Neurocomputing 492 (2022) 137–149

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Transformer with peak suppression and knowledge guidance for fine-grained image recognition

Xinda Liu [a], Lili Wang [a,*], Xiaoguang Han [b]

[a] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[b] Shenzhen Research Institute of Big Data, Shenzhen, China

ARTICLE INFO

Article history:
Received 6 November 2021
Revised 5 February 2022
Accepted 3 April 2022
Available online 6 April 2022
Communicated by Zidong Wang

Keywords:
Fine-grained image recognition
Food recognition
Knowledge guidance
Peak suppression
Vision transformer

ABSTRACT

Fine-grained image recognition is challenging because discriminative clues are usually fragmented, whether from a single image or multiple images. Despite their significant improvements, the majority of existing methods still focus on the most discriminative parts from a single image, ignoring informative details in other regions and lacking consideration of clues from other associated images. In this paper, we analyze the difficulties of fine-grained image recognition from a new perspective and propose a transformer architecture with the peak suppression module and knowledge guidance module, which respects the diversification of discriminative features in a single image and the aggregation of discriminative clues among multiple images. Specifically, the peak suppression module first utilizes a linear projection to convert the input image into sequential tokens. It then blocks the token based on the attention response generated by the transformer encoder. This module penalizes the attention to the most discriminative parts in the feature learning process, therefore, enhancing the information exploitation of the neglected regions. The knowledge guidance module compares the image-based representation generated from the peak suppression module with the learnable knowledge embedding set to obtain the knowledge response coefficients. Afterwards, it formalizes the knowledge learning as a classification problem using response coefficients as the classification scores. Knowledge embeddings and image-based representations are updated during training simultaneously so that the knowledge embedding includes a large number of discriminative clues for different images of the same category. Finally, we incorporate the acquired knowledge embeddings into the image-based representations as comprehensive representations, leading to significantly higher recognition performance. Extensive evaluations on the six popular datasets demonstrate the advantage of the proposed method in performance. The source code and models will be available online after the acceptance of the paper.

1. Introduction

Aiming to distinguish the objects belonging to multiple subcategories of the same meta-category, fine-grained image recognition has been one of the most fundamental problems in the computer vision and multimedia communities [1–5]. It is essential for a wide range of downstream applications such as rich image captioning [6], image generation [7], machine teaching [8], fine-grained image retrieval [9], food recognition [10,11], and food recommendation [12].

Fine-grained image recognition is challenging due to subtle inter-class differences and significant intra-class variances. Most existing methods only consider the problem of fine-grained recognition from the perspective of obtaining the discriminative characteristics of a single image but ignore the clues provided by multiple images. In order to take the clues of multiple images into consideration, we try to explain the difficulty of fine-grained image recognition with a new perspective and attribute it to fragmented discriminative clues.

The fragmentation here has two implications: (1) From the perspective of a single image, discriminative clues appear in different local areas since the inter-class differences could be subtle; As shown in Fig. 1 (a) and (b), the long-tailed jaeger is similar to the black tern overall, but the beak of long-tailed jaeger is curved, and the beak of the black tern is straight. This kind of discriminative part is usually tiny and distributed in different image regions, as shown in Fig. 1 (c). (2) From the perspective of multiple images, each image contains only a part of the discriminative information about the category depending on different poses, scales, and rotations, due to significant intra-class variances. As shown in Fig. 1 (b)

* Corresponding author.

https://doi.org/10.1016/j.neucom.2022.04.037
0925-2312/

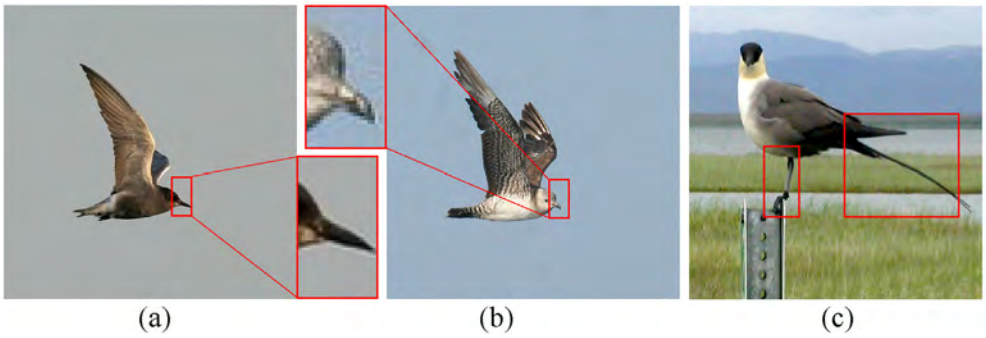

Fig. 1. Some fine-grained bird images sampling from the CUB-200-2011 dataset. (a) is Black Tern, both (b) and (c) are Long-tailed Jaeger. Discriminative parts are annotated by red boxes. The details in the red boxes of (a) and (b) are magnified next to the image. Discriminative clues are distributed in multiple regions of the image, and the clues of a single image are usually incomplete.

and (c), these two pictures are long-tailed jaeger, the beak of the bird in (c) is difficult to distinguish, and there are no bird claws in (b). Therefore, the discriminative information contained in each image is incomplete.

To find and aggregate fragmented clues is the key to fine-grained image recognition. Despite their impressive results, existing methods usually consider fine-grained image recognition only from few regions, ignoring many informative details in other regions and other associated images. For instance, if the beak of a specific bird is very different from other birds, the model may pay much attention to the beak of the bird while ignoring the claws and tail of the bird. When this happens, the model could easily make mistakes when the beak of the bird is not visible.

Given this challenge, we propose a Transformer with Peak Suppression and Knowledge Guidance (TPSKG) for fine-grained image recognition. The proposed Peak Suppression (PS) module uses the transformer architecture to integrate the local information and explores a training routine to increase the diversity of discriminative features. This PS module is designed to obtain as many discriminative clues as possible from a single image. Simultaneously the proposed Knowledge Guidance (KG) module incorporates the learnable knowledge embedding into the image-based representation for a comprehensive representation. This KG module is used to aggregate discriminative information from multiple images.

Specifically, we are inspired by Vision Transformer (ViT) [13] and use a transformer architecture to tackle the fine-grained image recognition problems. The input image is reshaped into a patch sequence without overlap and then linearly mapped to the sequential tokens. The transformer encoder uses the self-attention mechanism to integrate the information of the different tokens to obtain a global representation. Instead of integrating all the token information like the original ViT, we deliberately remove the most discriminative token based on the value of the attention weight map in training to penalize strongly discriminative learning and enforce the network to pay attention to other neglected informative areas for keeping the fine-grained representation diversity.

After that, we use a knowledge embedding set to explicitly express the discriminative clues of the same category from different images and formalize the learning of knowledge embedding as a classification task. The knowledge guidance module measures the similarity of the knowledge embeddings and image-based representations generated from the peak suppression module to obtain the knowledge response coefficients. We use the knowledge response coefficients as the classification scores directly and use the category label as ground truth to supervise the knowledge learning. The image-based representations become more discriminative through the joint training of fine-grained classification and knowledge learning tasks, and the knowledge embeddings are also concurrently updated through iterations. The learning procedure of knowledge embeddings covers the entire training dataset, therefore these embeddings become the comprehensive representations containing various subtle and slight characteristics of all categories. Finally, we obtain the knowledge-based representations computed from the knowledge embedding set along with the knowledge response coefficients, and inject them into the image-based representations. The proposed knowledge embedding learning and exploitation lead to a significant boost for recognition performance.

To verify the effectiveness of our method, we conduct extensive experiments on the six popular benchmarks for the fine-grained image recognition task. Quantitative experimental results demonstrate that the proposed method can achieve competitive performance compared to the state-of-the-art approaches. As shown in Fig. 2, qualitative experimental results demonstrate the advantages of our method in covering more informative areas and increasing the diversity of expression at the same time. The quantitative analysis and visualization of knowledge embedding also illustrate the effectiveness of category-related knowledge embedding learning.

