Feature Fusion Vision Transformer for Fine-Grained Visual Categorization

Jun Wang
jun.wang.3@warwick.ac.uk
University of Warwick, UK

Xiaohan Yu
xiaohan.yu@griffith.edu.au
Griffith University, Australia

Yongsheng Gao
yongsheng.gao@griffith.edu.au
Griffith University, Australia

Abstract

The core for tackling the fine-grained visual categorization (FGVC) is to learn subtle yet discriminative features. Most previous works achieve this by explicitly selecting the discriminative parts or integrating the attention mechanism via CNN-based approaches. However, these methods enhance the computational complexity and make the model dominated by the regions containing the most of the objects. Recently, vision transformer (ViT) has achieved SOTA performance on general image recognition tasks. The self-attention mechanism aggregates and weights the information from all patches to the classification token, making it perfectly suitable for FGVC. Nonetheless, the classification token in the deep layer pays more attention to the global information, lacking the local and low-level features that are essential for FGVC. In this work, we propose a novel pure transformer-based framework Feature Fusion Vision Transformer (FFVT) where we aggregate the important tokens from each transformer layer to compensate the local, low-level and middle-level information. We design a novel token selection module called mutual attention weight selection (MAWS) to guide the network effectively and efficiently towards selecting discriminative tokens without introducing extra parameters. We verify the effectiveness of FFVT on four benchmarks where FFVT achieves the state-of-the-art performance. Code is available at this link.

1 Introduction

Fine-grained visual categorization (FGVC) aims to solve the problem of differentiating subordinate categories under the same basic-level category, e.g., birds, cars and plants. FGVC has wide real-world applications, such as autonomous driving and intelligent agriculture. Some FGVC tasks are exceedingly hard for human beings due to the small inter-class variance and large intra-class variance, e.g., recognizing 200 subordinate plant leaves and 200 subordinate birds. Therefore, FGVC is an important and highly challenging task.

Owing to the decent designed networks and large-scale annotated datasets, FGVC has gained steady improvements in recent years. Current methods on FGVC can be roughly divided into localization-based methods and attention-based methods. The core for solving FGVC is to learn the discriminative features in images. Early localization-based methods

[1, 16, 41] achieve this by directly annotating the discriminative parts in images. However, it is costly and time-consuming to build bounding box annotations, hindering the applicability of these methods on real-world applications. To alleviate this problem, recent localization-based methods normally integrate the region proposal network (RPN) to obtain the potential discriminative bounding boxes. These selected proposals are then fed into the backbone network to gain the local features. After that, most methods often adopt a rank loss [7] on the classification outputs for all local features. However, [14] argues that RPN-based methods ignore the relationships among selected regions. Another problem is that this mechanism drives the RPN to propose large bounding boxes as they are more likely to contain the foreground objects. Confusion occurs when these bounding boxes are inaccurate and cover the background rather than objects. Besides, some discriminative regions, e.g., leaf vein in plant leaves, cannot be simply annotated by a rectangular [35].

Attention-based [40, 50, 54] methods automatically detect the discriminative regions in images via self-attention mechanism. These methods release the reliance on manually annotation for discriminative regions and have gained encouraging results. Recently, vision transformer has demonstrated potential performance on general image classification [6], image retrieval [9] and semantic segmentation [54]. This great success shows that the innate attention mechanism of a pure transformer architecture can automatically search the important parts in images that contribute to image recognition. However, few study investigate the performance of vision transformer in FGVC. As the first work to study the vision transformer on FGVC, [14] proposed to replace the inputs of the final transformer layer with some important tokens and gained improved results. Nonetheless, the final class token may concern more on global information and pay less attention to local and low-level features, defecting the performance of vision transformer on FGVC since local information plays an important role in FGVC. Besides, previous works focus on FGVC benchmarks containing more than ten thousands of annotated images, and no study explores the capability of vision transformer on small-scale and ultra-fine-grained visual categorization (ultra-FGVC) settings.

