*electronics*

*Article*

# MFVT: Multilevel Feature Fusion Vision Transformer and RAMix Data Augmentation for Fine-Grained Visual Categorization

**Xinyao Lv, Hao Xia, Na Li, Xudong Li and Ruoming Lan ***

School of Physics and Electronics, Shandong Normal University, Jinan 250358, China
* Correspondence: sdnu_lanruoming@163.com

**Abstract:** The introduction and application of the Vision Transformer (ViT) has promoted the development of fine-grained visual categorization (FGVC). However, there are some problems when directly applying ViT to FGVC tasks. ViT only classifies using the class token in the last layer, ignoring the local and low-level features necessary for FGVC. We propose a ViT-based multilevel feature fusion transformer (MFVT) for FGVC tasks. In this framework, with reference to ViT, the backbone network adopts 12 layers of Transformer blocks, divides it into four stages, and adds multilevel feature fusion (MFF) between Transformer layers. We also design RAMix, a CutMix-based data augmentation strategy that uses the resize strategy for crop-paste images and label assignment based on attention. Experiments on the CUB-200-2011, Stanford Dogs, and iNaturalist 2017 datasets gave competitive results, especially on the challenging iNaturalist 2017, with an accuracy rate of 72.6%.

**Keywords:** fine-grained visual categorization; Vision Transformer; feature fusion; data augmentation

**Citation:** Lv, X.; Xia, H.; Li, N.; Li, X.; Lan, R. MFVT: Multilevel Feature Fusion Vision Transformer and RAMix Data Augmentation for Fine-Grained Visual Categorization. *Electronics* **2022**, *11*, 3552. https://doi.org/10.3390/electronics11213552

Academic Editor: Xue (Shelley) Lin

Received: 20 September 2022
Accepted: 28 October 2022
Published: 31 October 2022

## 1. Introduction

Image classification is a classic problem in computer vision, whose goal is to categorize images [1]. Fine-grained visual categorization (FGVC) refers to more refined subcategory division on the basis of basic categories, such as distinguishing types of birds and dogs, which is essentially intra-class classification [1–3]. The problem is to achieve a more detailed division of the categories obtained in the traditional classification problem [4,5].

FGVC has common research needs and application scenarios in fields such as smart agriculture and unmanned retail. Therefore, the question of how to design an accurate and efficient FGVC algorithm is of great significance [6]. Compared with ordinary image classification tasks, the image data seen by FGVC have similar appearances and features, and there is interference, such as posture, illumination, perspective, occlusion, and background, resulting in small inter-class variations and large intra-class variations. These difficulties make FGVC a challenging research task [7–9].

Feature extraction is a key factor in determining the accuracy of image classification. Traditional FGVC methods are based on manually extracting image features [9]. However, the artificial feature description ability is limited, and the classification effect is not good. With the rapid development of deep learning, features extracted by neural network training have stronger predictive ability than those extracted manually, which promotes the rapid development of FGVC [10,11].

The self-attention neural network Transformer has become the object of much research in the CV field. The Vision Transformer (ViT) [12] makes full use of the modeling ability of the Transformer's attention mechanism by dividing an image into multiple patch tokens, and it promotes the development of vision tasks such as image classification [12,13]. A number of FGVC algorithms based on ViT, such as TransFG, AFTrans, and FFVT, have also been produced [14–16]. These methods improve the model structure to different degrees, to effectively improve the performance of FGVC algorithms. Nonetheless, there are issues to consider when applying ViT to FGVC tasks. ViT only uses the class token of the last layer for classification, and the deep class token is obtained based on all image patches in the self-attention mode, so it pays more attention to global information. In the FGVC task, landmark details and parts are the key information for classification. According to our experiments, class tokens can extract different levels of information features, and these are complementary. Therefore, better utilization of these levels of information in the FGVC task will allow the model to obtain more comprehensive information from the image for final prediction.

For the model to grasp more detailed information that is helpful for classification, instead of only the most discriminative information, CutMix data augmentation covers the image with part of another image, so that the model can discover more useful information [17,18]. However, this brings new problems. Random cropping and pasting of background image patches that do not contribute to classification will lead to a loss of object information and incorrect label assignment [18–22].

We therefore propose a ViT-based multilevel feature fusion vision transformer (MFVT) for the FGVC task. In addition to a ViT backbone, multilevel feature fusion (MFF) is included. To solve the problem of possible object information loss and label errors due to CutMix, referring to ResizeMix and TransMix, we design RAMix, a CutMix data enhancement strategy based on Resize for image cropping and pasting, and a Transformer-based attention mechanism for label assignment. This paper makes the following contributions.

- The MFVT algorithm has a backbone network that is consistent with ViT. The 12-layer Transformer block divides it into four stages, which require no additional labeling information such as bounding boxes, so as to achieve fine-grained visual categorization.
- The MFF module extracts the features of the last block output of different stages in the backbone network, and uses a lightweight method for fusion, introducing visual information at different levels and effectively improving feature expression.
- RAMix data augmentation reasonably mixes images, and the attention mechanism in ViT effectively guides image label assignment without introducing additional parameters.

## 2. Related Work

Fine-grained visual categorization has been extensively explored in convolutional neural network (CNN)-based methods [23]. Early work such as Part-Based R-CNN and Mask-CNN relied on bounding boxes and part annotation to locate and distinguish regions, but this requires extensive manual annotation, which limits practical application. Subsequent work used only image labels, with an attention mechanism to localize key regions in a weakly supervised manner. Typical methods include RA-CNN, MA-CNN, and DP-Net [2,24]. There is also a focus on enriching the feature representation to achieve better classification results. For example, in Bilinear-CNN [25], two networks coordinate with each other for overall and local detection and feature extraction with high accuracy.

