Fine-Grained Bird Image Classification Based on Counterfactual method of Vision Transformer Model

Tianhua Chen
Beijing Technology and Business University
Yanyue Li
Beijing Technology and Business University
Qinghua Qiao (qiaoqh@casm.ac.cn)
Chinese Academy of Surveying and Mapping

Research Article

Keywords: Vision transformer, fine-grained visual classification, deep learning, bird image classification

Posted Date: March 21st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2694231/v1

Additional Declarations: No competing interests reported.

Fine-Grained Bird Image Classification Based on Counterfactual method of Vision Transformer Model

Tianhua Chen[1], Yanyue Li[1] and Qinghua Qiao[2*]

[1]School of Artificial Intelligence, Beijing Technology and Business University, Fucheng Road, Beijing, 100048, China.
[2*]Natural Resources Survey and Monitoring Research Centre, Chinese Academy of Surveying and Mapping, Wanshou Road, Beijing, 100830,China.

*Corresponding author(s). E-mail(s): qiaoqh@casm.ac.cn;
Contributing authors: cth188@sina.com; liyanyue26@163.com;

Abstract

Identifying bird targets, especially scarce birds, is essential to reflect the level of ecological protection and diversity. A fine-grained bird identification model with hierarchical feature fusion and counterfactual feature selection is proposed due to the problems of similar features, slight differences, and difficulty of identification among different birds. The model comprises a feature extraction network, a hierarchical feature fusion module, and a counterfactual feature selection module. The feature extractor extracts local area image features and global image features. The hierarchical feature fusion module fuses fine-grained information from the shallow layers of the network and semantic information from the current layer. The counterfactual feature enhancement module sifts out the distinguishing features fed to the classifier by counterfactual intervention. The experimental results show that the method can achieve 91.9% and 91.4% accuracy on two available datasets, CUB-200-2011 and NABirds, respectively, which is higher than the current mainstream fine-grained bird recognition algorithms and shows excellent classification performance.

Keywords: Vision transformer, fine-grained visual classification, deep learning, bird image classification

1 Introduction

In recent years, as the Earth's ecology continues to deteriorate, the survival of birds is more threatened than ever before. Birds are indispensable members of the ecosystem, and their population size and diversity are essential indicators of environmental protection. Birds have an extremely sensitive nature to their environment. When the number and species of birds in a region change, it often also means that there are significant changes in factors internal to the current bird habitat. There are many species of birds, and different birds have different habits. Identifying the species of birds among them not only helps to protect the diversity of bird species but also helps to understand and protect the ecosystem of this habitat [1].Because of the diversity of species, different habits, and characteristics of birds inhabiting the ecosystem, the large intra-class variation of the same species and the slight inter-class variation of other species of birds bring significant challenges to bird image recognition.

Bird image recognition belongs to the fine-grained visual classification task. General image classification represents the classification of categories with significant differences, such as the classification of cats and dogs. While fine-grained visual classification aims to classify image objects with considerable similarities, such as sub-classes in birds [2][3]. Bird fine-grained image recognition has long been considered a challenging task. Firstly, photos of the same bird taken from different angles can vary considerably in color and shape. Secondly, the color, texture, and other features of different species of birds are very similar. Thirdly, fine-grained classification of birds usually requires specialized experts to tag the image data, which makes the image data more expensive.

Benefiting from the progress of deep neural networks [4][5][6], the performance of the fine-grained visual classification model has made steady progress in recent years. The models of the fine-grained visual classification model can be classified into intensely supervised learning models and weakly supervised learning models according to the amount of supervised information. When training the intensely supervised learning model, additional manual annotation information, such as annotation frames and local region locations, is used in addition to the category labels of the images [7][8][9]. Current fine-grained visual classification tasks focus on weakly supervised learning models with only image labels to avoid the labor-intensive local annotation problem. The weakly supervised learning model does not rely on bounding box or part annotation and uses only image category labels to accomplish the work of fine-grained image classification, significantly improving the availability and scalability of fine-grained recognition.

Most previous weakly supervised learning models use a convolutional neural network as the backbone network to extract subtle features in images. However, as the number of network layers increases, such techniques bring more complicated computation, are prone to interference from noise in non-featured regions when extracting deeper features, and tend to contribute to model overfitting during training. In recent years, Vision Transformer has achieved

state-of-the-art performance in image classification. This shows that the critical feature information of images can be extracted by using a pure Transformer directly in building the network and by using the attention map of multi-head self-attention as the basis for selecting regions. We found that the direct application of the Visual Transformer to image classification has achieved some success, but the recognition accuracy for similar categories needs to be further improved. The key to distinguishing two birds that look very similar is to find those subtle differences. Therefore, we make extensive adjustments to Vision Transformer based on the characteristics of fine-grained visual classification of birds to further improve the model's performance for bird recognition.