In this paper, we propose a novel feature fusion vision transformer (FFVT) for FGVC. FFVT aggregates the local information from low-level, middle-level and high-level tokens to facilitate the classification. We present a novel important token selection approach called Mutual Attention Weight Selection (MAWS) to select the representative tokens on each layer that are added as the inputs of the last transformer layer. In addition, we explore the performance of our method on four FGVC datasets to comprehensively verify the capability of our proposed FFVT on FGVC. In conclusion, our work has four main contributions.

1. To our best knowledge, we are the first study to explore the performance of vision transformer on both small-scale and ultra-FGVC settings. The two small-scale datasets in this paper are highly challenging due to the ultra-fine-grained inter-category variances and few training data available. Some examples are visualized in Figure 1.

2. We propose FFVT, a novel vision transformer framework for fine-grained visual categorization tasks that can automatically detect the distinguished regions and take advantage of different level of global and local information in images.

3. We present a novel important token selection approach called Mutual Attention Weight Selection (MAWS). MAWS can effectively select the informative tokens that are having high similarity to class token both in the contexts of the class token and the token itself without introducing extra parameters.

4. We verify the effectiveness of our method on four fine-grained benchmarks. Experimental results demonstrate that FFVT achieves state-of-the-art performance on them, offering an alternative to current CNN-based approaches. Ablation studies show that our

proposed method boost the performance of the backbone model by 5.42%, 4.67% and 0.80% on CottonCultivar80, SoyCultivarLocal and CUB datasets, respectively.

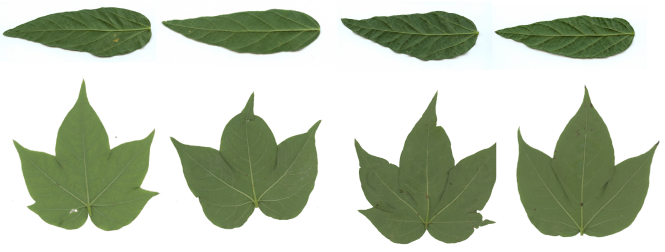

Figure 1: Examples of images in SoyCultivarLocal and Cotton datasets. Images in the first row come from four species of Soy.Loc, while examples in the second row are selected from four categorizes of Cotton.

2 Related Works

2.1 Fine-Grained Visual Categorization

Methods on FGVC can be coarsely divided into two groups: localization-based methods and attention-based methods. Similar to object detection task, localization-based methods often detect the foreground objects and perform classification based on them. Early works [1, 16, 41] achieve this by taking advantage of part annotation to supervise the learning of the detection branch. However, bounding box annotation requires large manual labor, hampering their real-world applications.

To alleviate above problem, recent localization-based methods introduce the weakly supervised object detection (WSOD) technique to predict the potential discriminative regions with only image-level label. Ge et al. [13] used WSOD and instance segmentation techniques to obtain the rough object instances, and then selected the important instances to perform classification. He et al. [15] presented two spatial constraints to select the discriminative parts obtained by the detection branch. Wang et al. [38] utilized correlations between regions to select distinguished parts. However, these methods require a well designed WSOD branch to propose potential discriminative regions. Moreover, the selected parts sent to the classification head often cover the whole object instead of the truly discriminative parts.

Alternatively, attention-based methods automatically localize the discriminative regions via self-attention mechanism without extra annotations. Zhao et al. [50] proposed a diversified visual attention network which uses the diversity of the attention to collect dicriminative information. Xiao et al. [40] presented a two-level attention mechanism to steadily filter out the trivial parts. Similar to [40], Zheng et al. [54] proposed a progressive-attention to progressively detect discriminative parts at multiple scales. However, these methods often suffer from huge computational cost.

2.2 Transformer

Transformer has achieved huge success in natural language processing [9, 30, 31]. Motivated by this, researchers try to exploit the transformers in computer vision. Recent work ViT [6]

achieves the state-of-the-art performance on image classification by employing a pure transformer architecture on a number of fix-sized image patches. Later, researchers explore the performance of the pure transformer in other computer vision tasks. Zheng [54] et al. developed a pure transformer SETR on semantic segmentation task. Alaaeldin et al. [9] exploited a transformer to generate the image descriptor for image retrieval task. Nonetheless, few studies explore the vision transformer on FGVC.