### *2.1. ViT-Based Image Classification*

The Vision Transformer (ViT) first realized the application of the Transformer to image classification. Experimental results have shown that the visual field does not necessarily rely on the CNN, and inputting the image patch sequence into the Transformer can also achieve a good image classification effect. ViT introduced the Transformer to the visual field.

Due to the excellent performance of ViT, many Transformer-based image classification models have been proposed, improving ViT from perspectives in five categories [26]: (1) Transformer-enhanced CNN, where a Transformer block is used to replace part of the convolution module in a convolutional neural network, e.g., VTs, BoTNet; (2) CNN-enhanced Transformer for image classification enhances the Transformer and accelerates its convergence with convolutional biases, e.g., DeiT, ConViT; (3) Transformer image classification with local attention enhancement adapts the image patch structure through a local self-attention mechanism, e.g., TNT, Swin Transformer; (4) Hierarchical Transformer image classification applies similar structures to Transformers, following a hierarchical CNN, e.g., CvT, PVT; (5) Deep Transformer image classification enables the network to learn more complex representations by increasing the depth of the model, e.g., CaiT, DeepViT.

### 2.2. Fine-Grained Visual Categorization Based on ViT

TransFG is said to be the first method to validate the effectiveness of visual Transformers on FGVC tasks. With the introduction and application of ViT, several ViT-based FGVC methods have been proposed, e.g., TransFG, AFTrans, FFVT, and $R^2$-Trans.

In TransFG [14], a region selection module (PSM) integrates all the original attention weights of the Transformer into an attention map to guide the efficient and accurate selection of discriminative image patches and compute the relationship between them. Repetition loss encourages multiple attention heads to focus on different regions, and contrastive loss is applied to further increase the distance between feature representations of similar subclasses. FFVT extends ViT to large-scale FGVC and small-scale ultra-fine-grained visual categorization, with a feature fusion visual Transformer that aggregates local information from low-, mid-, and high-level tokens for classification. Mutual attention weight selection (MAWS) selects a representative token on each layer and adds it as the input of the last Transformer layer.

AFTrans [15] and RAMS-Trans [27] use the same strategy. The most discriminative part of the image is extracted, and it is enlarged and re-input into the network for further learning. The selective attention collection module (SACM) in AFTrans leverages the attention weights in ViT and adaptively filters input patches based on their relative importance. Global and local multiscale pipelines are supervised by weight-sharing encoders, enabling end-to-end training.

To learn local discriminative regional attention, the strength of attention weights is used to measure a patch's importance to the original image, and a multiscale recurrent attention Transformer, RAMS-Trans, utilizes the Transformer's self-attention force mechanism to cyclically learn discriminative regional attention in a multiscale manner. The core of the algorithm, the dynamic patch proposal module (DPPM), guides region enlargement to integrate multiscale image patch blocks.

$R^2$-Trans [28] adaptively adjusts the masking threshold by calculating the proportion of high-weight regions in the segmentation, and moderately extracts background information in the input space. An information bottleneck approach guides the network to learn a minimum sufficient representation in the feature space. MetaFormer [10] is a simple, effective method for the joint learning of vision and various meta-information, using meta-information to improve the performance of fine-grained recognition, providing a strong baseline for FGVC.

## 3. Method

On the whole, our work is mainly divided into two parts. First, for the FGVC task, to extract more detailed image features, we designed a multilevel feature fusion module to improve the ViT. Second, we designed a data enhancement method called RAMix to enhance the network capability.

### 3.1. Algorithm Framework

Figure 1 shows the algorithm framework, which follows ViT and includes patch embedding, a Transformer encoder, a multilevel feature fusion module, classification head components, and data augmentation.

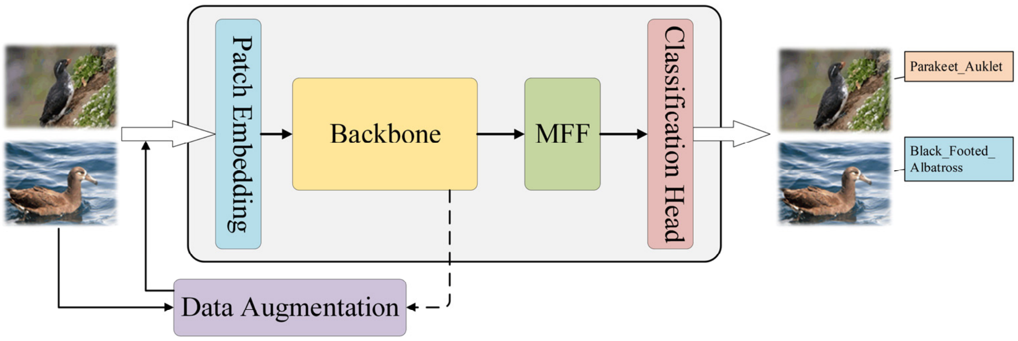

**Figure 1.** Algorithm framework.

The patch embedding module blocks and transforms the image, and embeds the class token and position. The backbone network is consistent with ViT, and includes 12 layers of Transformer blocks for basic feature extraction. A lightweight feature fusion module, MFF, further improves the expressiveness of features. Through parallel classification heads, the final classification is obtained from the weighted average method of three classifications.

In the model training stage, starting from the data, referring to ResizeMix [21] and TransMix [19], RAMix is based on Resize for image cropping and pasting, and a Transformer-based attention mechanism is used for label assignment.

