TransFG: A Transformer Architecture for Fine-Grained Recognition

Ju He[1], Jie-Neng Chen[1], Shuai Liu[3],
Adam Kortylewski[2], Cheng Yang[3], Yutong Bai[1], Changhu Wang[3]

[1] Johns Hopkins University
[2] Max Planck Institute for Informatics
[3] ByteDance Inc.

Abstract

Fine-grained visual classification (FGVC) which aims at recognizing objects from subcategories is a very challenging task due to the inherently subtle inter-class differences. Most existing works mainly tackle this problem by reusing the backbone network to extract features of detected discriminative regions. However, this strategy inevitably complicates the pipeline and pushes the proposed regions to contain most parts of the objects thus fails to locate the really important parts. Recently, vision transformer (ViT) shows its strong performance in the traditional classification task. The self-attention mechanism of the transformer links every patch token to the classification token. In this work, we first evaluate the effectiveness of the ViT framework in the fine-grained recognition setting. Then motivated by the strength of the attention link can be intuitively considered as an indicator of the importance of tokens, we further propose a novel Part Selection Module that can be applied to most of the transformer architectures where we integrate all raw attention weights of the transformer into an attention map for guiding the network to effectively and accurately select discriminative image patches and compute their relations. A contrastive loss is applied to enlarge the distance between feature representations of confusing classes. We name the augmented transformer-based model TransFG and demonstrate the value of it by conducting experiments on five popular fine-grained benchmarks where we achieve state-of-the-art performance. Qualitative results are presented for better understanding of our model.

Introduction

Fine-grained visual classification aims at classifying subclasses of a given object category, e.g., subcategories of birds (Wah et al. 2011; Van Horn et al. 2015), cars (Krause et al. 2013), aircrafts (Maji et al. 2013). It has long been considered as a very challenging task due to the small inter-class variations and large intra-class variations along with the deficiency of annotated data, especially for the long-tailed classes. Benefiting from the progress of deep neural networks (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014; He et al. 2016), the performance of FGVC has obtained a steady progress in recent years. To avoid labor-intensive part annotation, the community currently focuses on weakly-supervised FGVC with
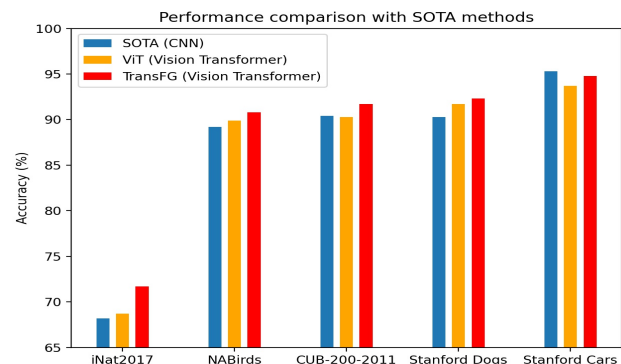
Figure 1: An overview of performance comparison of ViT and TransFG with state-of-the-art methods CNNs on five datasets. We achieve state-of-the-art performance on most datasets while performing a little bit worse on Stanford Cars possibly due to the more regular and simpler car shapes.

only image-level labels. Methods now can be roughly classified into two categories, i.e., localization methods and feature-encoding methods. Compared to feature-encoding methods, the localization methods have the advantage that they explicitly capture the subtle differences among subclasses which is more interpretable and yields better results.

Early works in localization methods rely on the annotations of parts to locate discriminative regions while recent works (Ge, Lin, and Yu 2019a; Liu et al. 2020; Ding et al. 2019) mainly adopt region proposal networks (RPN) to propose bounding boxes which contain the discriminative regions. After obtaining the selected image regions, they are resized into a predefined size and forwarded through the backbone network again to acquire informative local features. A typical strategy is to use these local features for classification individually and adopt a rank loss (Chen et al. 2009) to maintain consistency between the quality of bounding boxes and their final probability output. However, this mechanism ignores the relation between selected regions and thus inevitably encourages the RPN to propose large bounding boxes that contain most parts of the objects which fails to locate the really important regions. Sometimes these bounding boxes can even contain large areas of background

and lead to confusion. Additionally, the RPN module with different optimizing goals compared to the backbone network makes the network harder to train and the re-use of backbone complicates the overall pipeline.

Recently, the vision transformer (Dosovitskiy et al. 2020) achieved huge success in the classification task which shows that applying a pure transformer directly to a sequence of image patches with its innate attention mechanism can capture the important regions in images. A series of extended works on downstream tasks such as object detection (Carion et al. 2020) and semantic segmentation (Zheng et al. 2021; Xie et al. 2021; Chen et al. 2021) confirmed the strong ability for it to capture both global and local features.

These abilities of the Transformer make it innately suitable for the FGVC task as the early long-range "receptive field" (Dosovitskiy et al. 2020) of the Transformer enables it to locate subtle differences and their spatial relation in the earlier processing layers. In contrast, CNNs mainly exploit the locality property of image and only capture weak long-range relation in very high layers. Besides, the subtle differences between fine-grained classes only exist in certain places thus it is unreasonable to convolve a filter which captures the subtle differences to all places of the image.