This paper proposes a fine-grained bird recognition model with layer-level feature fusion and counterfactual feature selection. The model uses Vision Transformer as the backbone network to extract image features and fuses attention weights among transformer layers through the layer fusion module. Therefore, the fused attention weights contain the fine-grained information of the shallow layer and the semantic information of the current layer. Our bird recognition model was extensively evaluated on two famous bird visual classification benchmark datasets(CUB-200-2011 [2]and NABirds [3]). In summary, we have made several significant contributions to this work on bird identification:

The proposed layer fusion module can fuse the shallow fine-grained information between transformer layers and the semantic information of the current layer so that the features contain richer fine-grained information to ensure that the model captures more details.The proposed counterfactual feature enhancement module obtains effective attention weights by performing a counterfactual intervention on the fused attention weights. The effective attention weights select the distinguishing features and reduce the noise interference from redundant features.The proposed model achieved advanced accuracy on fine-grained classification datasets of birds, which did not require additional information and achieved 91.9% and 91.4% accuracy on two available datasets, CUB-200-2011 and NABirds, respectively.

2 Related Works

Domestic and foreign scholars have done many studies to solve the problem of fine-grained visual classification of birds. In this section, we briefly review the existing literature on fine-grained visual classification(FGVC). Current models can be roughly divided into two categories: the convolutional neural network-based FGVC model and Transformer based FGVC model.

2.1 Convolutional neural network-based FGVC model

A fine-grained visual classification model uses a convolutional neural network as a baseline [6] [10] [11]. And the baseline can be classified into intensely supervised learning models and weakly supervised learning models according

to the amount of supervised information. Strongly supervised learning models need to use additional manual annotation information, such as annotation frames and local region locations, in addition to inputting images and image-level labels to the model [12] [13]. Since such models are expensive to annotate and heavy manual labeling is not necessarily the best choice for model classification, current research focuses more on weakly supervised learning models with only image-level labeling.

Fine-grained visual classification models for weakly supervised learning can be subdivided into feature coding methods and localization classification sub-network methods. The feature encoding method learns more fine-grained features by computing higher-order information in the image and comparing the relationship between higher-order information [14] [15] [16]. Localization classification sub-network methods can be further divided into two categories, which are the discovery of noteworthy parts by means of Region Proposal Net (RPN) and attention mechanisms, respectively.

The first category of methods uses Region Proposal Net (RPN) [17] to propose bounding boxes containing distinguished regions. After obtaining the selected image regions, they are resized to a predefined size and trained again by the backbone network to get informative local features [18] [19] [20]. The second category of methods enhances the image feature map by attention, which is combined with a backbone network to enable the network to extract more discriminative features of critical parts to distinguish similar class images [21] [22] [23] [24] [25].

However, the region suggestion network generates a large number of bounding boxes during the training process, including many useless bounding boxes. It needs to filter out the critical parts from a large number of bounding boxes and then perform feature extraction. The attention mechanism must train the detection network to locate the distinguished parts after the feature extraction network. Parts worthy of model attention are then sent to the feature extraction network to learn the key features. Both approaches require a specially designed module to get discriminative regions, and these selected regions need to be forwarded through the backend again for final classification. Such complex networks tend to cause model overfitting during training.

2.2 Transformer-based FGVC model

Transformer has contributed significantly to the research of natural language processing and machine translation [26] [27] [28]. Inspired by this, many researches in recent years have tried to apply transformer to the field of computer vision. Originally, transformer was used to process the CNN's backbone network, extracting the video's sequence features [29]. Subsequently, the Transformer model was further extended to other popular computer vision tasks, such as image classification [30] [31], object detection [32] [33], and semantic segmentation [34] [35]. Recently, pure transformer models have become increasingly popular.

Vision Transformer(ViT) [36] is the first model that uses a pure transformer architecture for image classification, and it can obtain state-of-the-art image performance, so many approaches for FGVC tasks use ViT as the backbone network. FFVT [37] proposes mutual attention weight selection to efficiently guide the network to select differentiated tokens without introducing additional parameters. AF-Trans [31] exploits the attention weights in ViT and adaptively filters them based on the relative importance of the input patches. RAMS-Trans [38] uses transformer's multi-head self-attention mechanism to learn discriminative regional attention in a multi-scale cyclic manner. ViT-FOD [39] proposes to segment the information patches from two images to generate a new image and integrate the complementary information captured by the category token in different transformer layers. TransFG [30] integrates all the original attention weights into the attention mapping as a way to guide the model in selecting disparate picture regions. These methods all use the vision transformer as the backbone network and use the multi-head self-attention mechanism to search for discriminative parts of the image and finally process the features of these regions to complete the fine-grained recognition task.