### 3.2. Backbone and MFF

The key challenge of FGVC is to detect discriminative regions that significantly contribute to finding subtle differences between subordinate classes, which can be well met for the multi-head self-attention MSA mechanism in ViT. Deep MSA pays more attention to global information, while FGVC tasks require more attention to detail. Therefore, we propose multilevel feature fusion to ensure the use of high-level global information while obtaining mid- and low-level information. The backbone network and multilevel feature fusion module are the core parts of the algorithm and are, respectively, shown in the left and right of Figure 2.

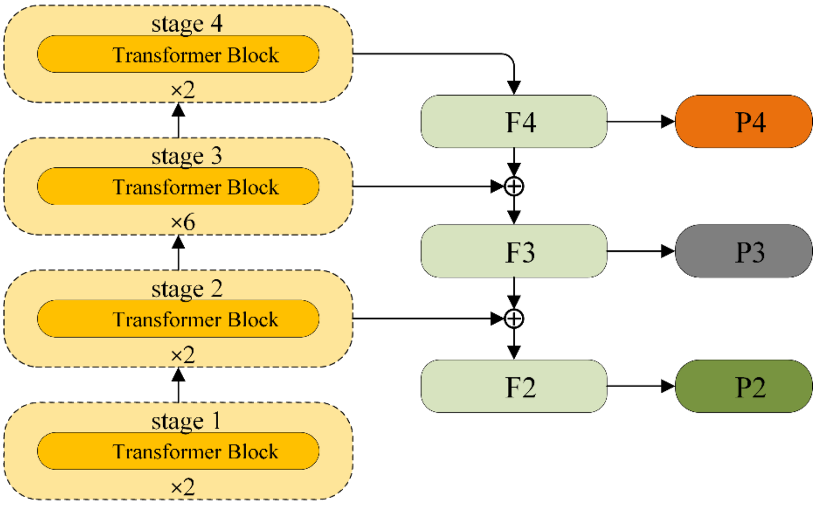

**Figure 2.** Backbone structure and multilevel feature fusion.

The backbone network is consistent with the original ViT. We use a division method similar to Swin Transformer [13] to divide the 12-layer Transformer block into stages 1–4. Stages 1, 2, and 4 have two blocks, and stage 3 has six blocks. The output of each stage contains certain feature information, which is effective and complementary, and the fused features have stronger expressiveness.

The Transformer block is the basic unit of the backbone network. Figure 3 shows the structure of two Transformer blocks connected in series.

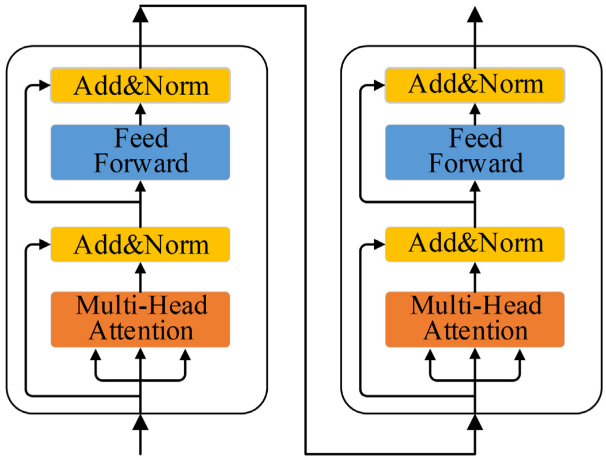

**Figure 3.** Transformer block.

A Transformer block includes multi-head self-attention ($MSA$), multilayer perceptron ($MLP$), and layer normalization ($LN$). The forward propagation of layer $l$ is calculated as

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1} \tag{1}$$

$$z_l = MLP\left(LN\left(z'_l\right)\right) + z'_l \tag{2}$$

and multi-head self-attention is calculated as

$$MSA(Q,K,V) = Concat(head_1,\ldots,head_h)W^o \tag{3}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{4}$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{5}$$

where $Q$, $K$, and $V$ refer to the query, key, and value, respectively; $d_k$ refers to the key vector dimension; $W$ refers to parameters when performing linear transformations on $Q$, $K$, and $V$; and $h$ is the number of heads. The *Concat* operation concatenates the outputs of multiple heads.

Feature fusion combines features from different layers or branches, and often fuses features of different scales to improve the deep learning performance. In the network architecture, low-level features have a higher resolution and more detailed information, and high-level features have stronger semantic information but a poorer perception of details [16]. Their efficient integration is the key to improving the classification model.

After the basic features are extracted from the backbone network, a lightweight feature fusion method is adopted. A consult feature pyramid (FPN) [29,30] in CNN, top-down pathways, and horizontal connections are added to the network structure. As shown on the right side of Figure 2, the last features of stages 4 and 3 are fused. We use three features for classification from different layers: P4 is the output of stage 4, P3 is obtained by fusing the output of stage 3 and P4, and P2 is obtained by fusing the output of stage 2 and P3.

### *3.3. Classification Head and Loss Function*

The detection performance can be improved by combining the detection results of different layers. Before the final fusion is completed, detection starts on the partially fused layer; there will be multiple layers of detection, and the multiple detection results are finally fused.

Instead of using the output features of the last layer for classification prediction, we use multilevel features and fuse the classification prediction results to obtain the final prediction. As shown in Figure 4, we use three classification heads for classification prediction, and each classification head consists of a fully connected layer.