In summary, we make the following main contributions:

(1) We provide a new perspective that the difficulty of fine-grained recognition lies in fragmented discriminative clues. This perspective helps consider not only multiple regions from a single image but also multiple images.
(2) We propose a vision transformer architecture with peak suppression and knowledge guidance for the fine-grained image recognition task. Peak suppression effectively increases the diversity of image representations via aggregating the local features from multiple regions from a single image. Knowledge guidance optimizes the final representations with the knowledge embeddings learning from multiple images.
(3) We formalize the knowledge learning as a classification problem and directly use the similarity between knowledge embeddings and image-based representations as the classification score to update the knowledge embeddings related to the category.
(4) We conduct extensive quantitative and qualitative experiments to demonstrate the effectiveness of the proposed method, which achieves competitive performance compared to the state-of-the-art approaches on six public datasets.

The rest of this paper is organized as follows. Section 2 reviews the related works. Section 3 elaborates on the proposed framework. Experimental results and analysis are reported in Section 4. Finally, we conclude the paper in Section 5.

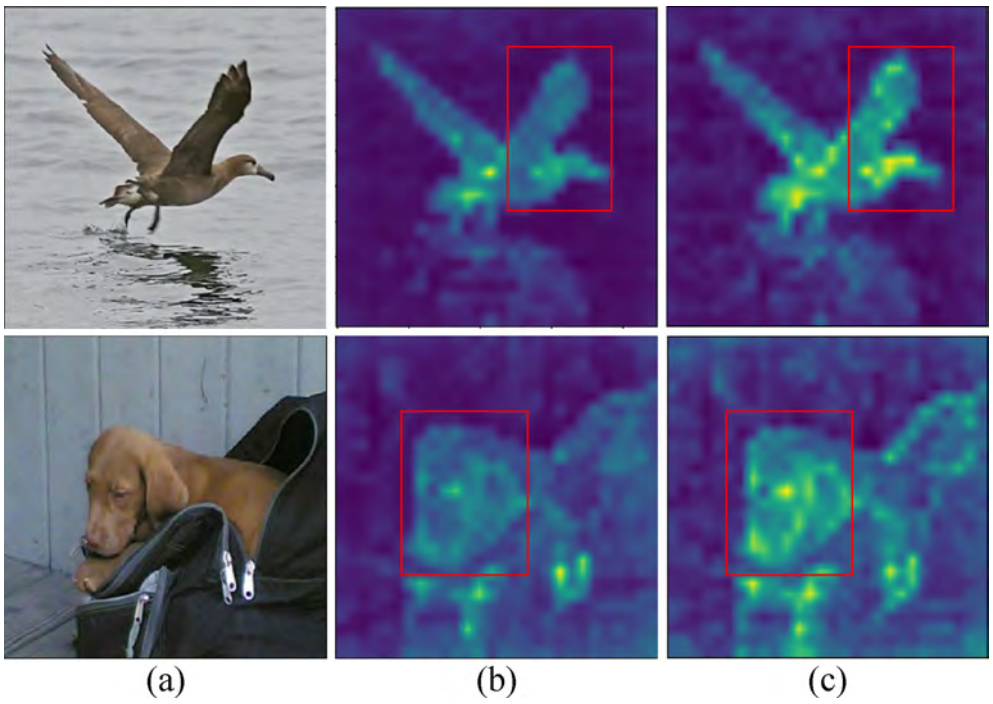

Fig. 2. The effects of the proposed approach for some samples from the CUB-200-2011 and Stanford Dogs datasets. (a) is the original fine-grained image, (c) and (b) are the attention weights obtained from the vision transformer with and without the proposed method (TPSKG). The parts of the proposed method that are significantly different from the original method are annotated by red boxes.

2. Related works

This section introduces the most related researches into the following categories: the fine-grained image recognition task and the vision transformer architecture.

2.1. Fine-grained image recognition

There are two prevailing paradigms in the current research in fine-grained image recognition. One is the local identification, and the other is the global discrimination.

Local-identification approaches focus on locating the discriminative semantic parts of fine-grained objects to identify the subtle differences among different object categories and construct mid-level representations corresponding to these parts for the final classification. Early works [14,15] used strong supervised mechanisms with part bounding box annotations to learn localizing the discriminative parts. However, the part annotation is time-consuming. Recent researches [16–19] focused on weakly supervised recognition methods with only image-level labels to obtain accurate part localization to solve this problem. Some patch-based methods [20–22] first initialize abundant region proposals and select the discriminative parts based on a specific strategy. There are also attention-based ways to localize the corresponding high areas related to the image label, such as [23–25,10,26].

Global-discrimination approaches generally learn the embeddings using a specific distance metric so that samples from the same category can be pulled close to each other while samples from different categories are pushed apart. For example, a bilinear model is used in [27] to learn the interacted feature of two independent CNNs, which achieves remarkable fine-grained recognition performance. However, the exceptionally high dimensionality of bilinear features still makes it impractical for realistic applications. Chen et al. [28] enforced the classification network to pay more attention to discriminative regions for spotting the differences by destructing and reconstructing the input image. Sun et al. [29] masked the most salient features for the input images to force the network to use more subtle clues for its correct classification. Zhuang et al.[30] learned a mutual feature vector to capture semantic differences in the input image pair.

Unlike the methods described above, we consider the discriminative but not the most significant part of a single image, but also emphasize the discriminative information aggregation in different images. Hence, we propose a vision transformer with peak suppression and knowledge guidance, which can effectively increase the richness of fine-grained representations in a local area and effectively aggregate patch features. Simultaneously, it emphasizes the learning and utilization of the knowledge of distinguishing characteristics between different image samples.

2.2. Vision transformer

The transformer architecture by [31] is proposed to deal with the sequential data in the field of natural language processing [32,33]. Inspired by the breakthroughs of transformer architectures in the field of natural language processing, researchers have recently applied transformer to computer vision tasks, such as image recognition [13,34], object detection [35,36], segmentation [37], image super-resolution [38]. For example, Cordonnier et al. [39] proved that a multi-head self-attention layer with a sufficient number of heads is at least as expressive as any convolutional layers. They extracted patches from the input image and applied full self-attention on top. iGPT [40] applies transformers to image pixels after reducing the image resolution and color space. It is worth noting that the Vision Transformer (ViT) [13] is a pure transformer that performs well on the image classification task when applied directly to the sequences of image patches. Based on ViT, Touvron et al.[34] proposed a data-efficient image transformer (DeiT) for the fine-grained visual categorization and achieved competitive performance.

To sum up, the vision transformer maps group pixels into a small number of visual tokens, representing a semantic concept in the image. These visual tokens are used directly for image classification, with the transformers being used to model the relationships among tokens. Our work is inspired by ViT and adopts the same method as ViT to build a transformer. Despite the efficiency of iGPT, ViT, and DeiT, these works only fine-tune the model on the fine-grained datasets directly to evaluate the effectiveness of the model for transfer learning and ignore the characteristics of the fine-grained image recognition task. Different from the above methods, we consider the characteristics of the fine-grained image and focus on the specific recognition task.

3. Framework

This section introduces the proposed framework, which is a transformer architecture for the fine-grained recognition task. As shown in Fig. 3, this framework mainly consists of two components, namely Peak Suppression (PS) and Knowledge Guidance (KG). PS takes images as input and outputs the suppressed image-based representations to the KG module. The KG module takes the image-based representations and learns the knowledge embeddings, finally uses the fusion representations for the recognition task. Section 3.1 introduces PS and Section 3.2 details KG.