The fine-grained visual classification model built on the Vision Transformer framework shows excellent classification performance. Its multi-head self-attention mechanism can effectively locate the discriminative parts of the image and obtain the global features of these parts. Inspired by the above view, our bird recognition model uses Vision Transformer as the baseline and proposes a new hierarchical feature fusion module and a counterfactual feature enhancement module to enable the model to extract more effective fine-grained features.

3 Methods

We first describe the general framework of a fine-grained bird identification model with transformer hierarchical feature fusion and counterfactual feature selection (TransHCM). The general architecture of Vision Transformer is then briefly described, demonstrating how to perform some pre-processing steps to extend it to fine-grained recognition. Lastly, we introduce our proposed the hierarchical feature fusion module and the counterfactual feature selection module.

3.1 Structure of TransHCM

An overview of the proposed TransHCM can be seen in Figure 1. Images are split into small patches and projected into the embedding space (a non-overlapping split is shown here). Patch embeddings and learnable position embeddings are used as input to the transformer encoder. The hierarchical feature fusion module will fuse shallow attention. Before the final transformer layer, the counterfactual feature enhancement module is used to pick tokens that correlate to discriminative image patches and then only use these tokens as input.

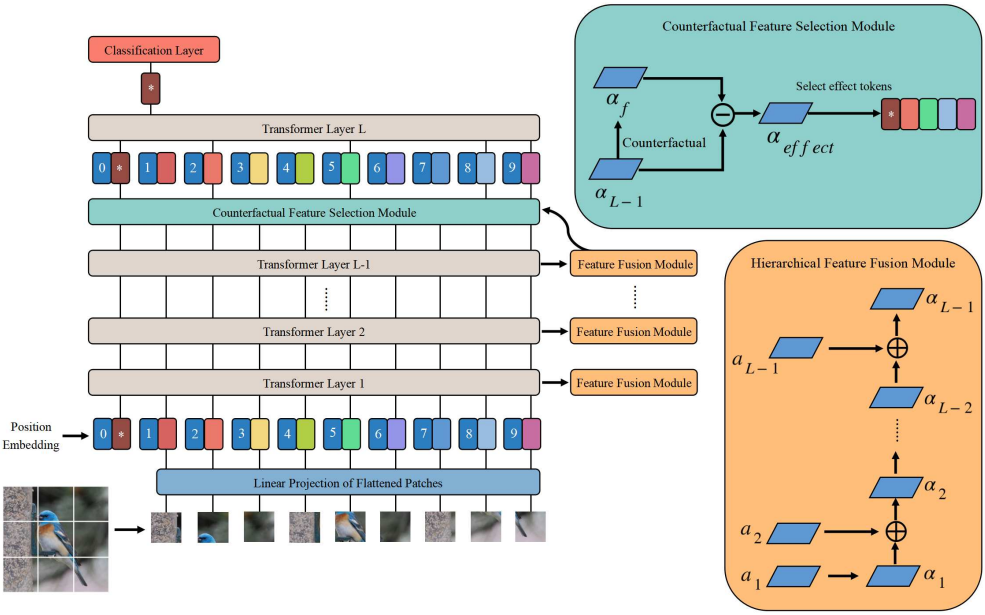

Fig. 1 The structure of TransHCM.

3.2 Vision Transformer-based robust baseline

We construct a bird recognition model based on the robust baseline of the visual transformer, following the general strong pipeline for FGVC [40] [41]. As shown in Figure 1, given an image $x \in \mathbb{R}^{H\times W\times C}$, where H,W,C are denoted as height, width, and number of channels, respectively, we divided it into N fixed-size patches then be expressed as follows: $\left\{x_p^i \mid i = 1, 2, \cdots, N\right\}$. The input sequences are supplemented with an additional learnable class embedding token marked as x_{cls}. The output class token serves as a global feature representation . By incorporating learnable location embedding, spatial information is incorporated. The input sequences fed into transformer layers can then be expressed as follows:

$$Z_0 = \left[x_{cls};\ F\left(x_p^1\right);\ F\left(x_p^2\right);\ \cdots F\left(x_p^N\right)\right] + P. \tag{1}$$

where Z_0 denotes the input sequence embedding and $P \in \mathbb{R}^{(N+1)\times D}$ denotes the location embedding. F is a linear projection mapping the patches to D dimensions. Since all transformer layers have a global field of view and no down-sampling operation, detailed feature information is saved.

3.2.1 Overlapping Patches

Purely transformer-based models, such as ViT [36], and DeiT [42], segment the images into non-overlapping patches, which lose local adjacent structures around the patches and may also cut off critical and discriminative regions.