Motivated by this opinion, in the paper, we present the first study which explores the potential of vision transformers in the context of fine-grained visual classification. We find that directly applying ViT on FGVC already produces satisfactory results while a lot of adaptations according to the characteristics of FGVC can be applied to further boost the performance. To be specific, we propose Part Selection Module which can find the discriminative regions and remove redundant information. A contrastive loss is introduced to make the model more discriminative. We name this novel yet simple transformer-based framework TransFG, and evaluate it extensively on five popular fine-grained visual classification benchmarks (CUB-200-2011, Stanford Cars, Stanford Dogs, NABirds, iNat2017). An overview of the performance comparison can be seen in Fig 1 where our TransFG outperforms existing SOTA CNN methods with different backbones on most datasets. In summary, we make several important contributions in this work:

1. To the best of our knowledge, we are the first to verify the effectiveness of vision transformer on fine-grained visual classification which offers an alternative to the dominating CNN backbone with RPN model design.

2. We introduce TransFG, a novel neural architecture for fine-grained visual classification that naturally focuses on the most discriminative regions of the objects and achieve SOTA performance on several benchmarks.

3. Visualization results are presented which illustrate the ability of our TransFG to accurately capture discriminative image regions and help us to better understand how it makes correct predictions.

Related Work

In this section, we briefly review existing works on fine-grained visual classification and transformer.

Fine-Grained Visual Classification

Many works have been done to tackle the problem of fine-grained visual classification and they can roughly be classified into two categories: localization methods (Ge, Lin, and Yu 2019a; Liu et al. 2020; Yang et al. 2021) and feature-encoding methods (Yu et al. 2018; Zheng et al. 2019; Gao et al. 2020). The former focuses on training a detection network to localize discriminative part regions and reuse them to perform classification. The latter targets at learning more informative features by either computing higher-order information or finding the relationships among contrastive pairs.

Localization FGVC Methods Previously, some works (Branson et al. 2014; Wei, Xie, and Wu 2016) tried to exploit the part annotations to supervise the learning procedure of the localization process. However, since such annotations are expensive and usually unavailable, weakly-supervised parts proposal with only image-level labels draw more attentions nowadays. Ge et al. (Ge, Lin, and Yu 2019a) exploited Mask R-CNN and CRF-based segmentation alternatively to extract object instances and discriminative regions. Yang et al. (Yang et al. 2021) proposed a re-ranking strategy to re-rank the global classification results based on the database constructed with region features. However, these methods all need a special designed module to propose potential regions and these selected regions need to be forwarded through the backbone again for final classification which is not required in our model and thus keeps the simplicity of our pipeline.

Feature-encoding Methods The other branch of methods focus on enriching the feature representations to obtain better classification results. Yu et al. (Yu et al. 2018) proposed a hierarchical framework to do cross-layer bilinear pooling. Zheng et al. (Zheng et al. 2019) adopted the idea of group convolution to first split channels into different groups by their semantic meanings and then do the bilinear pooling within each group without changing the dimension thus it can be integrated into any existed backbones directly. However, these methods are usually not interpretable such one does not know what makes the model distinguish subcategories with subtle differences while our model drops unimportant image patches and only keeps those that contain most information for the fine-grained recognition.