3.1. Peak suppression

Discriminative regions in fine-grained images are usually scattered over multiple regions of the image. Many existing methods focus on the most discriminative regions and ignore other discriminative regions. To address this issue, we force the network to pay less attention to the most discriminative regions and increase the information utilization of neglected regions, thereby increasing the diversity of obtained features. Inspired by the effectiveness of the diversification block on convolutional neural networks, we proposed the peak suppression module on the transformer architecture to pay more attention to the other informative parts and obtain more diverse expressions by suppressing the most discriminative regions. Different from the CNNs-based method of directly operating feature maps using the category-specific activation maps, the transformer-based method cannot achieve the goal by directly removing the most significant corresponding token in the last layer because all tokens have interacted during the feedforward process in multi-head attention layers. Therefore, we can only use the attention map to backtrack to the input image space and then mask salient image regions in the image space.

Formally, we follow the settings of [13] and use the ViT as the backbone. Let $x \in \mathbb{R}^{H\times W\times C}$ denotes a given training image where (H, W) is the resolution of the image, C is the number of channels. The image x is reshaped into a sequence of flattened 2D patches $x_p \in \mathbb{R}^{N\times P^2\times C}$, the resolution of each image patch is (P, P), and $N = HW/P^2$ is the resulting number of patches. These patches are converted to D dimensions embedding $x_pE \in \mathbb{R}^{N\times D}$ as input tokens through a trainable linear projection. Attaching the learnable embedding class token z_0^0, there are a total of $N + 1$ tokens. Position embeddings E_{pos} are added to the patch embeddings $E_{pos} \in \mathbb{R}^{(N+1)\times D}$ to retain the positional information. The transformer encoder takes the z_0 as input and outputs z_L and the attention weight M_L, where L means the transformer encoder is composed of a stack of L identical layers. Each layer consists of multi-head self-attention (MSA) and multilayer perceptron (MLP) blocks. Layer Normalization (LN) is applied before every block and residual connections after every block.

$$\begin{aligned} z_0 &= \left[x_{class}; x_p^1E; x_p^2E; \ldots; x_p^NE\right] + E_{pos}, \\ z_l' &= MSA(LN(z_{l-1})) + z_{l-1}, l = 1 \ldots L, \\ z_l &= MLP\left(LN(z_l')\right) + z_l', l = 1 \ldots L. \end{aligned} \tag{1}$$

We use the Attention Rollout technique [41] to acquire the attention map from the output token to the input space. Given a transformer with L layers, we need to flow the attention from all positions in the final layer L to all positions in layer 1. At every transformer layer, we average attention weights at each layer over all heads and get the weight matrix M_l that defines the attention value flows from all tokens in the previous layer to all tokens in the l layer. Considering that there are residual connections in the backbone, we deal with them by adding the identity matrix I to the attention matrices and re-normalize the attention weights to keep the total attention in the range of 0 to 1. Finally, we recursively multiply the weight matrices of all layers.

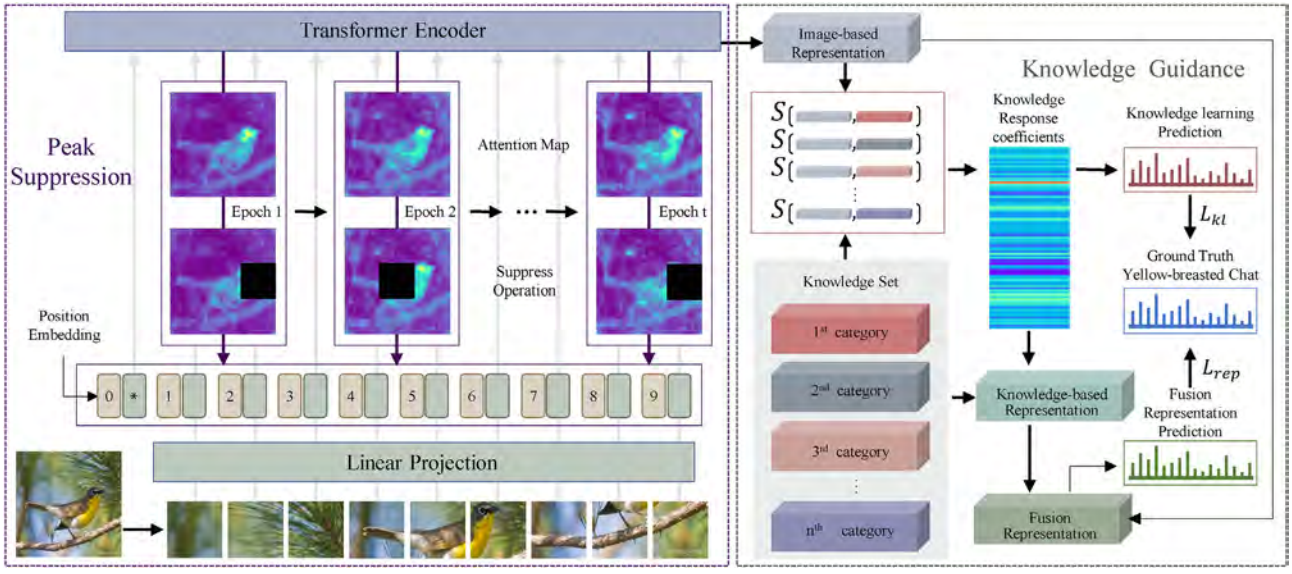

Fig. 3. Overview of our framework. Here we visualize the case of peak suppression and knowledge guidance given a training batch with a image and its corresponding label Yellow-breasted Chat. $S(\cdot)$ means the similarity function. Only the presentation label is used to predict in testing.

$$\widetilde{M}_l = \begin{cases} (M_l + I)\widetilde{M}_{l-1} & \text{if } l > 1, \\ M_l + I & \text{if } l = 1. \end{cases} \tag{2}$$

In this equation, $\widetilde{M}_l$ is attention rollout of the l layer, M_l is raw attention of the l layer, and the multiplication operation is matrix multiplication. The $\widetilde{M}_L$ illustrates the mixing of attention among tokens across all layers.

Let $B \in \mathbb{R}^{N+1}$ denote the binary suppressing mask for the input tokens. Each element in mask B is in the domain $\{0, 1\}$, where 0 indicates the corresponding location will be suppressed while 1 means that no suppression will take place. Note that the B^0 is always 1 because it corresponds to the class token.

After obtaining the attention map $\widetilde{M}_L \in \mathbb{R}^N$, we compute the B by traversing the entire attention map and finding the position of the largest response.

$$B^i = \begin{cases} 0 & \text{if } \widetilde{M}_L^{i-1} = max\left(\widetilde{M}_L\right) \\ 1 & otherwise. \end{cases} \tag{3}$$

where $i \in \{1, 2, \ldots, N\}$. In order to remove the influence of peaky token, we remove it from the forward process.

$$\hat{z}_0 = \left[x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] * B + E_{pos}, \tag{4}$$

where $*$ denotes element-wise multiplication. After the forward process, the transformer encoder outputs y as the image-based representation.

$$\begin{aligned} \hat{z}'_l &= MSA(LN(\hat{z}_{l-1})) + \hat{z}_{l-1}, l = 1 \ldots L, \\ \hat{z}_l &= MLP(LN(\hat{z}'_l)) + \hat{z}'_l, l = 1 \ldots L. \\ y &= LN(\hat{z}_L^0), \end{aligned} \tag{5}$$

where $\hat{z}_L^0$ denotes the class token vector of the output of L layer transformer encoder after peak suppression. We implement the remove of Eq. (4) by setting $-\infty$ to the suppressed token vectors. After the softmax layer of Eq. (5), the responses of these tokens will be close to 0.

To sum up, we suppress the peaky token based on the attention maps in the training phase. By suppressing the tokens, the network is forced to find the other informative regions instead of the most discriminative regions in the image. Although the proposed PS module increases the computational complexity in the training phase because of the double feed-forward, it does not increase the computational complexity at all in the inference phase. At the same time, the increase of the feature diversity can improve the performance of the network in the inference phase.