To ensure the region integrity, we use sliding windows to create patches with overlapping pixels. Using S to represent the step size and P to represent the patch size (e.g.16), the shape of the area where two neighboring patches meet is $(P - S) \times P$. The original picture will be divided and segmented into N segments with a resolution of $H \times W$.

$$N = N_H \times N_W = \left\lfloor \frac{H + S - P}{S} \right\rfloor \times \left\lfloor \frac{W + S - P}{S} \right\rfloor . \tag{2}$$

where $\lfloor \cdot \rfloor$ is the floor function and S is less than P . N_H and N_W indicate the number of splitting areas in height and breadth, respectively. The smaller S , the more segments the picture will be divided into, which typically results to better performance but also more computation.

3.2.2 Position Embedding

Transformer compensates for the lack of location information by location embedding. The location encoding is related to the location and input scale. Since the input image resolution is different for the FGVC task and the ViT task, it is not possible to embed the preprocessing directly using the location information on ImageNet. Therefore, we compute the positional encoding by bilinear $2D$ interpolation to support image inputs of different resolutions. Similar to ViT, the positional embedding is also learnable.

3.3 Hierarchical feature fusion module

The most critical problem in fine-grained image classification of birds is to locate parts with subtle differences between similar classes accurately. For example, Figure 2 shows four different birds with a very high degree of similarity. In order to distinguish these four birds, the model needs to be able to identify tiny differences, i.e., cheeks, eyebrows, and throat color.

The deep tokens of the original ViT model have stronger semantic information, but the perception of details is poor, and it is difficult to gather the important tokens infor-mation between the transformer layers. Therefore the original ViT model is not good at extracting fine-grained features. In contrast, the features extracted by the shallow network are more similar to the input image and contain more fine-grained information, such as the color, texture, edge, and corner information of the image. To compensate for the missing shallow detail information in the high-level features, we integrate the attention weights between the layers. The attention weights of the previous $L - 1$ layers in the model are expressed as:

$$a_l = \left[a_l^0, a_l^1, a_l^2, \cdots, a_l^K\right] \quad l \in 1, 2, \cdots, L - 1. \tag{3}$$

$$a_l^i = \left[a_l^{i_0}, a_l^{i_1}, a_l^{i_2}, \cdots, \cdots, a_l^{i_N}\right] \quad i \in 0, 1, \cdots K. \tag{4}$$

where K denotes the number of self-attention heads in the multi-head self-attention mechanism in the model and l denotes the number of layers in the

Fig. 2 Subtle differences between the white-browed gypsy, yellow-bellied titmouse, yellow-browed gypsy and robin gypsy.

model. In order to make full use of the shallow detail information, the attention weights between the layers are integrated as:

$$\alpha_l = \alpha_{l-1} \times a_l \quad l \in 2, \cdots, L-1. \tag{5}$$

where α_l represents the attention weight of the l layer after fusion. while $l = 1$, $\alpha_1 = a_1$. This method is applied to all transformer layers excluding the last layer. Specifically, by fusing the attention weights of the l layer and the $l-1$ layer by matrix multiplication. The hierarchical feature fusion module fuses the attention weights among the transformer layers. Finally, it uses the fused attention weights as the attention weights of the layer $L-1$ and the input of the counterfactual feature enhancement module. Since the product of fused attention weights and feature tokens contains not only all the information between each transformer layer but also the local information lost in the shallow network, it is easier to capture fine-grained features.

3.4 Counterfactual feature selection module

Due to the slight interclass differences of birds, there are a large number of redundant tokens in many tokens of image segmentation, and these redundant tokens are not conducive to the model learning fine-grained features of birds. To remove the useless, redundant tokens, we propose a counterfactual feature selection module.

The counterfactual feature enhancement module is derived from the field of Casual Inference, and its operation is shown in Figure 3. We use the nodes

of the Causal Directed Acyclic Graph to represent the variables of the attention model, including patch features X, learned attention weights A, and final prediction results Y. The link $X \rightarrow A$ indicates that patch features are used as input to the attention model, and the corresponding at-tention graph is the output. $(X, A) \rightarrow Y$ indicates that patch features and attention weights together determine the final prediction result. The causal relationships between nodes are encoded in link, so that X is the causal parent of A and Y is the causal child of X and A.

In the directed acyclic graph with counterfactual intervention, the link $X^* \rightarrow A$ indicates that the additionally irrelevant patches of features are used as input to the at-tention model, and the corresponding attention graph A^* is output. $(X, A^*) \rightarrow Y$ denotes the predicted outcome of the counterfactual determined by the current patches feature together with additional unrelated attention weights. A measure of how much the attention weights contribute to the prediction of Y can be obtained by using the original prediction results minus the counterfactual prediction results.