The most similar to our work is TransFG [14] which is the first study to extend the ViT into FGVC, while there are two notable differences between FFVT and TransFG. First, TransFG selects the discriminative tokens and directly send them to the last transformer layer (no feature fusion), while FFVT aims to aggregate the local and different level information from each layer to enrich the feature representation capability via feature fusion. Second, our proposed token selection strategy is totally different from that of TransFG which requires the attention information from all transformer layer to generate the selected token indexes via matrix multiplication. In contrast, our proposed MAWS utilize attention information from only one transformer layer to produce the corresponding indexes. Hence, MAWS is simple and efficient. Our work is also in accordance with the spirit of recent research [11, 33, 34, 35, 36, 43, 44, 45, 46, 47, 48, 49, 51, 52], which focuses on localizing subtle yet vital regions.

3 Methods

To better comprehend our method, we first briefly review the knowledge of vision transformer ViT in Section 3.1. Our proposed methods are then elaborately described in the following subsections.

3.1 ViT For Image Recognition

ViT follows the similar architecture of transformer in natural language processing with minor modification. Transformer in natural language processing takes a sequence of tokens as inputs. Similarly, given an image with resolution $H * W$, vision transformer first processes the image into $N = \lfloor \frac{H}{P} \rfloor * \lfloor \frac{W}{P} \rfloor$ fix-sized patches x_p. where P is the size for each patch.

The patches x_p are then linearly projected into a D-dimensional latent embedding space. To introduce the positional differences, a learnable vector called position embedding with the same size of patch embedding is directly added to patch embedding. Similar to the class token in BERT [4], an extra class token is added to interact with all patch embeddings and undertakes the classification task. The procedure is shown in Eq (1):

$$\boldsymbol{z}_0 = [\boldsymbol{x}_{class}; \boldsymbol{x}_p^1\boldsymbol{E}; \boldsymbol{x}_p^2\boldsymbol{E}; ...; \boldsymbol{x}_p^N\boldsymbol{E}] + \boldsymbol{E}_{pos} \tag{1}$$

Where $\boldsymbol{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ is the patch embedding projection and C is the number of the image channels. $\boldsymbol{E}_{pos}$ denotes the position embedding.

After that, these patch embeddings are fed into the transformer encoder containing several multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks. Note that all layers retain the same latent vector size D. The outputs of the l-th layer are calculated by Eqs (2) to (3):

$$\boldsymbol{z}_l^{'} = MSA(LN(\boldsymbol{z}_{l-1})) + \boldsymbol{z}_{l-1} \tag{2}$$

$$\boldsymbol{z}_l = MLP(LN(\boldsymbol{z}_l^{'})) + \boldsymbol{z}_l^{'}. \tag{3}$$

Where $LN(\cdot)$ is the layer normalization operation and z_l denotes the encoded image representation. Eventually, a classification head implemented by a MLP block is applied to the class token $\boldsymbol{z}_l^0$ to obtain the predicted category.

3.2 FFVT Architecture

[14] suggests that the ViT cannot capture enough local information required for FGVC. To cope with this problem, we propose to fuse the low-level features and middle-level features to enrich the local information. We present a novel token selection approach called mutual attention weight selection (MAWS) to determine the tokens to be aggregated in the deep layer. This section introduces the details of our proposed FFVT. The overall architecture of FFVT is illustrated in Fig 2.