Transformer

Transformer and self-attention models have greatly facilitated research in natural language processing and machine translation (Dai et al. 2019; Devlin et al. 2018; Vaswani et al. 2017). Inspired by this, many recent studies try to apply transformers in computer vision area. Initially, transformer is used to handle sequential features extracted by CNN backbone for the videos (Girdhar et al. 2019). Later, transformer models are further extended to other popular computer vision tasks such as object detection (Carion et al. 2020; Zhu et al. 2020), segmentation (Xie et al. 2021; Wang et al. 2021), object tracking (Sun et al. 2020). Most recently, pure transformer models are becoming more and more popular. ViT (Dosovitskiy et al. 2020) is the first work to show that applying a pure transformer directly to a sequence of image patches can yield state-of-the-art performance on image
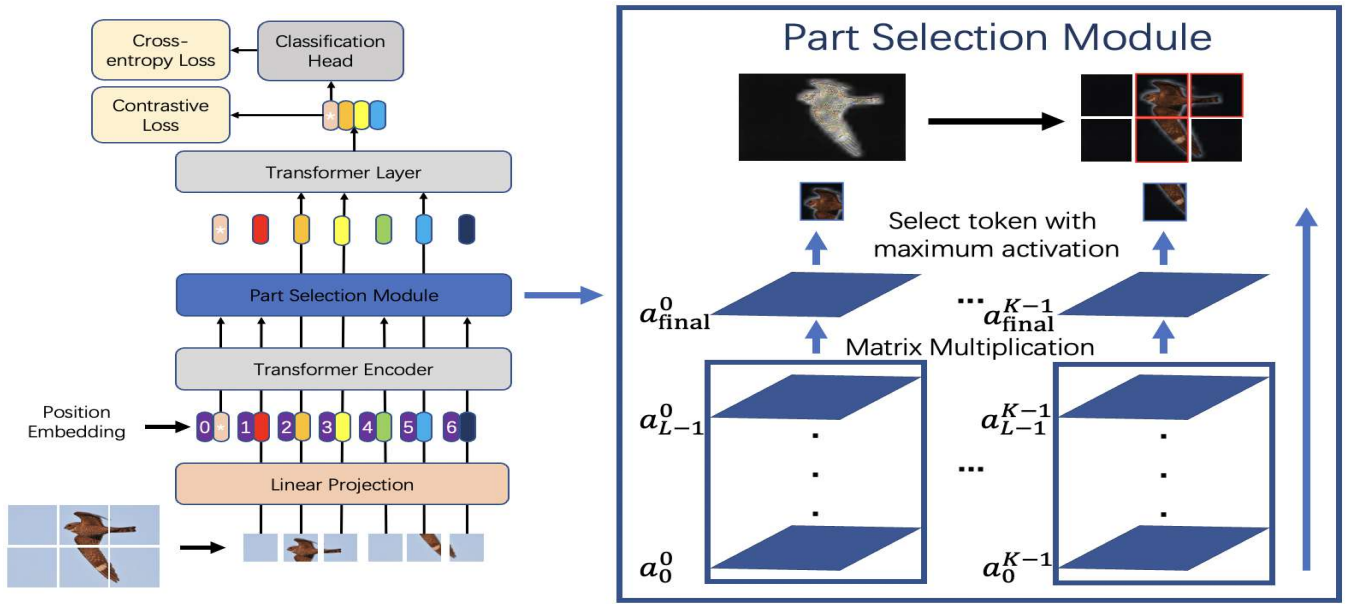
Figure 2: The framework of our proposed TransFG. Images are split into small patches (a non-overlapping split is shown here) and projected into the embedding space. The input to the Transformer Encoder consists of patch embeddings along with learnable position embeddings. Before the last Transformer Layer, a Part Selection Module (PSM) is applied to select tokens that corresponds to the discriminative image patches and only use these selected tokens as input. Best viewed in color.

classification. Based on that, Zheng et al. (Zheng et al. 2021) proposed SETR to exploit ViT as the encoder for segmentation. He et al. (He et al. 2021) proposed TransReID which embedded side information into transformer along with the JPM to boost the performance on object re-identification. In this work, we extend ViT to fine-grained visual classification and show its effectiveness.

Method

We first briefly review the framework of vision transformer and show how to do some preprocessing steps to extend it into fine-grained recognition. Then, the overall framework of TransFG will be elaborated.

Vision Transformer as Feature Extractor

Image Sequentialization. Following ViT, we first preprocess the input image into a sequence of flattened patches x_p. However, the original split method cut the images into non-overlapping patches, which harms the local neighboring structures especially when discriminative regions are split. To alleviate this problem, we propose to generate overlapping patches with sliding window. To be specific, we denote the input image with resolution $H * W$, the size of image patch as P and the step size of sliding window as S. Thus the input images will be split into N patches where

$$N = N_H * N_W = \lfloor \frac{H - P + S}{S} \rfloor * \lfloor \frac{W - P + S}{S} \rfloor \quad (1)$$

In this way, two adjacent patches share an overlapping area of size $(P - S) * P$ which helps to preserve better local region information. Typically speaking, the smaller the step

S is, the better the performance will be. But decreasing S will at the same time requires more computational cost, so a trade-off needs to be made here.

Patch Embedding. We map the vectorized patches x_p into a latent D-dimensional embedding space using a trainable linear projection. A learnable position embedding is added to the patch embeddings to retain positional information as follows:

$$\mathbf{z}_0 = [x_p^1 \mathbf{E}, x_p^2 \mathbf{E}, \cdots, x_p^N \mathbf{E}] + \mathbf{E}_{pos} \quad (2)$$

where N is the number of image patches, $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) * D}$ is the patch embedding projection, and $\mathbf{E}_{pos} \in \mathbb{R}^{N * D}$ denotes the position embedding.

The Transformer encoder (Vaswani et al. 2017) contains L layers of multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks. Thus the output of the l-th layer can be written as follows:

$$\mathbf{z}_l^{'} = MSA(LN(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1} \qquad l \in 1, 2, \cdots, L \quad (3)$$

$$\mathbf{z}_l = MLP(LN(\mathbf{z}_l^{'})) + \mathbf{z}_l^{'} \qquad l \in 1, 2, \cdots, L \quad (4)$$

where $LN(\cdot)$ denotes the layer normalization operation and $\mathbf{z}_l$ is the encoded image representation. ViT exploits the first token of the last encoder layer z_L^0 as the representation of the global feature and forward it to a classifier head to obtain the final classification results without considering the potential information stored in the rest of the tokens.