3.2. Knowledge guidance

After obtaining the diversified discriminative clues of a single image, the knowledge guidance module fuses the information of multiple images to get a more comprehensive feature representation. The knowledge guidance module first learns the knowledge embeddings related to the category. To this end, we propose a novel learning method that abstracts the learning of knowledge embeddings as a classification problem and directly uses the similarity between knowledge embeddings and image-based representations as the classification score. Subsequently, the knowledge guidance module injects the obtained knowledge embeddings into the image-based representations to get a more comprehensive expression. Therefore, the knowledge guidance module contains two tasks, one is knowledge learning, and the other is knowledge exploiting. We first introduce the knowledge learning task.

3.2.1. Knowledge learning

To enforce the networks to learn the knowledge embedding for each category, we treat the knowledge learning task as a classification task. Considering the multi-class fine-grained image recognition task, let $\mathscr{F} = \{f^g\}_{g=1}^G$ denotes the fine-grained label set containing all G fine-grained labels and $\mathscr{X} = \{x^j\}_{j=1}^J$ denotes the image training dataset containing the total J images. Through the peak suppression module of transformer encoder, we can acquire the representation set $\mathscr{Y} = \{y^j\}_{j=1}^J, y^j \in \mathbb{R}^D$ from the training dataset. Randomly initializing a knowledge embedding set of the D-dimension knowledge embedding $\mathscr{K} = \{k^g\}_{g=1}^G, k^g \in \mathbb{R}^D$, each k^g means the knowledge embedding of the category f^g.

Given a image x^j with the corresponding ground-truth fine-grained label f^g, the transformer encoder outputs its representation $y^j \in \mathbb{R}^D$. The knowledge embedding set tries to distinguish which category this representation y^j belongs to by judging the similarity between this representation y^j and every knowledge embedding k^g in the $\mathscr{K}$. We obtain a knowledge response coefficients $r^j \in \mathbb{R}^G$.

$$r^j = Softmax(S(y^j, \mathscr{K})), \tag{6}$$

where $S(\cdot)$ is a similarity function. We have tried a variety of methods to calculate similarity, such as the neural networks and element-wise multiplication, and the results are not significantly different. In addition, the computational complexity of element-wise multiplication is much smaller. Therefore, without loss of generality, the element-wise multiplication is adopted in our experiments.

We then directly convert the knowledge response coefficients r^j into one-hot for a supervised learning. We define the knowledge learning loss as

$$Loss_{kl} = CrossEntropy(Onehot(r^j), f^g). \tag{7}$$

We use $Loss_{kl}$ to supervise knowledge learning processing to update the knowledge embedding set. The knowledge embeddings have gradually become the common ground of different instances of the same category during the training procedure. As the representations become more discriminative in training, the knowledge embeddings become better to express the category. Furthermore, in this training procedure, the knowledge learning task considers the representations of all training images, making the knowledge embeddings more comprehensive in the expression of corresponding categories.

3.2.2. Knowledge exploiting

In order to effectively use knowledge, we first obtain a knowledge-based representation based on the knowledge response coefficients and the knowledge embeddings. The knowledge-based representation δ^j is computed as

$$\delta^j = \sum_{g \in G} r^j * k^g. \tag{8}$$

This knowledge-based representation includes a summary of the fine-grained features of different instances in the same category and a summary of the differences in different categories.

We use the knowledge to guide the classification by injecting the knowledge-based representation into the final fusion fine-grained representation:

$$u^j = FC\left(LN\left(y^j + \delta^j\right)\right), \tag{9}$$

where FC is the fully connective layer.

We define a representation learning loss to supervise the training to obtain the fusion representation:

$$Loss_{rep} = CrossEntropy(u^j, f^g). \quad (10)$$

We optimize the knowledge learning task and the knowledge exploiting task simultaneously so that the image representation and the knowledge can be updated iteratively and promote each other during the training process.

Thus, the total loss function of the whole network can be defined as

$$Loss = Loss_{kl} + \mu \times Loss_{rep}, \quad (11)$$

where μ is a hyperparameter used to adjust the different emphasis of the two tasks.

During the training, our method explicitly obtains the knowledge embeddings of the different fine-grained categories. These knowledge embeddings are distinguishable and comprehensive, so we insert them into the image-based representations to increase the comprehensiveness of features for the fine-grained recognition task. The proposed KG module does not significantly increase the computational complexity and the number of parameters in both training and testing phases.

Notably, the KG module essentially uses the correspondence between label information and different images to obtain embedding features, which we call knowledge embedding. Unlike the ordinary embedding module, we directly construct a classification task, using the label information as supervision, and fusing the information of multiple different images of the same category to obtain a comprehensive representation. For example, there is an embedding module in Prototypical Networks [42]. It first obtains the features and then calculates the prototype embeddings by averaging the features. As for the difference, features and knowledge embeddings are trained simultaneously in the proposed method.

In summary, the peak suppression module aims to consider more regions in a single image to obtain more diverse expressions. Based on the peak suppression module, the knowledge guidance module aims to extract and exploit the category-related embeddings based on the expression of multiple images. Therefore, these two modules are complementary in aggregating fragmented information at different levels. In the next Section 4, we conduct sufficient experiments to prove the effectiveness of the proposed method.

4. Experiments

4.1. Experimental setup

4.1.1. Datasets

We conduct our experiments on six fine-grained image recognition datasets, including two publicly available bird datasets CUB-200-2011 [43] and NABirds [44], one flower dataset Oxford 102 Flowers [45], one dog dataset Stanford Dogs [46], and two food datasets ISIA Food-200 [10] and ISIA Food-500 [47]. The detailed statistics about these six datasets including class numbers and train/test distributions are summarized in Table 1.

Table 1
Fine-grained image dataset statistics.

Dataset	# Class	# Training	# Testing
CUB-200-2011 [43]	200	5,994	5,794
Oxford Flowers [45]	102	1,020	6,149
Stanford Dogs [46]	120	12,000	8,580
NABirds [44]	555	23,929	24,633
ISIA Food-200 [10]	200	118,210	59,287
ISIA Food-500 [47]	500	239,379	120,143

4.1.2. Implementation details and comparison methods

Our method is implemented on the Pytorch platform with four Nvidia V100 GPUs. The input image size is 448 × 448 as most state-of-the-art fine-grained image recognition approaches. By following the settings of NTS-NET [20], we use data augmentations, including random cropping and horizontal flipping during the training procedure. Only the center cropping is involved in inference. The model is trained with the stochastic gradient descent (SGD) with a batch size of 8 and momentum of 0.9 for all datasets. The learning rate is set to 3e-2 initially, and the schedule applies a cosine decay function to an optimizer step. In all our experiments, we use ViT-B-16 pre-trained on ImageNet21k as the backbone. For all experiments, we adopt the top-1 accuracy as the evaluation metric. To demonstrate the advantages of our model, we list some methods for comparisons, as shown in Table 2. These methods use different backbones and input resolutions. Each method has its tendency to enlarge the influence of fine-grained objects by locating local regions or increase the recognition rate by learning a fine-grained representation globally. Different from these methods, the proposed method captures the information of multiple local areas through the peak suppression module. Furthermore, the proposed knowledge guidance module takes into account information from multiple images to obtain a comprehensive representation. Noteworthy, the proposed method is compared with the CNN-based approaches to illustrate that it achieves competitive performance to state-of-the-art, which is helpful to promote the application of transformers in the field of fine-grained image recognition. We also make comparisons under the same architecture to demonstrate the effectiveness of our method.