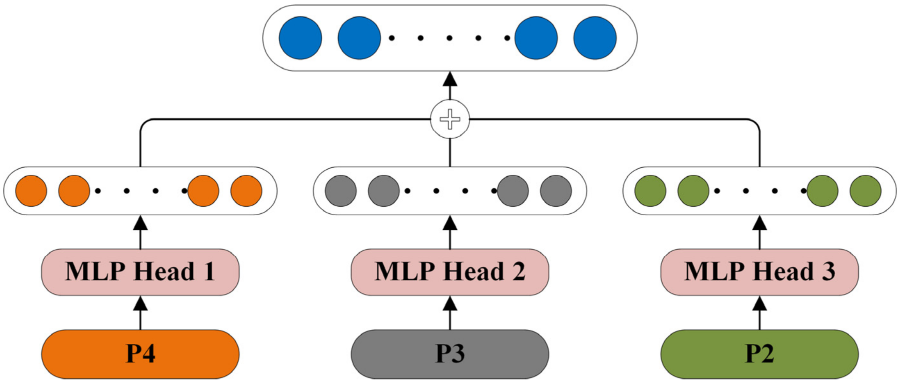

**Figure 4.** Classification header structure.

We obtained three different levels of features from MFF; we use three classification heads to classify the three levels of features. Classification heads 1, 2, and 3 were classified using features P4, P3, and P2, respectively, and the final classification result was obtained by averaging the three classification results.

We use soft target cross-entropy as the loss function in each classification head,

$$loss_h = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}(ylog(\widetilde{y}) + (1-y)\log(1-\widetilde{y})) \tag{6}$$

where $N$ is the number of samples, $n$ is the number of categories, $y$ represents the input label, $\widetilde{y}$ represents the prediction label, $loss_h$ is the loss function of $head_h$, and $h$ takes values from 1 to 3.

To adjust the influence of the classification results of different levels on the final classification, the overall loss function is the weighted average of three loss functions,

$$L = \frac{1}{\alpha+\beta+\gamma}(\alpha loss_1 + \beta loss_2 + \gamma loss_3) \tag{7}$$

where $\alpha$, $\beta$, and $\gamma$ are weight parameters.

### *3.4. RAMix Data Augmentation*

The Transformer has great expressive power, but, according to the experiments of ViT and DeiT [17], the network needs a large quantity of data for model training. Hence, data augmentation is an important part of model training, which can prevent overfitting and improve model performance. We use CutMix data augmentation in the FGVC task, but the random crop and paste operation adds no usefulness to the target image when the cropped area is in the background, without object information. However, when labels are calculated, they will be allocated according to the size ratio of the mixed images, resulting in the loss of object information and label allocation errors. This has a greater impact on the use of small datasets for training in the FGVC task. We design RAMix to address this problem, as Figure 5 shows.

In the training set, images A and B are randomly selected, and $x \in R^{W\times H\times C}$ and $y$ are used to represent the training image and its label, respectively. The goal is to generate a new training sample $(\widetilde{x}, \widetilde{y})$ by combining training samples $(x_A, y_A)$ and $(x_B, y_B)$.

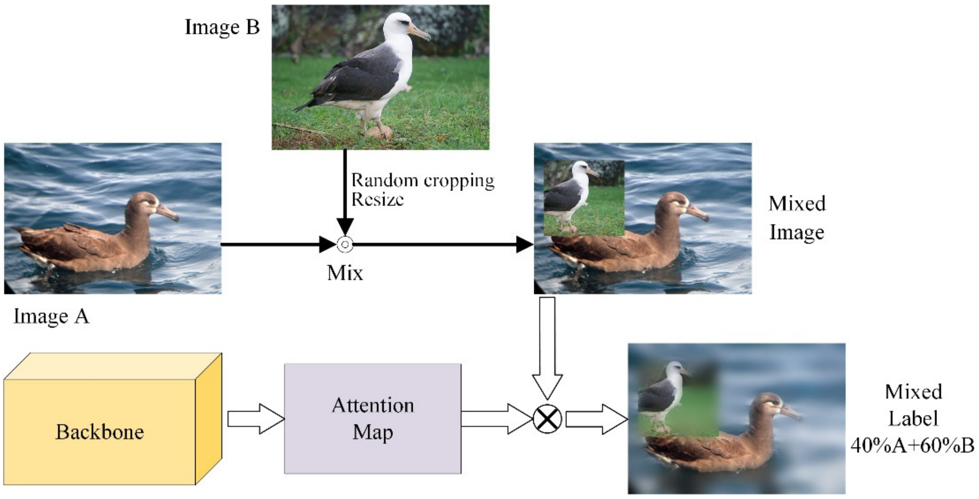

**Figure 5.** RAMix data augmentation.

Image A is randomly cropped to $W\mu \times H\mu$ through the crop value $\mu$ and is reduced to a small image block $P_A$ by the scale ratio $\tau$ and the resize operation, i.e., $P_A = T(x_A)$, where $T()$ represents the reduction operation, $\tau \sim U(\min, \max)$, which means that $\tau$ is evenly distributed between min and max, and minmax denotes the minimum and maximum of the image block $P_A$. To ensure that the image block contains as many objects as possible, and the objects are not too small to be distinguished, we set $\mu = 0.7$, min = 0.25, and max = 1. This means that the minimum image block $P_A$ is a quarter of the size of the input image, and the maximum of the $P_A$ is the same as the input image size.

Image block $P_A$ is pasted onto a random area $R_B$ of image B to generate a new image.

$$\widetilde{x} = Paste(P_A\ , B,\ M) \tag{8}$$

where $M \in \{0,1\}^{W \times H}$ is a binary mask representing location $R_B$. Since scale ratios and paste regions are obtained randomly, this mixing operation adds little cost.