TransFG Architecture

While our experiments show that the pure Vision Transformer can be directly applied into fine-grained visual classification and achieve impressive results, it does not well capture the local information required for FGVC. To this end,
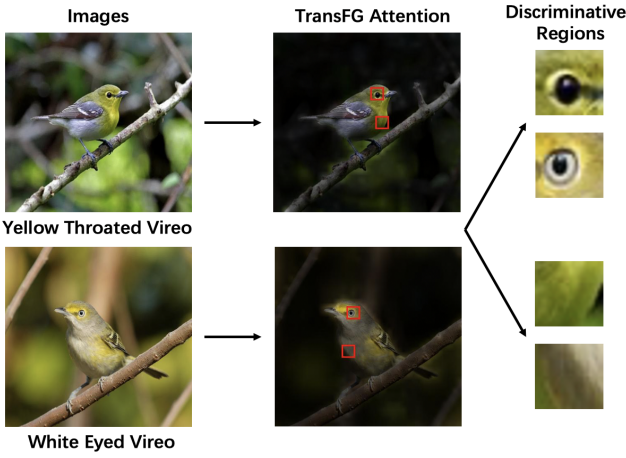
854

Figure 3: A confusing pair of instances from the CUB-200-2011 dataset. Model needs to has the ability to capture the subtle differences in order to classify them correctly. The second column shows the overall attention maps and two selected tokens of our TransFG method. Best viewed in color.

we propose the Part Selection Module (PSM) and apply contrastive feature learning to enlarge the distance of representations between confusing sub-categories. The framework of our proposed TransFG is illustrated in Fig 2.

Part Selection Module   One of the most important problems in fine-grained visual classification is to accurately locate the discriminative regions that account for subtle differences between similar sub-categories. For example, Fig 3 shows a confusing pair of images from the CUB-200-2011 (citation) dataset. The model needs to have the ability to capture the very small differences, i.e., the color of eyes and throat in order to distinguish these two bird species. Region proposal networks and weakly-supervised segmentation strategies are widely introduced to tackle this problem in the traditional CNN-based methods.

Vision Transformer model is perfectly suited here with its innate multi-head attention mechanism. To fully exploit the attention information, we change the input to the last Transformer Layer. Suppose the model has K self-attention heads and the hidden features input to the last layer are denoted as $\mathbf{z}_{L-1} = [z_{L-1}^0; z_{L-1}^1, z_{L-1}^2, \cdots, z_{L-1}^N]$. The attention weights of the previous layers can be written as follows:

$$\mathbf{a}_l = [a_l^0, a_l^1, a_l^2, \cdots, a_l^K] \qquad l \in 1, 2, \cdots, L-1 \quad (5)$$

$$a_l^i = [a_l^{i_0}; a_l^{i_1}, a_l^{i_2}, \cdots, a_l^{i_N}] \qquad i \in 0, 1, \cdots, K-1 \quad (6)$$

Previous works (Serrano and Smith 2019; Abnar and Zuidema 2020) suggested that the raw attention weights do not necessarily correspond to the relative importance of input tokens especially for higher layers of a model, due to lack of token identifiability of the embeddings. To this end, we propose to integrate attention weights of all previous layers. To be specific, we recursively apply a matrix multiplication to the raw attention weights in all the layers as

$$\mathbf{a}_{final} = \prod_{l=0}^{L-1} \mathbf{a}_l \qquad (7)$$

As $\mathbf{a}_{final}$ captures how information propagates from the input layer to the embeddings in higher layers, it serves as a better choice for selecting discriminative regions compared to the single layer raw attention weights a_{L-1}. We then choose the index of the maximum value $A_1, A_2, \cdots, A_K$ with respect to the K different attention heads in $\mathbf{a}_{final}$. These positions are used as index for our model to extract the corresponding tokens in $\mathbf{z}_{L-1}$. Finally, we concatenate the selected tokens along with the classification token as the input sequence which is denoted as:

$$\mathbf{z}_{local} = [z_{L-1}^0; z_{L-1}^{A_1}, z_{L-1}^{A_2}, \cdots, z_{L-1}^{A_K}] \qquad (8)$$

By replacing the original entire input sequence with tokens corresponding to informative regions and concatenate the classification token as input to the last Transformer Layer, we not only keep the global information but also force the last Transformer Layer to focus on the subtle differences between different sub-categories while abandoning less discriminative regions such as background or common features.

Contrastive Feature Learning   Following ViT, we still adopt the first token z_i of the PSM module for classification. A simple cross-entropy loss is not enough to fully supervise the learning of features since the differences between subcategories might be small. To this end, we adopt contrastive loss $\mathcal{L}_{con}$ which minimizes the similarity of classification tokens corresponding to different labels and maximizes the similarity of classification tokens of samples with the same label y. To prevent the loss being dominated by easy negatives (different class samples with little similarity), a constant margin α is introduced that only negative pairs with similarity larger than α contribute to the loss $\mathcal{L}_{con}$. Formally, the contrastive loss over a batch of size B is denoted as:

$$\mathcal{L}_{con} = \frac{1}{B^2} \sum_i^B [\sum_{j:y_i=y_j}^B (1 - Sim(z_i, z_j) +$$
$$\sum_{j:y_i \neq y_j}^B \max((Sim(z_i, z_j) - \alpha), 0)] \qquad (9)$$

where z_i and z_j are pre-processed with $l2$ normalization and $Sim(z_i, z_j)$ is thus the dot product of z_i and z_j.