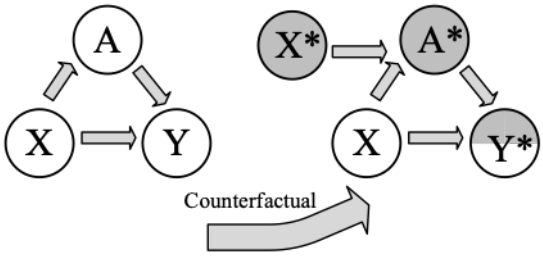

Fig. 3 The directed acyclic graphs represent structural counterfactual causal models.

Through changing the attention weights of key tokens in layer $L-1$ in the Vision Transformer model, the attention weights of classification tokens in layer L can be changed after doing multiple self-attention calculations in layer L, thus affecting the prediction results. We obtain the counterfactual attention weights α_f by randomly decreasing the values in the fused attention weight matrix α_l.The fused attention weights are subtracted from the attention weights of the counterfactuals to obtain the effective attention weights α_{effect}:

$$\alpha_{effect} = \alpha_l - \alpha_f. \tag{6}$$

Select the index $A_1, A_2, \cdots, A_K$ with the maximum weight for each of the K attention heads in α_{effect}.The critical tokens selected for extracting discriminatory tokens between levels by indexing are denoted as:

$$Z_{l-1} = \left[Z_{l-1}^{A_1}, Z_{l-1}^{A_2}, \cdots, Z_{l-1}^{A_K} \right]. \tag{7}$$

where $Z_{l-1}^{A_K}$ denotes the discriminative token extracted by the Kth index. Finally, the classification tokens are connected with valid discriminative tokens, and this process can be expressed as:

$$Z_{l-1} = \left[Z_{l-1}^{0}, Z_{l-1}^{A_1}, Z_{l-1}^{A_2}, \cdots, Z_{l-1}^{A_K} \right]. \tag{8}$$

where Z_{l-1}^{0} denotes a classification token. All the tokens are input to the final trans-former layer, and through the multi-head self-attention mechanism, tokens information interacts with each other. Finally, the class token containing valid attention is input to the Multi-layer Perceptron(MLP) head to get the prediction result of useful features.

Since the differences between subcategories are tiny and straightforward cross-entropy loss is insufficient to completely oversee feature learning, we use a contrast loss function in addition to the cross-entropy loss function to calculate the differences between predicted features and actual category features during the training of the model. When the prediction result is accurate, it maximizes the feature similarity between the prediction result and the actual category label and reduces the loss. On the contrary, when the prediction result is wrong, it can minimize the feature similarity between the pre-diction result and the actual category labels and increase the loss. The contrastive loss function is expressed as:

$$L_{con} = \frac{1}{B^2} \sum_{i}^{B} \Big(\sum_{j:y=y_{label}}^{B} (1 - Cosine(z_i, z_{label})) + \sum_{j:y \neq y_{label}}^{B} \max(Cosine(z_i, z_{label}) - m, 0) \Big). \tag{9}$$

where B is the number of samples and $Cosine(z_i, z_{label})$ denotes the cosine similarity between the predicted features and the actual class features. m is the set threshold, which affects the value of the contrast loss only if the cosine similarity is greater than m.

Our bird recognition model uses Vision Transformer as the feature extractor, the hierarchical feature fusion module to make the model extract fine-grained features, and the counterfactual feature enhancement module to select tokens that are useful for bird recognition. The valid tokens selected in the final transformer layer interact with the class token. Finally, the class token is sent to the MLP head to get the final classification result. The proposed hierarchical feature fusion module and counterfactual feature enhancement module can preserve the global attention information and get more accurate attention parts, forcing the last Transformer Layer to pay attention to the sub The proposed hierarchical feature fusion module and counterfactual feature enhancement module can retain the global attention information and obtain more accurate attention parts, requiring the last Transformer Layer to concentrate on the subtle distinctions between various sub-categories, while ignoring the regions that are not discriminative, effectively improving the model's ability to capture key tokens.

4 Experiments

We evaluate the effectiveness of our proposed fine-grained bird identification model with hierarchical feature fusion and counterfactual feature selection on bird identification tasks. The detailed experimental setup, including the dataset and training hyperpa-rameters, is first presented. Then a quantitative analysis is performed, and finally, an ablation study is conducted. We also provide detailed analysis and visualization findings to demonstrate the model's interpretability.

4.1 Datasets and implementation details

We evaluated our proposed TransHCM on two widely used bird identification datasets, CUB-200-2011 [2]and NABirds [3], for fine-grained bird classification. The CUB200-2011 dataset has 200 bird categories, including 5994 training images and 5794 test data. Each category contains about 30 training data. NABirds has 555 bird species, 23929 training images, and 24633 test images. Both datasets provide image-level annotations and key point locations, but only image-level annotations are used in this paper.