The last step is to assign labels. Due to the different sizes of pasting areas, the occlusion of the target in the original image will be different, as should the assignment of labels, so we have improved the label assignment. We utilize the attention map A instead of the size of the paste region to compute the mixing weight $\lambda$. Labels $y_A$ and $y_B$ of images A and B, respectively, are mixed according to $\lambda$ to obtain the label of the mixed image,

$$\widetilde{y} = \lambda y_A + (1 - \lambda) y_B \tag{9}$$

The calculation of $\lambda$ is guided by the attention map $A$ and calculates the weights that mix the labels of the two sample images,

$$\lambda = A \cdot N \downarrow (M) \tag{10}$$

where $A$ is the attention map from the class token to the image patch tokens, summarizing which patches are most useful to the final classifier. $N \downarrow ()$ denotes nearest-neighbor interpolation down-sampling to transform the original M from HW to $p$ pixels. In this way, the network can learn to dynamically reassign the weights of labels for each data point based on their responses in the attention map. An input that is better focused by the attention map will be assigned a higher value in the mixed label.

## 4. Algorithm Verification

We performed multiple experiments on three datasets in a Linux environment, using the PyTorch deep learning framework on an Nvidia 3090 GPU.

### *4.1. Dataset and Experimental Details*

We evaluated MFVT on the widely used fine-grained datasets iNaturalist 2017 [31], CUB-200-2011, and Stanford Dogs [32]. Details are shown in Table 1.

**Table 1.** Statistics of three datasets.

| Datasets | Training | Testing | Category |
|---|---|---|---|
| iNaturalist 2017 | 579,184 | 95,986 | 5089 |
| CUB-200-2011 | 5994 | 5794 | 200 |
| Stanford Dogs | 12,000 | 8580 | 120 |

In the preprocessing stage, we used the same training strategy for the relatively small CUB-200-2011 and Stanford Dogs. The input image was resized to 600 × 600 and randomly cropped to 448 × 448. For iNaturalist 2017, to reduce the training time, we resized images to 400 × 400 and randomly cropped them to 304 × 304. Finally, random horizontal flipping and RAMix data augmentation were employed for the three datasets.

We chose the SGD optimizer for training, with momentum = 0.9, weight decay = $5 \times 10^{-4}$, and cosine annealing to adjust the learning rate. The initial learning rate was 0.03 for CUB-200-2011 and Stanford Dogs, and 0.02 for iNaturalist 2017. The batch size for all three datasets was 16. The parameters of weighted summation in the loss function were set to $\alpha = \beta = 1$, $\gamma = 0.5$.

### *4.2. Ablation Experiment*

To evaluate the effectiveness and impact of MFF and two-level data augmentation, we conducted ablation studies on the CUB-200-2011 dataset, and the same performance could be observed on the other datasets.

To confirm the validity and complementarity of information at each level, we used different layer classification heads to make final predictions and tried to use the features of multiple stages for fusion. The results are shown in Tables 2 and 3. Experiments were performed without RAMix data enhancement.

**Table 2.** Accuracy of different layers.

| Layer | Accuracy |
|---|---|
| Stage 4 (12th layer) | 90.8 |
| Stage 3 (10th layer) | 89.9 |
| Stage 2 (4th layer) | 68.2 |
| Stage 1 (2nd layer) | 32.1 |

**Table 3.** Multilevel feature fusion ablation test.

| Feature Fusion Layers | Accuracy |
|---|---|
| None | 90.8 |
| Stage 4 + stage 3 | 91.2 |
| Stage 4 + stage 3 + stage 2 | 91.3 |
| Stage 4 + stage 3 + stage 2 + stage 1 | 91.2 |

From Table 2, we can see that even if only the class token of the fourth layer is used for the final classification, an accuracy rate of more than 68% can be achieved, which means that the class token of this layer has features that are effective for classification. It can be seen from Table 3 that feature fusion significantly improves the performance of the model. With the increase in the number of fusion layers, the performance is stronger, which shows the effectiveness and complementarity of the features at each level. However, after merging the features of the first stage, the accuracy of the model decreases. Combined with the results in Table 2, we believe that the underlying features have more noise; hence, the model cannot satisfactorily extract useful information.

We used stage 4 + stage 3 + stage 2 for feature fusion, which resulted in a 0.5% accuracy improvement and a 0.4% increase in computation due to the increased number of classification heads, which we consider reasonable given the performance gain.

As shown in Table 4, we conducted experiments using MFF. Compared with the original ViT, after simple CutMix data enhancement, the accuracy rate is improved, which is consistent with the results in DeiT, but the improvement effect is not obvious. We used RAMix data augmentation to further improve the accuracy, making it more efficient and reliable.

**Table 4.** Accuracy of different data augmentation methods.

| Method | Accuracy |
|---|---|
| ViT | 90.8 |
| ViT + CutMix | 91.1 |
| ViT + RAMix | 91.6 |

When using RAMix, the attention map is the most important point. We obtained this in two ways: obtaining it from the last layer, and obtaining it from each layer and taking the average. The experimental results of these two methods on the CUB-200-2011 dataset are shown in Table 5, which shows little difference between them. To reduce the network complexity and improve the training speed, we choose the first method, which is relatively simple. The results of the complete ablation trial are shown in Table 6.