In summary, our model is trained with the sum of cross-entropy loss L_{cross} and contrastive L_{con} together which can be expressed as:

$$\mathcal{L} = \mathcal{L}_{cross}(y, y') + \mathcal{L}_{con}(z) \qquad (10)$$

where $\mathcal{L}_{cross}(y, y')$ is the cross-entropy loss between the predicted label y' and the ground-truth label y.

Experiments

In this section, we first introduce the detailed setup including datasets and training hyper-parameters. Quantitative analysis is then given followed by ablation studies. We further give qualitative analysis and visualization results to show the interpretability of our model.

Method	Backbone	CUB	Cars
ResNet-50	ResNet-50	84.5	-
NTS-Net	ResNet-50	87.5	93.9
Cross-X	ResNet-50	87.7	94.6
DBTNet	ResNet-101	88.1	94.5
FDL	DenseNet-161	89.1	94.2
PMG	ResNet-50	89.6	95.1
API-Net	DenseNet-161	90.0	**95.3**
StackedLSTM	GoogleNet	90.4	-
DeiT	DeiT-B	90.0	93.9
ViT	ViT-B_16	90.3	93.7
TransFG	ViT-B_16	**91.7**	94.8

Table 1: Comparison of different methods on CUB-200-2011, Stanford Cars.

Experiments Setup

Datasets. We evaluate our proposed TransFG on five widely used fine-grained benchmarks, i.e., CUB-200-2011 (Wah et al. 2011), Stanford Cars (Krause et al. 2013), Stanford Dogs (Khosla et al. 2011), NABirds (Van Horn et al. 2015) and iNat2017 (Van Horn et al. 2018). We also exploit its usage in large-scale challenging fine-grained competitions.

Implementation details. Unless stated otherwise, we implement TransFG as follows. First, we resize input images to $448 * 448$ except $304 * 304$ on iNat2017 for fair comparison (random cropping for training and center cropping for testing). We split image to patches of size 16 and the step size of sliding window is set to be 12. Thus the H, W, P, S in Eq 1 are 448, 448, 16, 12 respectively. The margin α in Eq 9 is set to be 0.4. We load intermediate weights from official ViT-B_16 model pretrained on ImageNet21k. The batch size is set to 16. SGD optimizer is employed with a momentum of 0.9. The learning rate is initialized as 0.03 except 0.003 for Stanford Dogs dataset and 0.01 for iNat2017 dataset. We adopt cosine annealing as the scheduler of optimizer.

Quantitative Analysis

We compare our proposed method TransFG with state-of-the-art works on above mentioned fine-grained datasets. The experiment results on CUB-200-2011 and Stanford Cars are shown in Table 1. From the results, we find that our method outperforms all previous methods on CUB dataset and achieve competitive performance on Stanford Cars.

To be specific, the third column of Table 1 shows the comparison results on CUB-200-2011. Compared to the best result StackedLSTM (Ge, Lin, and Yu 2019b) up to now, our TransFG achieves a **1.3%** improvement on Top-1 Accuracy metric and 1.4% improvement compared to our base framework ViT (Dosovitskiy et al. 2020). Multiple ResNet-50 are adopted as multiple branches in (Ding et al. 2019) which greatly increases the complexity. It is also worth noting that StackLSTM is a very messy multi-stage training model which hampers the availability in practical use, while our TransFG maintains the simplicity.

The fourth column of Table 1 shows the results on Stanford Cars. Our method outperforms most existing methods

Method	Backbone	Dogs
MaxEnt	DenseNet-161	83.6
FDL	DenseNet-161	84.9
Cross-X	ResNet-50	88.9
API-Net	ResNet-101	90.3
ViT	ViT-B_16	91.7
TransFG	ViT-B_16	**92.3**

Table 2: Comparison of different methods on Stanford Dogs.

Method	Backbone	NABirds
Cross-X	ResNet-50	86.4
API-Net	DenseNet-161	88.1
CS-Parts	ResNet-50	88.5
FixSENet-154	SENet-154	89.2
ViT	ViT-B_16	89.9
TransFG	ViT-B_16	**90.8**

Table 3: Comparison of different methods on NABirds.

while performs worse than PMG (Du et al. 2020) and API-Net (Zhuang, Wang, and Qiao 2020) with small margin. We argue that the reason might be the much more regular and simpler shape of cars. However, even with this property, our TransFG consistently gets **1.1%** improvement compared to the standard ViT model.