- DVAN [4]: Diversified visual attention network which relieves the dependency on supervised information.
- MaxEnt [48]: Maximum entropy approach which provides a training routine to maximize the entropy of the output probability distributions.
- NTS-NET [20]: Navigator-teacher-scrutinizer network which finds consistent informative regions through multi-agent cooperation.
- Cross-X [49]: Multi-scale feature learning with exploiting the relationships between different images and layers.
- GHNS [50]: A framework that generates features of hard negative samples.
- CSC-Net [51]: Category-specific semantic coherency network which semantically aligns the discriminative regions of the same subcategory.
- CDL [52]: Correlation-guided discriminative learning model which mines and exploits the discriminative potentials of correlations.
- MLA-CNN [53]: Multi-level attention model which uses the neural activations to generate multi-scale regions which are helpful for the fine-grained categorization.
- BiM-PMA [54]: Progressive mask attention model by leveraging both visual and language modalities.
- CIN [55]: Channel interaction network models channel-wise interplay within an image and across images.
- DB [29]: End-to-end network with a diversification block to use more subtle clues.
- FDL [22]: Filtration and distillation learning model with region proposing and feature learning.
- PMG [56]: Progressive multi-granularity method exploiting information based on the smaller granularity information found at the last step and the previous stage.

Table 2
The differences among the state-of-the-art methods. "Local-identification" column means the methods locate the discriminative semantic parts of fine-grained objects. "Global-discrimination" column means the methods generally learn the fine-grained feature embeddings. "# Images to consider" column means how many images the methods take into account per iteration.

Method	Backbone	Input Resolution	Local- identification	Global- discrimination	# Images to consider
DVAN [4]	VGG-16	224 × 224	✓	×	one
BiM-PMA [54]	VGG-16	448 × 448	✓	×	one
MLA-CNN [53]	VGG-19	448 × 448	✓	×	one
CSC-Net [51]	ResNet-50	224 × 224	✓	×	one
SCAPNet [60]	ResNet-50	224 × 224	✓	×	one
NTS-NET [20]	ResNet-50	448 × 448	✓	×	one
Cross-X [49]	ResNet-50	448 × 448	×	✓	multiple
GHNS [50]	ResNet-50	448 × 448	×	✓	one
CDL [52]	ResNet-50	448 × 448	✓	×	one
DB [29]	ResNet-50	448 × 448	×	✓	one
GaRD [57]	ResNet-50	448 × 448	✓	✓	one
HGNet [59]	ResNet-50	448 × 448	×	✓	one
CTF-CapsNet [61]	ResNet-50	448 × 448	✓	×	one
MSEC [62]	ResNet-50	448 × 448	✓	×	one
PMG [56]	ResNet-50	550 × 550	×	✓	one
CIN [55]	ResNet-101	448 × 448	×	✓	two
SnapMix [58]	ResNet-101	448 × 448	✓	×	one
MaxEnt [48]	DenseNet-161	-	×	✓	one
FDL [22]	DenseNet-161	448 × 448	✓	×	one
API-NET [30]	DenseNet-161	512 × 512	×	✓	two
CPM [17]	GoogLeNet	over 800	✓	×	one
ViT-ResNet-50 [13]	ViT-B-6 & ResNet-50	448 × 448	×	×	one
ViT [13]	ViT-B-16	448 × 448	×	×	one
TPSKG (ours)	ViT-B-16	448 × 448	✓	✓	multiple

- API-NET [30]: Pairwise interaction network which can progressively recognize a pair of images by interaction.
- CPM [17]: Complementary part model in a weakly supervised manner to retrieve information suppressed by dominant object parts detected by CNNs.
- GaRD [57]: Graph-based relation discovery approach to grasp the stronger contextual details.
- SnapMix [58]: Semantically proportional mixing which exploits CAM to lessen the label noise in augmenting fine-grained data.
- HGNet [59]: Hierarchical gate network to exploit the interconnection among hierarchical categories.
- SCAPNet [60]: Scale-consistent attention part network to guide part selection across multi-scales and keep the selection scale consistent.
- CTF-CapsNet [61]: Coarse-to-fine capsule network to shape an increasingly specialized description.
- MSEC [62]: Multi-Scale Erasure and Confusion which realizes confusion at different scales in images and sub-regions.

4.2. Ablation analysis

In this section, we conduct a series of ablation studies on the CUB-200-2011, Stanford Dog, Oxford 102 Flowers, NABirds, ISIA Food-200, and ISIA Food-500 datasets to better understand the designation of the proposed TPSKG. We use the performance of the original ViT-B-16 as the ablation baseline.

4.2.1. Impact of different components

To investigate the contribution of each component in the proposed method, we omit different components of TPSKG and report the corresponding top-1 recognition accuracy. From the results reported in Table 3, we can draw the following conclusions:

(1) The recognition accuracy on the CUB-200-2011 dataset drops from 91.3% to 91.0% and 90.9% when omitting the PS module and the KG module respectively, which demonstrates the effectiveness of both of the components for the fine-grained image recognition task. The experimental results on the other five datasets also have similar trends to the results of the CUB-200-2011 dataset, indicating that both the PS module and the KG module can effectively improve the recognition performance.

(2) The network with only the PS module improves the recognition accuracy of baseline by 0.5% (90.4% vs. 90.9%) on CUB-200-2011 dataset, shows that details other than the most significant part are helpful for the fine-grained image recognition task. At the same time, the network with the KG module improves 0.6% (90.4% vs. 91.0%) on the CUB-200-2011 dataset, showing that integrating the discriminative information of multiple images can effectively improve the recognition performance of the network. This result is consistent with our analysis of discriminative information fragmentation in fine-grained image recognition tasks.

(3) The PS module improves the recognition accuracy of baseline by 1.3% (59.9% vs. 61.2%) and the KG module improves the recognition accuracy of baseline by 2.1% (59.9% vs. 62.0%) on the ISIA Food-500 dataset, the combination of these two modules can improve the recognition accuracy of baseline by 5.5% (59.9% vs. 65.4%). This experimental phenomenon shows that the PS module and the KG module are complementary and can promote each other. The more diverse the expression obtained by the peak suppression module, the stronger the expression ability of embedding learned by the knowledge guidance module.

(4) We conduct ablation experiments to verify the effectiveness of knowledge learning and knowledge exploiting. We randomly initialized the knowledge embedding and fixed it to explore the case without the knowledge learning, resulting in a significant drop in network recognition performance. The overall performance of the method drops slightly without the knowledge exploiting.

4.2.2. Visualizations of knowledge embedding

The knowledge embedding is a category-related feature representation learning with the similarity as the classification score directly. The embedding becomes the most similar to the feature

Table 3
Ablation results of the proposed TPSKG on the fine-grained image datasets.

Model	CUB-200-2011	Stanford Dog	Oxford Flowers	NABirds	ISIA Food-200	ISIA Food-500
ViT-B-16	90.4	91.4	99.2	89.6	67.4	59.9
w/o Peak Suppression	91.0	91.8	99.3	89.9	69.3	62.0
w/o Knowledge learning	90.7	91.5	99.2	89.6	67.6	60.8
w/o Knowledge exploiting	91.0	91.8	99.3	89.9	69.3	63.0
w/o Knowledge Guidance	90.9	91.8	99.3	89.8	68.3	61.2
All TPSKG	91.3	92.5	99.5	90.1	69.5	65.4

expression in the corresponding category and the most dissimilar to the feature expression in other categories. As shown in Fig. 4: (1) The knowledge-based representation and the image-based representation are in the same subspace, which is convenient for feature fusion. (2) These two representations of the same category are close in the subspace, indicating that knowledge-based representation can express category-related information. (3) The knowledge-based representation is separable in the subspace, so it is helpful for recognition tasks.

