. Wang, W. Huang, and M. R. Scott, "Channel interaction networks for fine-grained image categorization," in *AAAI*, 2020, pp. 10 818–10 825.

[52] S. Huang, X. Wang, and D. Tao, "Snapmix: Semantically proportional mixing for augmenting fine-grained data," *CoRR*, vol. abs/2012.04846, 2020.

[53] R. Ji, L. Wen, L. Zhang, D. Du, Y. Wu, C. Zhao, X. Liu, and F. Huang, "Attention convolutional binary neural tree for fine-grained visual categorization," in *CVPR*, 2020, pp. 10 465–10 474.

[54] Y. Cui, Y. Song, C. Sun, A. Howard, and S. J. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *CVPR*, 2018, pp. 4109–4118.

[55] C. Liu, H. Xie, Z. Zha, L. Ma, L. Yu, and Y. Zhang, "Filtration and distillation: Enhancing region attention for fine-grained visual categorization," in *AAAI*, 2020, pp. 11 555–11 562.

[56] T. Zhang, D. Chang, Z. Ma, and J. Guo, "Progressive co-attention network for fine-grained visual classification," *CoRR*, 2021.

[57] M. V. Conde and K. Turgutlu, "Exploring vision transformers for fine-grained classification," *CoRR*, 2021.

[58] X. Liu, L. Wang, and X. Han, "Transformer with peak suppression and knowledge guidance for fine-grained image recognition," *CoRR*, 2021.

[59] Y. Zhang, J. Cao, L. Zhang, X. Liu, Z. Wang, F. Ling, and W. Chen, "A free lunch from vit: Adaptive attention multi-scale fusion transformer for fine-grained visual recognition," *CoRR*, vol. abs/2110.01240, 2021.

[60] Y. Wang, S. Ye, S. Yu, and X. You, "R2-trans: Fine-grained visual categorization with redundancy reduction," *CoRR*, vol. abs/2204.10095, 2022.

[61] D. Korsch, P. Bodesheim, and J. Denzler, "Classification-specific parts for improving fine-grained visual categorization," in *DAGM GCPR*, 2019, pp. 62–75.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[63] A. Recasens, P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba, "Learning to zoom: A saliency-based sampling layer for neural networks," in *ECCV*, 2018, pp. 52–67.

[64] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in *CVPR*, 2020, pp. 8659–8669.

[65] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017, pp. 4278–4284.

[66] H. Zheng, J. Fu, Z. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *CVPR*, 2019, pp. 5012–5021.
[67] G. Sun, H. Cholakkal, S. H. Khan, F. S. Khan, and L. Shao, "Fine-grained recognition: Accounting for subtle differences between similar classes," in *AAAI*, 2020, pp. 12 047–12 054.
[68] W. Luo, H. Zhang, J. Li, and X. Wei, "Learning semantically enhanced feature for fine-grained image classification," *IEEE Signal Process. Lett.*, pp. 1545–1549, 2020.

Ruyi Ji received the Ph.D. degree in software engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2021. He now is a post doctor researcher at the Institute of Software Chinese Academy of Sciences, working with Prof. Yanjun Wu. His current research interests includes machine learning and computer vision, with a focus on image processing and pattern recognition.

Jiaying Li received the B.Eng. degree from the China Jiliang University, in 2019. She is currently pursuing the M.Sc. degree with the Institute of Automation, Beijing Information Science and Technology University. Her research directions include pattern recognition, specifically focusing on fine-grained image classification.

Libo Zhang received the Ph.D. degree in computer software and theory from the University of Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Associate Research Professor with the Institute of Software Chinese Academy of Sciences, Beijing. He is selected as a Member of Youth Innovation Promotion Association, Chinese Academy of Sciences, and Outstanding Youth Scientist of Institute of Software Chinese Academy of Sciences. His current research interests include image processing and pattern recognition.

Jing Liu (Member, IEEE) received the B.E. and M.S. degrees from Shandong University, Shandong, in 2001 and 2004, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2008. She is currently a professor with School of Artificial Intelligence, University of Chinese Academy of Sciences and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her current research interests include deep learning, image content analysis and classification, and multimedia understanding and retrieval.

Yanjun Wu received the B.Eng. degree in computer science from Tsinghua University, in 2006, and the Ph.D. degree in computer science from the Institute of Software Chinese Academy of Sciences (ISCAS), Beijing, China. He is currently a Research Professor with ISCAS. Also, he is the Director with Intelligent Software Research Center, ISCAS. His current research interests include computer vision, operating systems, and system security.