In the experiment, the network is trained by loading the weights of the official ViT-B_16 model pre-trained on ImageNet21k, resizing the original image to 448448, and segmenting the image to make it 1616 patches. The training period for each dataset is set to 90. In the training phase, data enhancement is performed by Random Crop, Random HorizontalFlip, and Random Gaussian Blur, while in the testing phase, Center Crop is used. In the training phase, stochastic gradient descent (SGD) is used to optimize the network with a momentum value of 0.9, and the initial learning rate is set to 0.02. The learning rate is adjusted in an orderly manner using cosine annealing. The batch size was set to 16. The Pytorch framework was used as the experimental platform to accelerate the training in FP16 data format with the APEX toolkit.

4.2 Quantitative analysis

We compared our proposed method TransHCM with the recent work on the fi-ne-grained dataset described above. Table 1 displays the findings of the CUB-200-2011 trials. According to the findings, our approach outperforms all prior methods on the CUB dataset. Compared to the best model CAP so far, our TransHCM improves by 0.1% in Top-1 accuracy and 1.6% compared to our baseline Vision Transformer. We observed that most methods obtained good results by having different branches to get multiple backbones or using a rather deep CNN structure to extract better features. For example, CAP uses Xception as the backbone, and the network employs various branches and can achieve the highest previous accuracy. However, this significantly increases the com-putational complexity, and such a model hinders practical usability. In contrast, our TransHCM maintains simplicity and achieves the highest accuracy rate available.

Table 1 Results on CUB-200-2011 with different pre-trained models.

Method	Backbone	Pretrain	CUB-200-2011
Cross-X	ResNet-50	ImageNet-1k	87.7
FixSENet	SENet-154	ImageNet-1k	88.7
DSTL	Xception-v3	iNat17	89.3
API-Net	DenseNet-161	ImageNet-1k	90.0
ViT	ViT-B_16	ImageNet-21k	90.3
CPM	ResNet-50	ImageNet-1k	90.4
CAL	ResNet101	ImageNet-1k	90.6
TransFG	ViT-B_16	ImageNet-21k	91.7
CAP	Xception	ImageNet-1k	91.8
TransHCM	ViT-B_16	ImageNet-21k	**91.9**

Table 2 Results on NABirds with different pre-trained models.

Method	Backbone	Pretrain	NABirds
Cross-X	ResNet-50	ImageNet-1k	86.2
DSTL	Xception-v3	iNat17	87.9
API-Net	DenseNet-161	ImageNet-1k	88.1
MGE-CNN	ResNet-101	ImageNet-1k	88.6
FixSENet	SENet-154	ImageNet-1k	89.2
ViT	ViT-B_16	ImageNet-21k	89.9
TransFG	ViT-B_16	ImageNet-21k	90.8
CAP	Xception	ImageNet-1k	91.0
TransHCM	ViT-B_16	ImageNet-21k	**91.4**

NABirds is a much bigger avian dataset with 355 bird groups, more subtle variations between categories, and substantial differences in samples from the same category, making fine-grained visual categorization more difficult. We show our results in Table 2. Compared with the best model CAP to date, our model improves by 0.4% on the Top-1 accuracy metric. Our model is able to achieve 91.4% accuracy in identifying NABirds, which is a 1.5% improvement compared to our base framework Vision Transformer. It indicates that the backbone Xception of the CAP model can extract fine-grained bird features well, but the backbone ViT-B_16 of our model extracts bird fine-grained features more accurately in more species of bird datasets.

Our proposed fine-grained bird recognition model with hierarchical feature fusion and counterfactual feature selection can effectively extract important fine-grained image features of birds and achieve advanced accuracy in both standard bird datasets.

4.3 Ablation studies

We performed an ablation study on TransHCM to analyze the proposed method's impact on birds' fine-grained classification accuracy. All ablation

experiments were conducted on the CUB-200-2011 dataset, and the same phenomenon was observed on the NABirds dataset.

4.3.1 The impact of integrating different layers of characteristics

Table 3 Ablation study of fusing different layers of features on the CUB-200-2011 dataset.

Method	Layer Num	CUB-200-2011
TransHCM	3	91.6
TransHCM	5	91.7
TransHCM	7	91.7
TransHCM	9	91.8
TransHCM	11	91.9

Table 3 shows the effect of fusing the attention weights extracted from Transformer layers with different numbers of layers on the final classification effect. The fused attention weights are passed through the counterfactual feature enhancement module to obtain effective attention weights and effective class token features as input to the MLP head to achieve the classification task. As can be seen from Table 3, fusing 11 layers, the model learns more fine-grained features and does a better job of classifying CUB-200-2011 images.