4.2.3. Choice of μ:

Since Eq. 11 requires selecting a hyperparameter μ, it is essential to study the influence of classification performance on the choice of μ. We conduct this experiment for four different μ on the CUB-200-2011 dataset. As shown in Fig. 5, the experimental results show that (1) The performance is relatively robust to the choice of μ generally. (2) The model performs better when the weight of $Loss_{rep}$ is slightly larger than $Loss_{kl}$. The probable reason is that the learning of knowledge relies on the combined effect of representation and label. A slightly larger weight of $Loss_{rep}$ allows the network to learn the discriminative feature representation preferentially. Because the performance of the model is not sensitive to the choice of μ, the weighting coefficient in Eq. 11 is empirically chosen to be μ=2 in all the following experiments.

4.3. Comparison with state-of-the-art

For further verification for the TPSKG, we compare our method to the state-of-the-art methods on the six publicly available fine-grained datasets in this section.

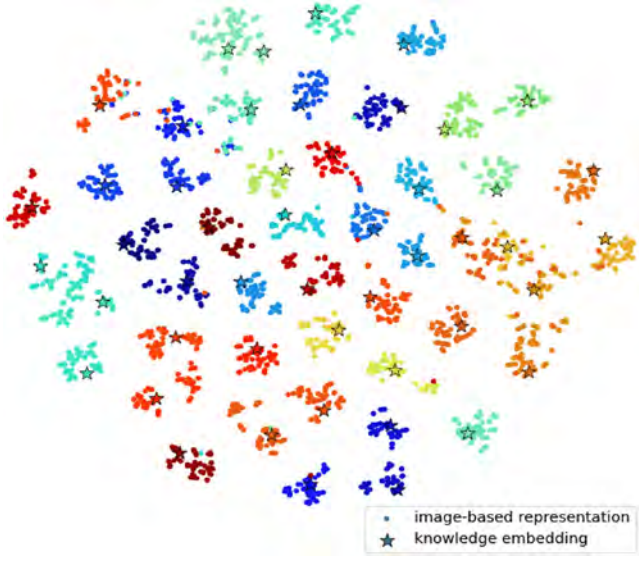

Fig. 4. The t-SNE visualizations of image representations and knowledge embedding of 50 sample categories from the CUB-200-2011 dataset. Each dot stands for an image representation, and each star represents the knowledge embedding of a category. The color indicates the categories.

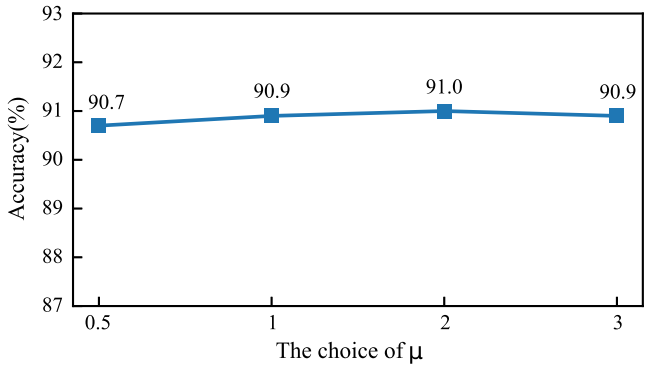

Fig. 5. Performance on the choice of μ on the CUB-200-2011 dataset.

4.3.1. CUB-200-2011

We compare the proposed method against many state-of-the-art fine-grained recognition models on CUB-200-2011, as shown in Table 4. The results show the following conclusions.

(1) Overall, the proposed TPSKG performs better than the state-of-the-art fine-grained methods, including the global-discrimination approaches methods MaxEnt [48] and DB [29], and the part-based methods NTS-NET [20] and CPM [17].
(2) Images with higher resolutions usually contain richer information and subtle details that are important for the fine-grained image recognition task. According to the literature

Table 4
Comparison results on CUB-200-2011 dataset.

Method	Backbone	Resolution	Accuracy(%)
MaxEnt [48]	DenseNet-161	-	84.9
MLA-CNN [53]	VGG-19	448 × 448	85.7
DVAN [4]	VGG-16	224 × 224	87.1
NTS-NET [20]	ResNet-50	448 × 448	87.5
Cross-X [49]	ResNet-50	448 × 448	87.7
HGNet [59]	ResNet-50	448 × 448	87.9
CIN [55]	ResNet-101	448 × 448	88.1
MSEC [62]	ResNet-50	448 × 448	88.3
CDL [52]	ResNet-50	448 × 448	88.4
DB [29]	ResNet-50	448 × 448	88.6
CTF-CapsNet [61]	ResNet-50	448 × 448	88.6
HGNet [59]	ResNet-50	448 × 448	88.7
GHNS [50]	ResNet-50	448 × 448	89.1
FDL [22]	DenseNet-161	448 × 448	89.1
CSC-Net [51]	ResNet-50	224 × 224	89.2
SCAPNet [60]	ResNet-50	224 × 224	89.5
PMG [56]	ResNet-50	550 × 550	89.6
GaRD [57]	ResNet-50	448 × 448	89.6
SnapMix [58]	ResNet-101	448 × 448	89.6
API-NET [30]	DenseNet-161	512 × 512	90.0
CPM [17]	GoogLeNet	over 800	90.4
ViT-ResNet-50	ViT&ResNet-50	448 × 448	89.2
ViT [13]	ViT-B-16	448 × 448	90.4
TPSKG	ViT-B-16	448 × 448	**91.3**

[63], higher resolution input images will produce better performance generally. Our method uses a smaller resolution than PMG [56], API-NET [30], and CPM [17] but achieves a better performance. At the same time, our method is possible to perform better with higher resolution.
(3) CPM [17] has good performance using the stacked BiLSTMs to integrate the patch features. The performance of the original ViT method is equivalent to the CPM, which proves the effectiveness of the transformer in feature aggregation and its potential in the fine-grained recognition task.
(4) Although the original ViT model has satisfactory performance, we have improved it by 0.9%.
(5) Both CIN [55] and API-NET [30] have achieved good results based on the contrastive learning mechanism. It is worth noting that the improvement from our framework is orthogonal to those works, so the proposed TPSKG can also benefit from these methods.

4.3.2. Stanford dog

As can be seen from Table 5, our method shows more performance improvement on the Stanford Dog dataset, which is 2.2% higher than the current state-of-the-art method API-NET [30] without using the contrastive learning mechanism and the high-resolution input. It is worth noting that the performance of the original ViT also exceeds API-NET by 1.1%, which reflects that the transformer architecture can be well migrated to fine-grained recognition task.

4.3.3. Oxford 102 flowers

Unlike BiM-PMA [54], which uses all 2040 images in the training set and validation set for training, we follow the settings of PC-CNN [64] and PBC [65] and only use 1020 images in training set for training to ensure a relatively fair comparison. As can be seen from Table 6, although using fewer images, our method still achieves a 2.1% performance improvement compared to BiM-PMA. At the same time, our method still improves the recognition performance when the recognition performance of ViT is excellent.

4.3.4. NABirds

The NABirds dataset is a larger fine-grained dataset than the CUB-200-2011 dataset containing 48,562 North American bird images. Many methods with complex operations are not easy to experiment on a data set of this order of magnitude. If an image generates thousands of proposals, it means that tens of millions of proposals need to be processed. Table 7 reports the performance of several methods on the NABirds dataset. DSTL [63] uses the transfer learning strategy for the fine-grained image recognition task consisting of more than one dataset. The result of our method trained on a separate dataset exceeds DSTL by 2.2% in accuracy.

Table 5
Comparison results on Stanford Dog dataset.

Method	Backbone	Resolution	Accuracy(%)
DVAN [4]	VGG-16	224 × 224	81.5
MaxEnt [48]	DenseNet-161	-	83.6
PC-CNN [64]	DenseNet-161	224 × 224	83.8
FDL [22]	DenseNet-161	448 × 448	84.9
MSEC [62]	ResNet-50	448 × 448	85.6
MLA-CNN [53]	VGG-19	448 × 448	86.8
DB [29]	ResNet-50	448 × 448	87.7
Cross-X [49]	ResNet-50	448 × 448	88.9
API-NET [30]	DenseNet-161	512 × 512	90.3
ViT-ResNet-50	ViT&ResNet-50	448 × 448	87.7
ViT [13]	ViT-B-16	448 × 448	91.4
TPSKG	ViT-B-16	448 × 448	**92.5**

Table 6
Comparison results on Oxford 102 Flowers dataset.