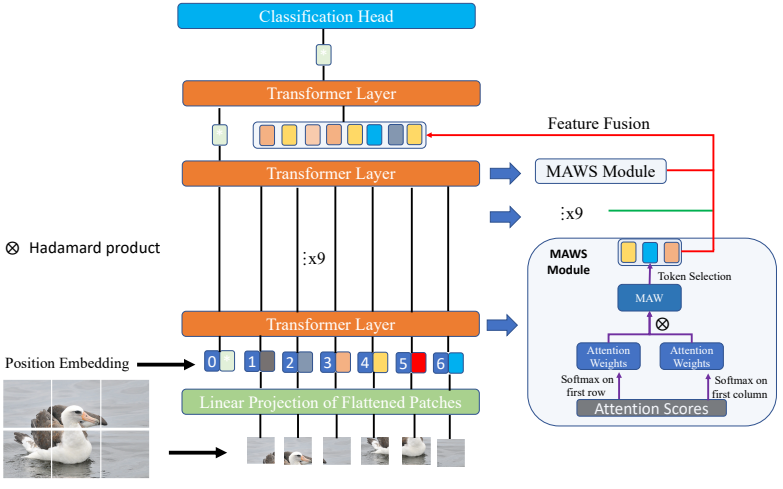

Figure 2: The overall architecture of the proposed FFVT. Images are split into a sequence of fix-sized patches which are then linearly projected into the embedding space. Combined with the position embedding, the patch embeddings are fed into the Transformer Encoder to learn the patch features. Feature fusion is exploited before the last transformer layer to aggregate the important local, low-level and middle level information from previous layers. This is implemented by replacing the inputs (exclude classification token) of the last transformer layer with the tokens selected by the MAWS Module.

3.2.1 Feature Fusion Module

The key challenge of the FGVC is to detect the discriminative regions that significantly contribute to figuring out the subtle differences among subordinate categories. Previous works often achieve this by manually annotating the discriminative regions or integrating the RPN module. However, these methods suffer from some problems discussed in Section 1&2, limiting their performance on real-world applications.

The MSA mechanism in vision transform can perfectly meet the above requirement, whereas MSA in deep layer is likely to pay more attention to the global information. Therefore, we propose a feature fusion module to compensate local information. As shown in

figure 2, given the important tokens (hidden features) from each layer selected by MAWS module, we replace the inputs (except for the class token) of the last transformer layer with our selected tokens. In this way, the class token in the last transformer layer fully interacts with the low-level, middle level and high-level features from the previous layers, enriching the local information and feature representation capability.

Specifically, we denote the tokens selected by MAWS module in l-layer as:

$$\boldsymbol{z}_l^{local} = [z_l^1,\ z_l^2,\ ...,\ z_l^K] \tag{4}$$

Where K is the number of selected features. The fused features along with the classification token fed into the last transformer layer L are:

$$\boldsymbol{z}_{ff} = [\boldsymbol{z}_{L-1}^0;\ \boldsymbol{z}_1^{local};\ \boldsymbol{z}_2^{local};...;\ \boldsymbol{z}_{L-1}^{local}] \tag{5}$$

Eventually, following the ViT, the classification token of the final transformer layer is sent to the classification head to perform categorization. The problem turns to how to select the important and discriminative tokens. To that end, we propose an effective and efficient token selection approach described in the next section.

3.2.2 Mutual Attention Weight Selection Module

Since an image is split into many patches, token selection turns to be an important problem. Noise is added when the background patches are frequently selected, while discriminative patches can boost the model performance. Hence, we propose a token selection approach which directly utilizes the attention scores generated by multi-head self-attention module.

To be specific, an attention score matrix for one attention head $A \in \mathbb{R}^{(N+1)\times(N+1)}$ is denoted as:

$$\boldsymbol{A} = [\boldsymbol{a}^0;\ \boldsymbol{a}_1;\ \boldsymbol{a}_2;\ ...;\ \boldsymbol{a}_i;\ ...;\ \boldsymbol{a}_N] \tag{6}$$

$$\boldsymbol{a}_i = [a_{i,0},\ a_{i,1},\ a_{i,2},\ ...,\ a_{i,j},\ ...,\ a_{i,N}] \tag{7}$$

Where $a_{i,j}$ is the attention score between token i and j in the context of token i, i.e., dot-product between the query of token i and the key of token j.

One simple strategy is to pick the tokens having the higher attention scores with the classification token as the classification token contains rich information for categorization. We can do this by sorting the $\boldsymbol{a}_0$ and picking the K tokens with the bigger value. We denote this strategy as single attention weight selection (SAWS). However, SAWS may introduce noisy information since the selected tokens could aggregate much information from noisy patches. Taking a three-patch attention score matrix γ shown below as an example:

$$\gamma = \begin{bmatrix} \boldsymbol{a}^0 \\ \boldsymbol{a}^1 \\ \boldsymbol{a}^2 \\ \boldsymbol{a}^3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 1 & 1 \end{bmatrix}$$