4.3.2 Impact of counterfactual feature enhancement module

Table 4 Ablation study of applying counterfactual feature enhancement module on CUB-200-2011 dataset.

Method	CFM	Accuracy
ViT	-	90.3
ViT	✓	90.8
TransHCM	-	91.7
TransHCM	✓	91.9

As shown in Table 4, the performance of the ViT model was improved from 90.3% to 90.8% by applying the Counterfactual Feature Enhancement Module (CFM) to constrain the attention calculation and adding the Counterfactual Feature Enhancement Module to the original ViT model. The performance of the TransHCM model improved from 91.7% to 91.9% by adding the counterfactual Feature Enhancement Module to the TransHCM model incorporating the first 11 layers of attention weights. We think this is because we choose the

most discriminative tokens as input, explicitly discarding some useless tokens and forcing the network to learn from the essential portions.

4.3.3 Contrast the impact of losses

Table 5 Ablation study of contrast loss in the CUB-200-2011 dataset.

Method	Contrastive Loss	Accuracy
ViT	-	90.3
ViT	✓	90.7
TransHCM	-	91.6
TransHCM	✓	91.9

Table 5 shows the performance comparison of ViT and TransHCM frameworks with and without using the contrast loss function. We observed that both models were able to obtain a significant performance gain using the contrast loss function. In terms of quantity, it improved the accuracy of ViT from 90.3% to 90.7 % and TransHCM's accuracy from 91.6 % to 91.9%. We think this is because the contrast loss can effectively expand the repre-sentation distance between different categories while decreasing the distance between the same categories, making the resulting feature matrix more discriminative. We show the visualization finding of the suggested TransHCM on the CUB-2011-200 dataset in Figure 4. The first and third rows are the original images, and the second and fourth rows are their visualization results. Our model can filter out the redundant noise well and accurately locate the birds' critical parts.

5 Conclusion

In this work, we propose a hierarchical feature fusion module and a counterfactual feature enhancement module to efficiently guide the model in learning the fine-grained features of birds. We design a new fine-grained recognition framework, TransHCM, to capture attentional relationships among patches using a multi-head self-attention mechanism and integrate attentional weights with shallow fine-grained information and deep semantic information through a fusion of cascading attentional weights, enabling the model to learn rich fine-grained features among cascades effectively. We also propose the counterfactual feature enhancement module by analyzing the effect of factual attention and counterfactual attention on the final prediction results. This module introduces counterfactual attention weights and subtracts the counterfactual attention weights from the fused attention weights to get the effective attention weights of the model so as to obtain more effective patches and selects the patches with the largest attention weights as the input of the last transformer layer, which reduces the interference of redundant noise to the network. To encourage the

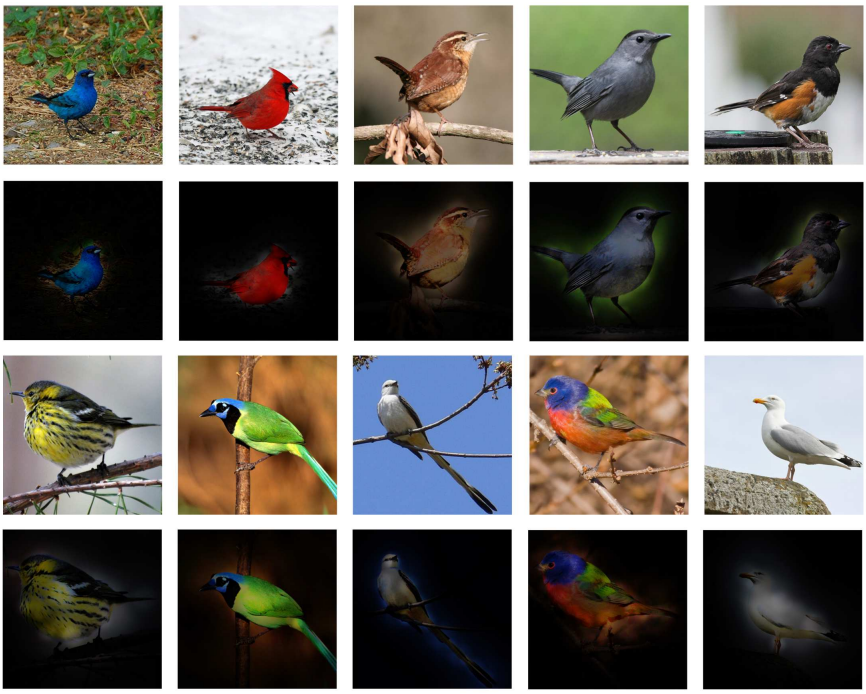

Fig. 4 Visualization results of TransHCM on the CUB-2011-200 dataset.