The results of experiments on Stanford Dogs are shown in Table 2. Stanford Dogs is a more challenging dataset compared to Stanford Cars with its the more subtle differences between certain species and the large variances of samples from the same category. Only a few methods have tested on this dataset and our TransFG outperforms all of them. While ViT (Dosovitskiy et al. 2020) outperforms other methods by a large margin, our TransFG achieves 92.3% accuracy which outperforms SOTA by **2.0%** with its discriminative part selection and contrastive loss supervision.

NABirds is a much larger birds dataset not only from the side of images numbers but also with 355 more categories which significantly makes the fine-grained visual classification task more challenging. We show our results on it in Table 3. We observe that most methods achieve good results by either exploiting multiple backbones for different branches or adopting quite deep CNN structures to extract better features. While the pure ViT (Dosovitskiy et al. 2020) can directly achieve 89.9% accuracy, our TransFG constantly gets 0.9% performance gain compared to ViT and reaches 90.8% accuracy which outperforms SOTA by **1.6%**.

iNat2017 is a large-scale dataset for fine-grained species

Method	Backbone	iNat2017
ResNet152	ResNet152	59.0
IncResNetV2	IncResNetV2	67.3
TASN	ResNet101	68.2
ViT	ViT-B_16	68.7
TransFG	ViT-B_16	**71.7**

Table 4: Comparison of different methods on iNat2017.

Method	Patch Split	Accuracy	Train Time (h)
ViT	N-Overlap	90.3	1.30
ViT	Overlap	**90.5**	3.38
TransFG	N-Overlap	91.5	1.98
TransFG	Overlap	**91.7**	5.38

Table 5: Ablation study on split way of image patches on CUB-200-2011 dataset.

recognition. Most previous methods do not report results on iNat2017 because of the computational complexity of the multi-crop, multi-scale and multi-stage optimization. With the simplicity of our model pipeline, we are able to scale TransFG well to big datasets and evaluate the performance which is shown in Table 4. This dataset is very challenging for mining meaningful object parts and the background is very complicated as well. We find that Vision Transformer structure outperforms ResNet structure a lot in these large challenging datasets. ViT outperformes ResNet152 by nearly 10% and similar phenomenon can also be observed in iNat2018 and iNat2019. Our TransFG is the only method to achieve above 70% accuracy with input size of 304 and outperforms SOTA with a large margin of **3.5%**.

For the just ended iNat2021 competition which contains 10,000 species, 2.7M training images, our TransFG achieves very high single model accuracy of 91.3%. (The final performance was obtained by ensembling many different models along with multi-modality processing) As far as we know, at least two of the Top5 teams in the final leaderboard adopted TransFG as one of their ensemble models. This clear proves that our model can be further extended to large-scale challenging scenarios besides academy datasets.

Ablation Study

We conduct ablation studies on our TransFG pipeline to analyze how its variants affect the fine-grained visual classification result. All ablation studies are done on CUB-200-2011 dataset while the same phenomenon can be observed on other datasets as well.

Influence of image patch split method. We investigate the influence of our overlapping patch split method through experiments with standard non-overlapping patch split. As shown in Table 5, both on the pure Vision Transformer and our improved TransFG framework, the overlapping split method bring consistently improvement, i.e., 0.2% for both frameworks. The additional computational cost introduced by this is also affordable as shown in the fourth column.

Method	Accuracy (%)
ViT	90.3
TransFG	**91.0**

Table 6: Ablation study on Part Selection Module (PSM) on CUB-200-2011 dataset.

Influence of Part Selection Module. As shown in Table 6, by applying the Part Selection Module (PSM) to select discriminative part tokens as the input for the last Transformer

Method	Contrastive Loss	Acc (%)
ViT		90.3
ViT	✓	**90.7**
TransFG		91.0
TransFG	✓	**91.5**

Table 7: Ablation study on contrastive loss on CUB-200-2011 dataset.

Method	Value of α	Accuracy (%)
TransFG	0	91.1
TransFG	0.2	91.4
TransFG	0.4	**91.7**
TransFG	0.6	91.5

Table 8: Ablation study on value of margin α on CUB-200-2011 dataset.

layer, the performance of the model improves from 90.3% to 91.0%. We argue that this is because in this way, we sample the most discriminative tokens as input which explicitly throws away some useless tokens and force the network to learn from the important parts.

Influence of contrastive loss. The comparisons of the performance with and without contrastive loss for both ViT and TransFG frameworks are shown in Table 7 to verify the effectiveness of it. We observe that with contrastive loss, the model obtains a big performance gain. Quantitatively, it increases the accuracy from 90.3% to 90.7% for ViT and 91.0% to 91.5% for TransFG. We argue that this is because contrastive loss can effectively enlarge the distance of representations between similar sub-categories and decrease that between the same categories which can be clearly seen in the comparison of confusion matrix in Fig 4.
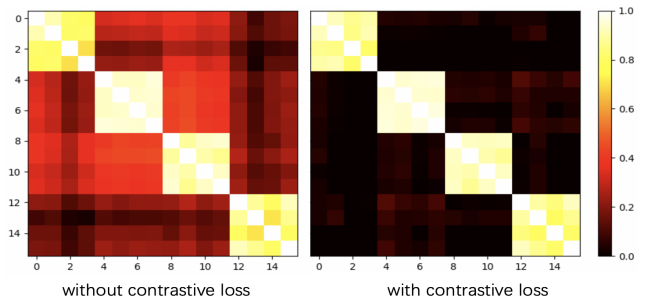

Figure 4: Illustration of contrastive loss. Confusion matrices without and with contrastive loss of a batch with four classes where each contains four samples are shown. The metric of confusion matrix is cosine similarity. Best viewed in color.