Token three is selected as it has the biggest value in the attention score vector for classification token. However, token three aggregates much information from token one (the maximum attention score in $\boldsymbol{a}_3$) thus may introduce noises assuming token one is a noisy token. To cope with this problem, we develop a mutual attention weight selection module

Datasets	Category	Training	Testing
CUB-birds	200	5994	5794
Stanford Dogs	196	8144	8041
Soy.Loc	200	600	600
Cotton	80	240	240

Table 1: Statistics of CUB-200-2011, SoyCultivarLocal (Soy.Loc), CottonCultivar80 (Cotton) and Stanford Dogs datasets.

which requires the selected tokens to be similar to the classification token both in the contexts of the classification token and the tokens themselves.

In particular, we denote the first column in the attention score matrix as $\boldsymbol{b}_0$. Note that $\boldsymbol{b}_0$ is the attention score vector between classification token and other tokens in the context of **other tokens** compared with $\boldsymbol{a}_0$ in the context of **classification token**. The mutual attention weight $\boldsymbol{ma}_i$ between the classification token and token i is then calculated by Eqs (8) to (9):

$$\boldsymbol{ma}_i = a'_{0,i} * b'_{i,0} \tag{8}$$

$$a'_{0,i} = \frac{e^{a_{0,i}}}{\sum_{j=0}^{N} e^{a_{0,j}}}, \; b'_{i,0} = \frac{e^{b_{i,0}}}{\sum_{j=0}^{N} e^{b_{j,0}}} \tag{9}$$

For multi-head self-attention, we first average the attention scores of all heads. After obtaining the mutual attention weight, the indexes of important tokens are collected according to the mutual attention values. Our approach does not introduce extra learning parameters. It is simple and efficient compared with the matrix multiplication in [14].

4 Experiments

4.1 Datasets

We explore the effectiveness of FFVT on two widely used FGVC dataset and two small-scale ultra-fine-grained datasets, i.e., CUB-200-2011 [32], Stanford Dogs [18], SoyCultivar-Local [46] and CottonCultivar80 [46]. The SoyCultivarLocal and CottonCultivar80 are two highly challenging datasets as they further reduce the granularity of categorization, e.g. from species to cultivar, and with few training data available. The statistics of four datasets are shown in Table 1.

4.2 Implementation Details

The same as the most current transformer-based approaches, the backbone network (ViT) of FFVT is pretrained on the ImageNet21K dataset. Following the same data augmentation methods on most existing works, input images are first resized to 500×500 for Soy.Loc and Cotton datasets, and 600×600 for CUB and Stanford Dogs. We then crop the image into 384×384 for Soy.Loc and Cotton, and 448×448 for CUB and Stanford Dogs (Random cropping in training and center cropping in testing). Random horizontal flipping is adopted and an extra color augmentation is applied for CUB. K in Eq (4) is set to 12 for CUB, Soy.Loc and Cotton, and 24 for Stanford Dogs.

We select the SGD optimizer to optimize the network with a momentum of 0.9. The initial learning rate is 0.02 with the cosine annealing scheduler for FFVT on CUB, Soy.Loc Cotton datasets, and 0.003 on the Stanford Dogs dataset. The batch size is set to 8 for all datasets except for the Stanford Dogs with a batch size of 4. For fair comparisons, we reimplement the experiments of ViT and TransFG on the Stanford Dogs benchmark with their default settings and the same batch size as FFVT. Experiments are conducted on four Nvidia 2080Ti GPUs using PyTorch deep learning framework.

4.3 Comparison with the State-Of-The-Art

Here, we demonstrate the experimental results on four datasets and compare our method with a number of state-of-the-art works. As shown in Table 2, FFVT obtains the second best-performed method on CUB with an accuracy of 91.6%, beating other methods by a large margin except for the most recent state-of-the-art fine-grained method TransFG (-0.1%). Note that FFVT achieves a comparable accuracy against TransFG with much less computation cost and GPU memory consumption since the overlapping strategy of TransFG significantly increases the number of the input patches from 784 to 1296. Besides, limited by our computation resources, the batch size of TransFG on the experiment of CUB dataset is two times larger than FFVT. This may also account for the relative performance differences. FFVT outperforms all the listed approaches on Stanford Dogs with an accuracy of 91.5%, strongly exceeding the second best-performed TransFG by 0.9%.