network to learn more effective patches of features, we also used a contrast loss function to maximize the difference between different samples while minimizing the difference between the same samples. TransHCM achieves satis-factory results that illustrate the great potential of Vision Transformer-based models for fine-grained image classification tasks of birds. However, some potential problems will need to be resolved in the future. First, our model only outputs single-scale feature representations and cannot thus handle mul-ti-scale variations, and some feature information may be lost in the tokens that the model finally feeds into the classification network. Second, our model does not enlarge the es-sential parts to extract more subtle features. In view of this, we will model multi-scale changes and integrate multi-scale hierarchical features in our future work. And construct a parts detector for clipping and amplifying the relevant parts and connecting the features of these parts for recognition, reducing the loss of fine-grained features.

Acknowledgments. The authors are grateful to the reviewers for their valuable suggestions.

Declarations

Ethical Approval. Not applicable.

Competing interests. The authors declare no conflict of interest.

Authors' contributions. Tianhua Chen, Yanyue Li and Qinghua Qiao contributed to the conception of the study; Tianhua Chen and Yanyue Li performed the experiment; Tianhua Chen and Yanyue Li contributed significantly to analysis and manuscript preparation; Tianhua Chen, Yanyue Li and Qinghua Qiao performed the data analyses and wrote the manuscript; Tianhua Chen, Yanyue Li and Qinghua Qiao helped perform the analysis with constructive discussions.

Funding. Not applicable.

Availability of data and materials. Publicly available datasets were analyzed in this study. The CUB-200-2011 dataset is available at http://www.vision.caltech.edu/datasets/cub_200_2011/, accessed in 2011. The NABirds dataset is available at https://dl.allaboutbirds.org/nabirds, accessed in 2015.

References

[1] Socolar, J.B., Gilroy, J.J., Kunin, W.E., Edwards, D.P.: How should beta-diversity inform biodiversity conservation. Trends in ecology & evolution **31**(1), 67–80 (2016)

[2] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)

[3] Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: 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, pp. 595–604

[4] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017)

[5] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[6] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778

[7] Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based r-cnns for fine-grained category detection. In: Computer VisionECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, pp. 834–849. Springer

[8] Huang, S., Xu, Z., Tao, D., Zhang, Y.: Part-stacked cnn for fine-grained visual categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1173–1182

[9] He, K., Gkioxari, G., Dollr, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969

[10] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR

[11] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708

[12] Branson, S., Van Horn, G., Belongie, S., Perona, P.: Bird species categorization using pose normalized deep convolutional nets. arXiv preprint arXiv:1406.2952 (2014)

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

[14] Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 317–326

[15] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)

[16] Cui, Y., Zhou, F., Wang, J., Liu, X., Lin, Y., Belongie, S.: Kernel pooling for convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2930

[17] Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448

[18] Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 420–435

[19] Liu, C., Xie, H., Zha, Z.J., Ma, L., Zhang, Y.: Filtration and distillation: Enhancing region attention for fine-grained visual categorization. Proceedings of the AAAI Conference on Artificial Intelligence **34**(7), 11555–11562 (2020)

[20] 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, pp. 3034–3043

[21] Rao, Y., Chen, G., Lu, J., Zhou, J.: Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1025–1034

[22] Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5209–5217

[23] Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 805–821

[24] Zhuang, P., Wang, Y., Qiao, Y.: Learning attentive pairwise interaction for fine-grained classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13130–13137

[25] Hu, T., Qi, H., Huang, Q., Lu, Y.: See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:1901.09891 (2019)

[26] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019)

[27] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, ., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[29] Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 244–253

[30] He, J., Chen, J.-N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., Wang, C.: Transfg: A transformer architecture for fine-grained recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 852–860

[31] 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: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3234–3238. IEEE

[32] Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2906–2917

[33] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)

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

[35] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

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

[37] Wang, J., Yu, X., Gao, Y.: Feature fusion vision transformer for fine-grained visual categorization. arXiv preprint arXiv:2107.02341 (2021)

[38] 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, pp. 4239–4248

[39] Zhang, Z.-C., Chen, Z.-D., Wang, Y., Luo, X., Xu, X.-S.: Vit-fod: A vision transformer based fine-grained object discriminator. arXiv preprint arXiv:2203.12816 (2022)

[40] Lin, T.-Y., Dollr, 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, pp. 2117–2125

[41] Korsch, D., Bodesheim, P., Denzler, J.: End-to-end learning of fisher

vector encodings for part features in fine-grained recognition. In: Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021, Bonn, Germany, September 28October 1, 2021, Proceedings, pp. 142–158. Springer

[42] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jgou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. PMLR