Method	Backbone	Resolution	Accuracy(%)
PC-CNN [64]	DenseNet-161	224 × 224	93.6
PBC [65]	GoogleNet	224 × 224	96.1
BiM-PMA [54]	VGG-16	448 × 448	96.9
ViT-ResNet-50	ViT&ResNet-50	448 × 448	98.5
ViT [13]	ViT-B-16	448 × 448	99.2
TPSKG	ViT-B-16	448 × 448	**99.5**

4.3.5. ISIA Food-200

In order to further verify the effectiveness and explore the scope of application of our method, we explored the food recognition task on the ISIA Food-200 dataset. Unlike other fine-grained objects, many types of food are non-rigid and lack a fixed spatial structure and semantic pattern. Therefore, it is challenging to capture specific semantic information from food images. Our method attempts to increase the diversity of representations to cover more distinguished areas, which is more effective for non-rigid objects. Simultaneously, our method injects the extracted knowledge into the image-based representation so that a more comprehensive understanding of food categories can be used in the recognition task.

Table 8 reports the performance of several methods on the ISIA Food-200 dataset. The ViT is also mediocre on this task. IG-CMAN [10] is a patch-based method sequentially localizing multiple informative image regions with multi-scale from category level to ingredient-level guidance in a coarse-to-fine manner. Our method achieves the best 69.5% without using the multi-scale strategy and outperforms the state-of-the-art method IG-CMAN by 2.0% in accuracy. This result proves that our method has a more significant performance improvement in complex recognition problems.

4.3.6. ISIA Food-500

The ISIA Food-500 is a more comprehensive food dataset than the ISIA Food-200 with a larger data volume and higher diversity. We evaluated the proposed TPSKG against different fine-grained methods in Table 9. The performance of ViT-ResNet-50 and ViT show a pronounced decline. The possible reason is that the volume and complexity of the ISIA Food-500 dataset are much higher than that of the ISIA Food-200 dataset. This increase in complexity makes the impact of the loss of local region semantics more significant. It can also be seen that the proposed method exceeds the original ViT significantly, with a gain of 5.5% in accuracy. When the performance of the original ViT is poor, our method still achieves competitive performance compared to the state-of-the-art method, which proves that our method does not rely heavily on the performance of ViT. The proposed method obtains better accuracy than the SGLANet without a complicated multi-scale mechanism and spatial-channel attention. We will try to leverage the multi-scale information to improve performance in future work.

Table 7
Comparison results on NABirds dataset.

Method	Backbone	Resolution	Accuracy(%)
PC-CNN [64]	DenseNet-161	224 × 224	82.8
MaxEnt [48]	DenseNet-161	-	83.0
Cross-X [49]	ResNet-50	448 × 448	86.4
HGNet [59]	ResNet-50	448 × 448	86.4
DSTL [63]	Inception-v3	560 × 560	87.9
GaRD [57]	ResNet-50	448 × 448	88.0
ViT-ResNet-50	ViT&ResNet-50	448 × 448	86.7
ViT [13]	ViT-B-16	448 × 448	89.6
TPSKG	ViT-B-16	448 × 448	**90.1**

Table 8
Comparison results on ISIA Food-200 dataset.

Method	Backbone	Resolution	Accuracy(%)
ResNet-152	ResNet-152	224 × 224	61.1
DenseNet-161	DenseNet-161	224 × 224	62.6
IG-CMAN [10]	DenseNet-161	224 × 224	67.5
ViT-ResNet-50	ViT&ResNet-50	448 × 448	62.5
ViT [13]	ViT-B-16	448 × 448	67.4
TPSKG	ViT-B-16	448 × 448	**69.5**

Table 9
Comparison results on ISIA Food-500 dataset.

Method	Backbone	Resolution	Accuracy(%)
ResNet-152	ResNet-152	224 × 224	57.0
WRN-50 [66]	WRN-50	224 × 224	60.1
WS-DAN [67]	Inception-v3	299 × 299	60.7
NAS-NET [68]	ResNet-152	331 × 331	60.7
NTS-NET [20]	ResNet-152	448 × 448	63.7
SENet-154	SENet-154	224 × 224	63.8
DCL [28]	ResNet-152	448 × 448	64.1
SGLANet [47]	SENet-154	224 × 224	64.7
ViT-ResNet-50	ViT&ResNet-50	448 × 448	52.7
ViT [13]	ViT-B-16	448 × 448	59.9
TPSKG	ViT-B-16	448 × 448	**65.4**

4.3.7. Overall

We summarize the results of all datasets to obtain an overall understanding of the proposed method. We can find that (1) The recognition results of the different methods for the six different datasets show a very high degree of correspondence, indicating the strong reproducibility. (2) The performance of the hybrid model directly generated by the simple combination of ViT and ResNet-50 generally performs poorly on fine-grained image recognition task and even has performance degradation compared to the ViT model. A possible explanation for these results may be the lack of adequate semantic information in small regions. (3) The ViT model is generally suitable for simple fine-grained recognition tasks and obtains close to state-of-the-art results on multiple datasets, but it does not perform well for more complex food recognition tasks. (4) The proposed PS module and KG module have effectively improved the recognition performance, and the proposed method has achieved very superior performance on all datasets. (5) The KG module improves the model performance more significantly than the PS module. The possible explanation for this might be that the benefits of integrating discriminative information in multiple images are more significant than the coverage of more discriminative information areas in one single image.

4.4. Qualitative visualization

To show what happen during training, we visualize the attention maps of the TPSKG for the 1000th and 5000th iterations, as shown in Fig. 7. It can be clearly observed that the proposed method reduces the focus on the most discriminative parts of the image and increases the attention on other discriminative parts, which are significant but often overlooked.

In order to show the effectiveness of the method more intuitively, we visualize the attention maps of the original ViT and TPSKG models for sample images from six different datasets. As shown in Fig. 6, we find that although the original ViT can also perform localization and recognition, our method is better in both aspects. The attention map of the proposed TPSKG can not only locate the essential parts well but also cover more discriminative areas, which shows that the method is more robust than the original ViT. For the CUB-200-2011, Stanford Dog, Oxford Flowers and NABirds datasets with relatively simple scenes, features without the diversity can complete the recognition task. For relatively complex food recognition tasks on the ISIA Food-200 and ISIA Food-500 datasets, the diversity of features is more important, and our method has improved more obviously, which is consistent with the results of quantitative analysis.