Table 2: Comparison of different methods on CUB-200-2011 datasets. The best accuracy is highlighted in bold and the second best accuracy is underlined.

Method	Backbone	Accuracy
ResNet50 [15]	ResNet50	84.5
GP-256 [39]	VGG16	85.8
MaxEnt [8]	DenseNet161	86.6
DFL-CNN [37]	ResNet50	87.4
NTS-Net [42]	ResNet50	87.5
Cross-X [23]	ResNet50	87.7
DCL [3]	ResNet50	87.8
CIN [12]	ResNet101	88.1
DBTNet [53]	ResNet101	88.1
ACNet [17]	ResNet50	88.1
S3N [5]	ResNet50	88.5
FDL [22]	DenseNet161	89.1
PMG [7]	ResNet50	89.6
API-Net [55]	DenseNet161	90.0
StackedLSTM [13]	GoogleNet	90.4
ViT [6]	ViT-B_16	90.8
TransFG [14]	ViT-B_16	**91.7**
FFVT	ViT-B_16	<u>91.6</u>

SoyCultivarLocal and CottonCultivar80 are two extremely challenging ultra-fine-grained datasets. The difficulty lies in two folds, i.e., super-subtle inter-class differences and few training images (three for each category). Some examples are visualized in figure 1. Therefore, locating the discriminative regions plays an essential role in accurate classification.

Table 3: Comparison of different methods on Stanford Dogs (Dogs) dataset. The best accuracy is highlighted in bold and the second best accuracy is underlined. Values in parentheses are reported results in their papers.

Method	Backbone	Dogs
MaxEnt [8]	DenseNet161	83.6
FDL [22]	DenseNet161	84.9
RA-CNN [10]	VGG19	87.3
DB [27]	ResNet50	87.7
SEF [24]	ResNet50	88.8
Cross-X [23]	ResNet50	88.9
API-Net [65]	DenseNet161	90.3
ViT [6]	ViT-B_16	90.2
TransFG [14]	ViT-B_16	90.6 (92.3)
FFVT	ViT-B_16	**91.5**

The results of experiments on SoyCultivarLocal and CottonCultivar80 are shown in Table 3. FFVT obtains the highest accuracy of 57.92% on CottonCultivar80, outperforming the second best-performed method by a large margin (+4.17%). Similarly, our proposed FFVT beats all methods with an accuracy of 44.17% on SoyCultivarLocal.

Table 4: Comparison of different methods on SoyCultivarLocal (Soy.Loc) and CottonCultivar80 (Cotton) datasets. The best accuracy is highlighted in bold and the second best accuracy is underlined.

Method	Backbone	Cotton	Soy.Loc
AlexNet [19]	AlexNet	22.92	19.50
VGG16 [26]	VGG16	50.83	39.33
ResNet50 [15]	ResNet50	52.50	38.83
InceptionV3 [28]	GoogleNet	37.50	23.00
MobileNetV2 [25]	MobileNet	49.58	34.67
Improved B-CNN [21]	VGG16	45.00	33.33
NTS-Net [42]	ResNet50	51.67	42.67
fast-MPN-COV [20]	ResNet50	50.00	38.17
ViT [6]	ViT-B_16	51.25	39.33
DeiT-B [29]	ViT-B_16	53.75	38.67
TransFG [14]	ViT-B_16	45.84	38.67
FFVT	ViT-B_16	**57.92**	**44.17**

4.4 Ablation Studies

We perform the ablation studies on CottonCultivar80, SoyCultivarLocal and CUB to further validate the effectiveness of our proposed methods. SAWS is the single attention weight selection strategy designed in Section 3.2.2. As shown in Table 5, even the simple SAWS strategy can remarkably boost the performance by 4.58%, 3.50% and 0.64% on CottonCultivar80, SoyCultivarLocal and CUB, respectively. The results confirm the necessity of aggregating the local and different level information for vision transformer on FGVC. A bigger

improvement can be seen when applying the MAWS strategy (+6.67%, 4.84% and 0.80% on CottonCultivar80, SoyCultivarLocal and CUB, respectively), showing that MAWS better exploits the attention information. MAWS explicitly selects the most useful tokens thus forces the model to learn from these informative parts.