**Table 5.** Impact of different attention maps.

| Method | Attention Map | Accuracy |
|---|---|---|
| MFVT | Last layer | 91.62 |
| MFVT | Average | 91.59 |

**Table 6.** Overall ablation experiments.

| Method | Accuracy |
|---|---|
| ViT | 90.8 |
| ViT + MFF | 91.3 |
| ViT + MFF + RAMix | 91.6 |

### *4.3. Comparison with State-of-the-Art Methods*

We present experimental results on three datasets and compare our method with some state-of-the-art algorithms. On the iNaturalist 2017 dataset, our method achieves the best results with the same data preprocessing method, with a huge improvement of 0.9%. The method performs competitively on CUB-200-2011 and Stanford Dogs.

iNaturalist 2017 is a large dataset for fine-grained image recognition. The pictures feature visually similar species from around the world, captured in a wide variety of situations. The images are acquired with different types of cameras, differ in image quality, and have a large class imbalance, making them quite challenging. As shown in Table 7, similarly to TransFG and RAMS-Trans, our ViT-based method far outperforms the CNN-based method. Compared to the ViT baseline, we achieved a 3.9% improvement. Consistent with TransFG and RAMS-Trans, our method uses ViT-B_16 as the backbone, and the image input is 304; we achieve a 4.2% improvement compared to RAMS-Trans, and a 0.9% improvement compared to TransFG using the overlapping strategy, which is better than all SOTA methods, proving the effectiveness of our method.

**Table 7.** Accuracy of different methods on iNaturalist 2017.

| Method | Backbone | Accuracy |
|---|---|---|
| SSN | ResNet101 | 65.2 |
| Huang et al. | InResNetV2 | 66.8 |
| IncResNetV2 | InResNetV2 | 67.3 |
| TASN | ResNet101 | 68.2 |
| ViT | ViT-B_16 | 68.7 |
| TransFG | ViT-B_16 | 66.6 |
| TransFG&overlap | ViT-B_16 | 71.7 |
| RAMS-Trans | ViT-B_16 | 68.5 |
| AFTrans | ViT-B_16 | 68.9 |
| MFVT | ViT-B_16 | 72.6 |

As shown in Table 8, MFVT performs competitively on the CUB-200-2011 dataset, with 91.6% accuracy, which is obviously better than the CNN algorithm. Compared to ViT-based algorithms, we achieve comparable results to FFVT, and they are second only to TransFG using an overlapping strategy.

**Table 8.** Accuracy of different methods on CUB-200-2011.

| Method | Backbone | Accuracy |
|---|---|---|
| MA-CNN | VGG-19 | 86.5 |
| FDL | DenseNet161 | 89.1 |
| PMG | ResNet50 | 89.6 |
| API-Net | DenseNet161 | 90.0 |
| StackedLSTM | GoogleNet | 90.4 |
| ViT | ViT-B_16 | 90.8 |
| TransFG | ViT-B_16 | 90.9 |
| TransFG&overlap | ViT-B_16 | 91.7 |
| RAMS-Trans | ViT-B_16 | 91.3 |
| $R^2$-Trans | ViT-B_16 | 91.5 |
| AFTrans | ViT-B_16 | 91.5 |
| FFVT | ViT-B_16 | 91.6 |
| MFVT | ViT-B_16 | 91.6 |

Using the overlapping strategy, TransFG increases the number of patches from 784 to 1296, which increases the computational cost and GPU memory consumption. MFVT achieves similar accuracy without an overlapping strategy.

As shown in Table 9, our method achieves a 0.8% improvement over the ViT baseline on the Stanford Dogs dataset.

**Table 9.** Accuracy of different methods on Stanford Dogs.

| Method | Backbone | Accuracy |
|---|---|---|
| RA-CNN | VGG-19 | 87.3 |
| SEF | ResNet50 | 88.8 |
| Cross-X | ResNet50 | 88.9 |
| API-Net | ResNet101 | 90.3 |
| ViT | ViT-B_16 | 91.2 |
| TransFG | ViT-B_16 | 90.4 |
| TransFG&overlap | ViT-B_16 | 92.3 |
| AFTrans | ViT-B_16 | 91.6 |
| RAMS-Trans | ViT-B_16 | 92.4 |
| MFVT | ViT-B_16 | 92.0 |

### 4.4. Visualization

In order to analyze the effectiveness of the algorithm design and data enhancement, we show the visualization results of RAMix during the training phase and the attention map in the testing phase.

In Figure 6, we show the mixed images in lines 1, 3, and 5. To highlight which image patches have more influence on the classification results, we cover those patches whose influence on the final classification results is less than the threshold, and we display the covered images below the mixed images. It can be seen from the figure that most of the background is covered, while the object part is preserved. As with the object in the original image, the objects in the mixed-in patches receive attention, which means that the strategy of label assignment using the attention map is reasonable and efficient.