Influence of margin α. The results of different setting of the margin α in Eq 9 is shown in Table 8. We find that a small value of α will lead the training signals dominated by easy negatives thus decrease the performance while a high value of α hinder the model to learn sufficient information for increasing the distances of hard negatives. Empirically, we find 0.4 to be the best value of α in our experiments.
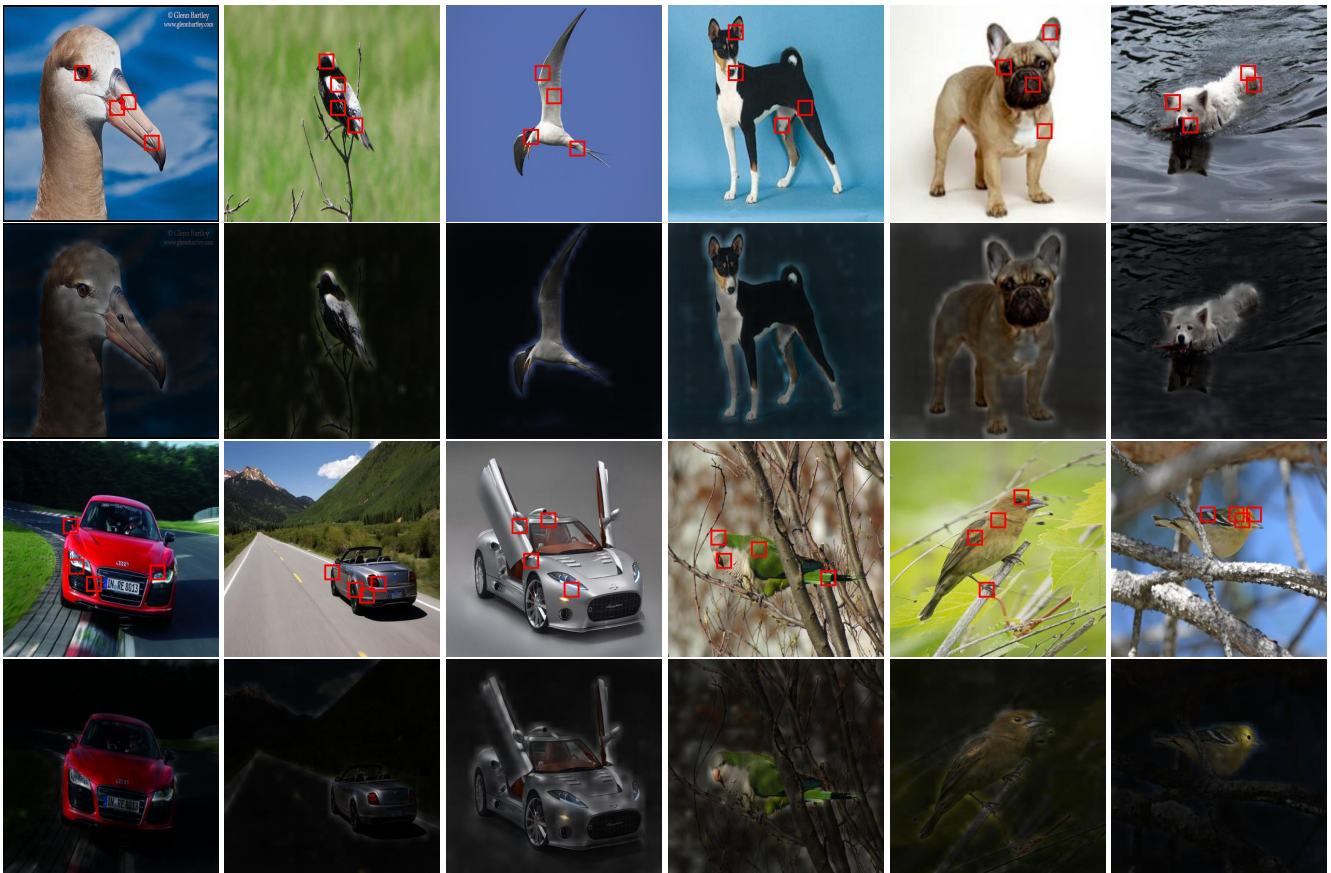
Figure 5: Visualization results of TransFG on CUB-200-2011, Stanford Dogs, Stanford Cars and NABirds datasets. Two kinds of visualization are given, where the first and the third row show the selected Top-4 token positions while the second and fourth rows show the overall global attention maps. See examples from NABirds dataset where birds are sitting on twigs. The bird parts are lighted while the occluded twigs are ignored. Best viewed in color.