Table 5: Ablation studies on CottonCultivar80 (Cotton), SoyCultivarLocal (Soy.Loc), and CUB datasets. The best accuracy is highlighted in bold.

Method	Cotton	Soy.Loc	CUB
ViT [6]	51.25	39.33	90.85
ViT+Feature Fusion+SAWS	55.83	42.83	91.49
FFVT(ViT+Feature Fusion+MAWS)	**57.92**	**44.17**	**91.65**

We then investigate the influence of the hyper-parameter K. Table 6 summarizes the results of FFVT on the SoyCultivarLocal dataset with the value of K ranging from 10 to 14. FFVT achieves the best performance when there are 12 tokens selected for each layer. One possible reason is that the tokens focused by each attention head are selected by the proposed MAWS module and contribute positively to the classification since this value (12) is in accordance with the number of the attention heads. As K increases from 10 to 12, the accuracy steadily enhances from 43.17% to 44.17%. A different pattern can be seen when K continues increasing to 14, where the accuracy slightly reduces to 42.5%. The performance drop may due to that large K introduces the noisy tokens while small K value lead to insufficient discriminative information for classification. Note that results of all K settings show a significant improvements over that backbone ViT (39.33%), indicating that FFVT is not very sensitive to the value of K.

Table 6: Ablation studies of the hyper-parameter K on SoyCultivarLocal benchmark. The best accuracy is highlighted in bold.

K	10	11	12	13	14
Accuracy(%)	43.17	43.83	**44.17**	43.00	42.50

5 Conclusion

This paper proposes a novel fine-grained visual categorization architecture FFVT and achieves state-of-the-art performance on four benchmarks. To facilitate the performance of vision transformer in FGVC, we propose a feature fusion approach to enrich the local, low-level and middle-level information for the classification token. To select the discriminative tokens that to be aggregated, we develop a novel token selection module MAWS which explicitly takes advantage of the attention scores produced by self-attention mechanism. Experimental results show that FFVT significantly improve the classification accuracy of standard ViT on different fine-grained settings, i.e., normal-scale, small-scale and ultra-fine-grained settings. We observe that FFVT is very effective on the challenging datasets, confirming its capability of capturing subtle differences and discriminative information.

Based on our encouraging results, we believe that the pure-transformer model has the huge potential on different FGVC settings, even in the small-scale datasets without the induction bias like convolutional neural networks.

References

[1] Thomas Berg and Peter N Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 955–962, 2013.

[2] Wei Chen, Tie-Yan Liu, Yanyan Lan, Zhi-Ming Ma, and Hang Li. Ranking measures and loss functions in learning to rank. *Advances in Neural Information Processing Systems*, 22:315–323, 2009.

[3] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6599–6608, 2019.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, pages 153–168. Springer, 2020.

[8] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine-grained classification. *arXiv preprint arXiv:1809.05934*, 2018.

[9] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *arXiv preprint arXiv:2102.05644*, 2021.

[10] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017.

[11] Yongsheng Gao and Maylor KH Leung. Face recognition using line edge map. *IEEE transactions on pattern analysis and machine intelligence*, 24(6):764–779, 2002.

[12] Yu Gao, Xintong Han, Xun Wang, Weilin Huang, and Matthew Scott. Channel interaction networks for fine-grained image categorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10818–10825, 2020.

[13] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3034–3043, 2019.

[14] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, and Alan Yuille. Transfg: A transformer architecture for fine-grained recognition. *arXiv preprint arXiv:2103.07976*, 2021.

[15] Xiangteng He and Yuxin Peng. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[16] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1173–1182, 2016.

[17] Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, and Feiyue Huang. Attention convolutional binary neural tree for fine-grained visual categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10468–10477, 2020.

[18] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[20] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 947–955, 2018.

[21] Tsung-Yu Lin and Subhransu Maji. Improved bilinear pooling with cnns. *arXiv preprint arXiv:1707.06772*, 2017.