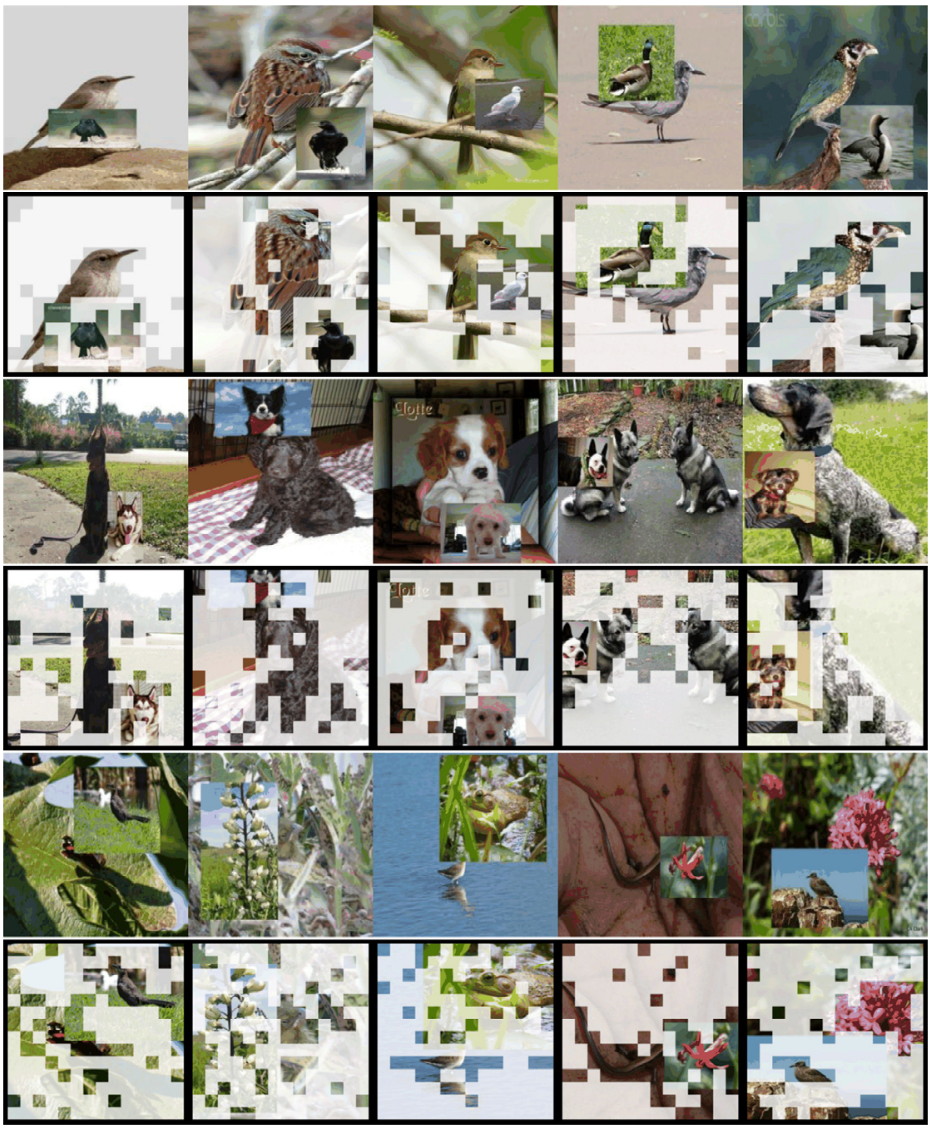

**Figure 6.** Visualization results of RAMix.

In Figure 7, we show the attention heatmap obtained during the inference phase. We can see that the attention is distributed over the entire object, rather than focusing on the most discriminative parts, which validates the effectiveness of our design.

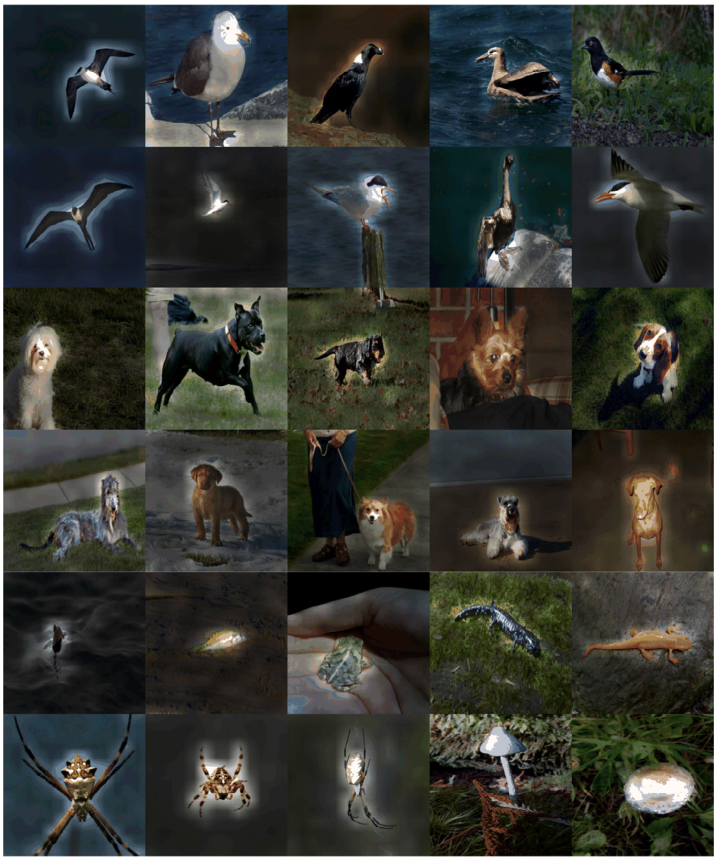

**Figure 7.** Visualization results of attention map.

## 5. Conclusions

We proposed an improved fine-grained visual categorization method, MFVT, based on ViT. To improve the performance of the visual Transformer in FGVC, the backbone network adopted 12 layers of Transformer blocks, divided into four stages, and feature fusion was added between Transformer layers. The feature fusion mechanism integrated high-level information and low-level features. For more accurate and reliable data enhancement, the RAMix data enhancement method was designed.

Experiments on the CUB-200-2011, Stanford Dogs, and iNaturalist 2017 datasets showed that MFVT significantly improved the classification accuracy of the standard ViT in fine-grained environments. We achieved 91.6% and 92.0% accuracy on the CUB-200-2011 and Stanford Dogs datasets; meanwhile, on the more challenging iNaturalist 2017 dataset, the accuracy rate of MFVT reached 72.6%. Based on the experimental results, we believe that the ViT model still has research potential in the field of FGVC.

**Author Contributions:** Conceptualization, X.L. (Xinyao Lv) and R.L.; methodology, X.L. (Xudong Li); software, H.X.; validation, X.L. (Xudong Li), N.L. and R.L.; formal analysis, H.X.; investigation, X.L. (Xudong Li); resources, N.L.; data curation, H.X.; writing—original draft preparation, X.L. (Xinyao Lv); writing—review and editing, R.L.; visualization, X.L. (Xudong Li); project administration, N.L.; funding acquisition, R.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Natural Science Foundation of Shandong Province (ZR2020MF119) and the Shandong Provincial Key Laboratory of Medical Physics and Image Processing Technology.