Qualitative Analysis

We show the visualization results of proposed TransFG on the four benchmarks in Fig 5. We randomly sample three images from each dataset. Two kinds of visualizations are presented. The first and the third row of Fig 5 illustrated the selected tokens positions. For better visualization results, we only draw the Top-4 image patches (ranked by the attention score) and enlarge the square of the patches by two times while keeping the center positions unchanged. The second and fourth rows show the overall attention map of the whole image where we use the same attention integration method as described above to first integrate the attention weights of all layers followed by averaging the weights of all heads to obtain a single attention map. The lighter a region is, the more important it is. From the figure, we can see that our TransFG successfully captures the most important regions for an object, i.e., head, wings, tail for birds; ears, eyes, legs for dogs; lights, doors for cars. At the same time, our overall attention map maps the entire object precisely even in complex backgrounds and it can even serves as a segmentation mask in some simple scenarios. These visualization results clearly prove the interpretability of our proposed method.

Conclusion

In this work, we propose a novel fine-grained recognition framework TransFG and achieve state-of-the-art results on four common fine-grained benchmarks. We exploit self-attention mechanism to capture the most discriminative regions. Compared to bounding boxes produced by other methods, our selected image patches are much smaller thus becoming more meaningful by showing what regions really contribute to the fine-grained classification. The effectiveness of such small image patches also comes from the Transformer Layer to handle the inner relationships between these regions instead of relying on each of them to produce results separately. Contrastive loss is introduced to increase the discriminative ability of the classification tokens. Experiments are conducted on both traditional academy datasets and large-scale competition datasets to prove the effectiveness of our model in multiple scenarios. Qualitative visualizations further show the interpretability of our method.

With the promising results achieved by TransFG, we believe that the transformer-based models have great potential on fine-grained tasks and our TransFG could be a starting point for future works.

References

Abnar, S.; and Zuidema, W. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

Branson, S.; Van Horn, G.; Belongie, S.; and Perona, P. 2014. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 213–229. Springer.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, W.; Liu, T.-Y.; Lan, Y.; Ma, Z.-M.; and Li, H. 2009. Ranking measures and loss functions in learning to rank. *Advances in Neural Information Processing Systems*, 22: 315–323.

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, R.; Chang, D.; Bhunia, A. K.; Xie, J.; Ma, Z.; Song, Y.-Z.; and Guo, J. 2020. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, 153–168. Springer.

Gao, Y.; Han, X.; Wang, X.; Huang, W.; and Scott, M. 2020. Channel Interaction Networks for Fine-Grained Image Categorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10818–10825.

Ge, W.; Lin, X.; and Yu, Y. 2019a. 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*, 3034–3043.

Ge, W.; Lin, X.; and Yu, Y. 2019b. 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 (CVPR)*.

Girdhar, R.; Carreira, J.; Doersch, C.; and Zisserman, A. 2019. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 244–253.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Fei-Fei, L. 2011. Novel Dataset for Fine-Grained Image Categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*. Sydney, Australia.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F.; Burges, C. J. C.; Bottou, L.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Liu, C.; Xie, H.; Zha, Z.-J.; Ma, L.; Yu, L.; and Zhang, Y. 2020. Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11555–11562.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Serrano, S.; and Smith, N. A. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, P.; Jiang, Y.; Zhang, R.; Xie, E.; Cao, J.; Hu, X.; Kong, T.; Yuan, Z.; Wang, C.; and Luo, P. 2020. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*.

Van Horn, G.; Branson, S.; Farrell, R.; Haber, S.; Barry, J.; Ipeirotis, P.; Perona, P.; and Belongie, S. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 595–604.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, H.; Zhu, Y.; Adam, H.; Yuille, A.; and Chen, L.-C. 2021. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5463–5474.

Wei, X.-S.; Xie, C.-W.; and Wu, J. 2016. Mask-cnn: Localizing parts and selecting descriptors for fine-grained image recognition. *arXiv preprint arXiv:1605.06878*.

Xie, E.; Wang, W.; Wang, W.; Sun, P.; Xu, H.; Liang, D.; and Luo, P. 2021. Segmenting transparent object in the wild with transformer. *arXiv preprint arXiv:2101.08461*.

Yang, S.; Liu, S.; Yang, C.; and Wang, C. 2021. Re-rank Coarse Classification with Local Region Enhanced Features for Fine-Grained Image Recognition. *arXiv preprint arXiv:2102.09875*.

Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; and You, X. 2018. Hierarchical bilinear pooling for fine-grained visual recognition. In *Proceedings of the European conference on computer vision (ECCV)*, 574–589.

Zheng, H.; Fu, J.; Zha, Z.-J.; and Luo, J. 2019. Learning deep bilinear transformation for fine-grained image representation. *arXiv preprint arXiv:1911.03621*.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6881–6890.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zhuang, P.; Wang, Y.; and Qiao, Y. 2020. Learning attentive pairwise interaction for fine-grained classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13130–13137.