[22] Chuanbin Liu, Hongtao Xie, Zheng-Jun Zha, Lingfeng Ma, Lingyun Yu, and Yongdong Zhang. Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11555–11562, 2020.

[23] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S Davis, Jun Li, Jian Yang, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8242–8251, 2019.

[24] Wei Luo, Hengmin Zhang, Jun Li, and Xiu-Shen Wei. Learning semantically enhanced feature for fine-grained image classification. *IEEE Signal Processing Letters*, 27:1545–1549, 2020.

[25] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] Guolei Sun, Hisham Cholakkal, Salman Khan, Fahad Khan, and Ling Shao. Fine-grained recognition: Accounting for subtle differences between similar classes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12047–12054, 2020.

[28] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[29] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[30] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[32] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[33] Jun Wang, Qianying Liu, Haotian Xie, Zhaogang Yang, and Hefeng Zhou. Boosted efficientnet: detection of lymph node metastases in breast cancer using convolutional neural networks. *Cancers*, 13(4):661, 2021.

[34] Jun Wang, Zhao Yang, Linglong Qian, Xiaohan Yu, and Yongsheng Gao. EAR-NET: Error Attention Refining Network For Retinal Vessel Segmentation. In *2021 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2021.

[35] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Mask guided attention for fine-grained patchy image classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1044–1048, 2021.

[36] Jun Wang, Hefeng Zhou, and Xiaohan Yu. PGTRNet: Two-phase Weakly SupervisedObject Detection with Pseudo Ground Truth Refining. *arXiv preprint arXiv:2104.00231*, 2021.

[37] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4148–4157, 2018.

[38] Zhihui Wang, Shijie Wang, Pengbo Zhang, Haojie Li, Wei Zhong, and Jianjun Li. Weakly supervised fine-grained image classification via correlation-guided discriminative learning. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1851–1860, 2019.

[39] Xing Wei, Yue Zhang, Yihong Gong, Jiawei Zhang, and Nanning Zheng. Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 355–370, 2018.

[40] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015.

[41] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, and Bo Zhang. Hierarchical part matching for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1641–1648, 2013.

[42] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 420–435, 2018.

[43] Xiaohan Yu, Shengwu Xiong, and Yongsheng Gao. Leaf image retrieval using combined feature of vein and contour. In *2015 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2015.

[44] Xiaohan Yu, Shengwu Xiong, Yongsheng Gao, Yang Zhao, and Xiaohui Yuan. Multiscale crossing representation using combined feature of contour and venation for leaf image identification. In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6. IEEE, 2016.

[45] Xiaohan Yu, Yongsheng Gao, Shengwu Xiong, and Xiaohui Yuan. Multiscale contour steered region integral and its application for cultivar classification. *IEEE Access*, 7: 69087–69100, 2019.

[46] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Shengwu Xiong, and Xiaohui Yuan. Patchy image structure classification using multi-orientation region transform. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12741–12748, 2020.

[47] Xiaohan Yu, Yang Zhao, Yongsheng Gao, and Shengwu Xiong. Maskcov: A random mask covariance network for ultra-fine-grained visual categorization. *Pattern Recognition*, page 108067, 2021.

[48] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Shengwu Xiong, and Xiaohui Yuan. Benchmark platform for ultra-fine-grained visual categorization beyond human performance. In *International Conference on Computer Vision (ICCV)*, 2021.

[49] Baochang Zhang, Yongsheng Gao, Sanqiang Zhao, and Jianzhuang Liu. Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. *IEEE transactions on image processing*, 19(2):533–544, 2009.

[50] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan. Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia*, 19(6):1245–1256, 2017.

[51] Yang Zhao, Chunhua Shen, Xiaohan Yu, Hao Chen, Yongsheng Gao, and Shengwu Xiong. Learning deep part-aware embedding for person retrieval. *Pattern Recognition*, 116:107938, 2021.

[52] Yang Zhao, Xiaohan Yu, Yongsheng Gao, and Chunhua Shen. Learning discriminative region representation for person retrieval. *Pattern Recognition*, 121:108229, 2022.

[53] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning deep bilinear transformation for fine-grained image representation. *arXiv preprint arXiv:1911.03621*, 2019.

[54] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021.

[55] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13130–13137, 2020.