**Data Availability Statement:** All the details of this work, including data and algorithm codes, are available from the author: 2020020592@stu.sdnu.edu.cn.

**Acknowledgments:** The authors would like to thank the reviewers for their helpful suggestions, which have considerably improved the quality of the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Ge, W.; Lin, X.; Yu, Y. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 3034–3043.
2. Agnes, S.A.; Anitha, J.; Pandian, S.; Peter, J.D. Classification of mammogram images using multiscale all convolutional neural network (MA-CNN). *J. Med. Syst.* **2020**, *44*, 30. [CrossRef] [PubMed]
3. Du, R.; Chang, D.; Bhunia, A.K.; Xie, J.; Ma, Z.; Song, Y.-Z.; Guo, J. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 153–168.
4. Chen, H.; Cheng, L.; Huang, G.; Zhang, G.; Lan, J.; Yu, Z.; Pun, C.-M.; Ling, W.-K. Fine-grained visual classification with multi-scale features based on self-supervised attention filtering mechanism. *Appl. Intell.* **2022**, *52*, 15673–15689. [CrossRef]
5. Fu, J.; Zheng, H.; Mei, T. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 4438–4446.
6. Qiu, C.; Zhou, W. A survey of recent advances in CNN-based fine-grained visual categorization. In Proceedings of the 2020 IEEE 20th International Conference on Communication Technology (ICCT), Nanning, China, 28–31 October 2020; pp. 1377–1384.
7. Ju, M.; Ryu, H.; Moon, S.; Yoo, C.D. GAPNet: Generic-Attribute-Pose Network For Fine-Grained Visual Categorization Using Multi-Attribute Attention Module. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2020; pp. 703–707.
8. Xu, S.; Chang, D.; Xie, J.; Ma, Z. Grad-CAM guided channel-spatial attention module for fine-grained visual classification. In Proceedings of the 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), Gold Coast, Australia, 20–23 September 2021; pp. 1–6.
9. Zhuang, P.; Wang, Y.; Qiao, Y. Learning attentive pairwise interaction for fine-grained classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 13130–13137.
10. Diao, Q.; Jiang, Y.; Wen, B.; Sun, J.; Yuan, Z. MetaFormer: A Unified Meta Framework for Fine-Grained Recognition. *arXiv* **2022**, arXiv:2203.02751.
11. Zhang, F.; Li, M.; Zhai, G.; Liu, Y. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In Proceedings of the International Conference on Multimedia Modeling, Prague, Czech Republic, 22–24 January 2021; pp. 136–147.
12. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
13. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
14. He, J.; Chen, J.-N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C.; Yuille, A. Transfg: A transformer architecture for fine-grained recognition. *arXiv* **2021**, arXiv:2103.07976. [CrossRef]
15. Zhang, Y.; Cao, J.; Zhang, L.; Liu, X.; Wang, Z.; Ling, F.; Chen, W. A free lunch from ViT: Adaptive attention multi-scale fusion Transformer for fine-grained visual recognition. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 22–27 May 2022; pp. 3234–3238.
16. Wang, J.; Yu, X.; Gao, Y. Feature fusion vision transformer for fine-grained visual categorization. *arXiv* **2021**, arXiv:2107.02341.
17. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Chongqing, China, 18–24 July 2021; pp. 10347–10357.
18. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October 2019; pp. 6023–6032.
19. Chen, J.-N.; Sun, S.; He, J.; Torr, P.H.; Yuille, A.; Bai, S. Transmix: Attend to mix for vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 12135–12144.
20. Walawalkar, D.; Shen, Z.; Liu, Z.; Savvides, M. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. *arXiv* **2020**, arXiv:2003.13048.
21. Qin, J.; Fang, J.; Zhang, Q.; Liu, W.; Wang, X.; Wang, X. Resizemix: Mixing data with preserved object information and true labels. *arXiv* **2020**, arXiv:2012.11101.
22. Harris, E.; Marcu, A.; Painter, M.; Niranjan, M.; Prügel-Bennett, A.; Hare, J.S. Understanding and Enhancing Mixed Sample Data Augmentation. *arXiv* **2020**, arXiv:2002.12047.
23. Zheng, H.; Fu, J.; Zha, Z.-J.; Luo, J. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5012–5021.

24. Ding, Y.; Ma, Z.; Wen, S.; Xie, J.; Chang, D.; Si, Z.; Wu, M.; Ling, H. AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. *IEEE Trans. Image Process.* **2021**, *30*, 2826–2836. [CrossRef] [PubMed]
25. Lin, T.-Y.; RoyChowdhury, A.; Maji, S. Bilinear CNN models for fine-grained visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1449–1457.
26. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. A survey of visual transformers. *arXiv* **2021**, arXiv:2111.06091.
27. Hu, Y.; Jin, X.; Zhang, Y.; Hong, H.; Zhang, J.; He, Y.; Xue, H. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 4239–4248.
28. Wang, Y.; Ye, S.; Yu, S.; You, X. R2-Trans: Fine-Grained Visual Categorization with Redundancy Reduction. *arXiv* **2022**, arXiv:2204.10095.
29. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
30. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12595–12604.
31. Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; Belongie, S. The inaturalist species classification and detection dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8769–8778.
32. Khosla, A.; Jayadevaprakash, N.; Yao, B.; Li, F.-F. Novel dataset for fine-grained image categorization: Stanford dogs. In Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC), Colorado Springs, CO, USA, 25 June 2011.