To visually analyze the influence of the knowledge guidance module, we visualize the feature representations before/after injecting the knowledge embedding by employing t-SNE, on the six fine-grained image datasets as shown in Fig. 8. The visualized data includes all test set images of CUB-200-2011, Stanford Dogs, Oxford Flowers, NAbirds, ISIA Food-200 datasets and 50 sample categories from the ISIA Food-500 dataset due to excessive data volume. Since the classification performance of the first four datasets is very high, the visualized images are not much different. But the latter two more complex food recognition datasets show signif-

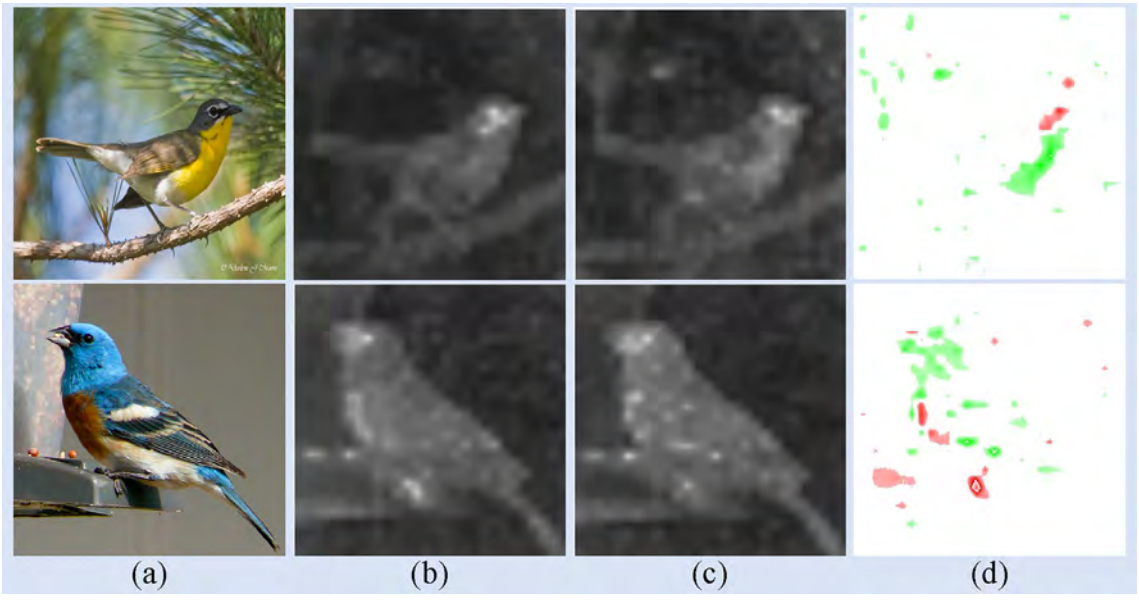

Fig. 7. The attention map comparison between 1000th and 5000th iteration. Left to right: (a) original image, (b) attention map of 1000th iteration, (c) attention map of 5000th iteration, (d) difference between (b) and (c). The red means the weights reduce and the green means the weights increase.

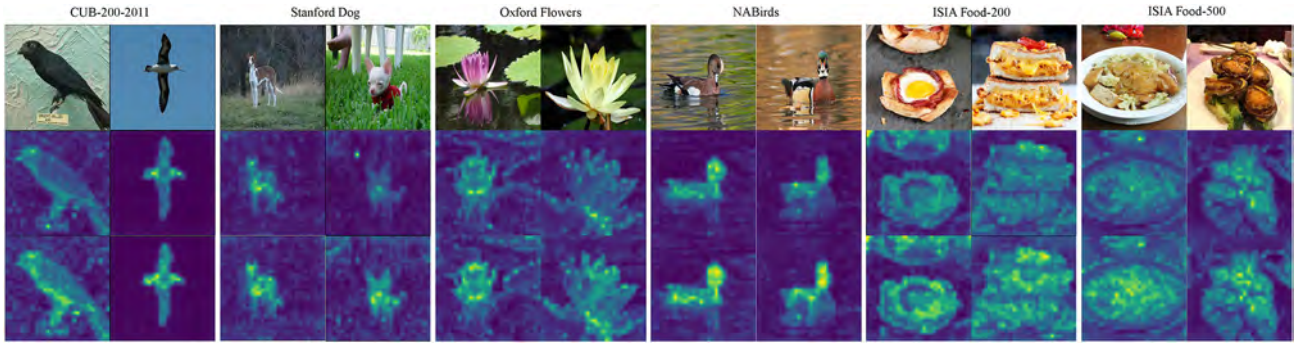

Fig. 6. The attention map comparison between our method and the baseline in different datasets. Top to below: original image, attention map of the ViT, attention map of our method. The yellow means high weights and the blue means relatively low weights.

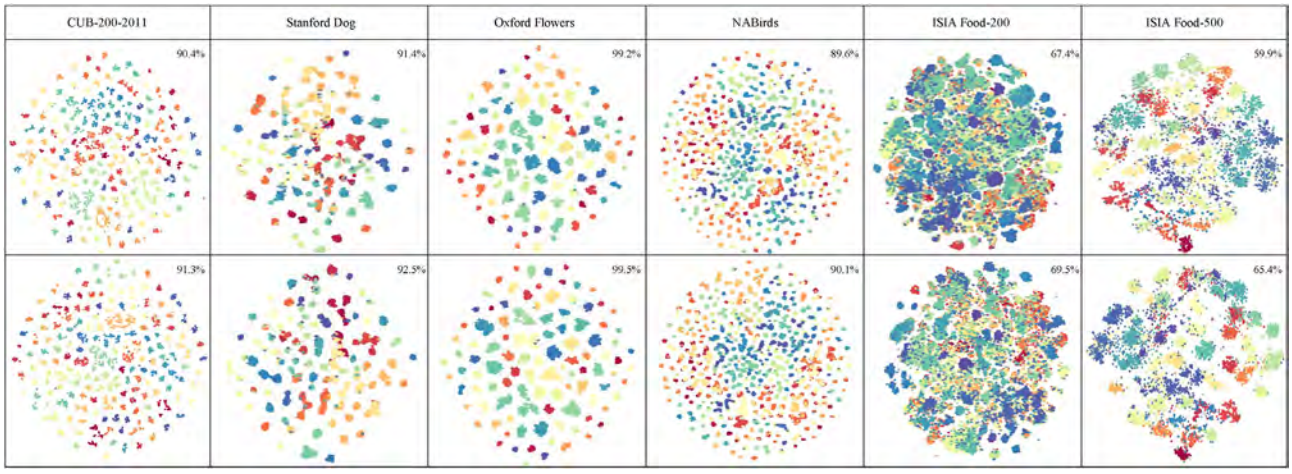

Fig. 8. The t-SNE visualization of fine-grained image feature representations of (top row) before injecting the knowledge embedding, (bottom row) after injecting the knowledge embedding on the six fine-grained datasets. Each color represents a different class. The upper right corner shows the accuracy of the corresponding method.

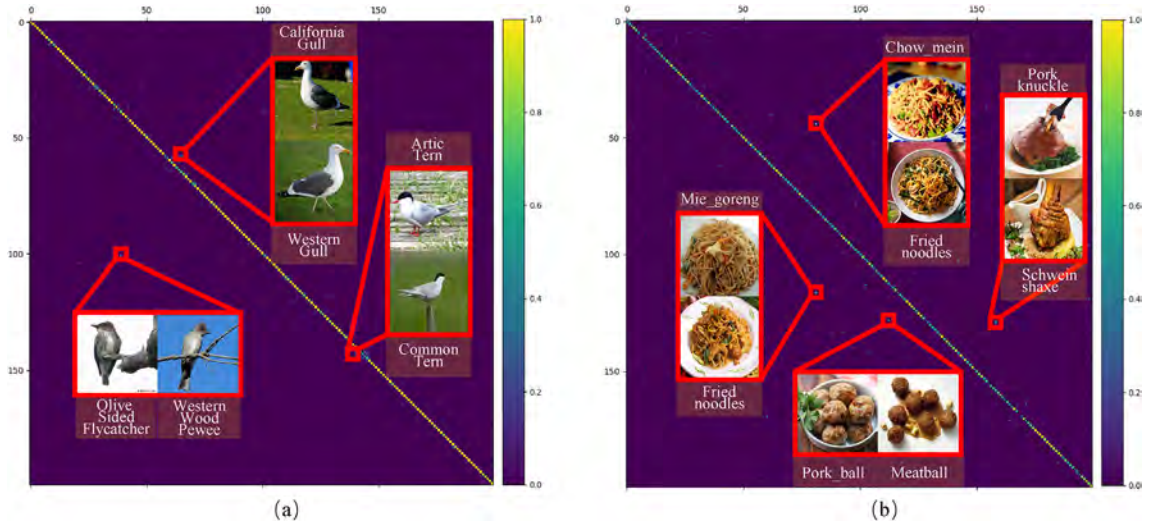

Fig. 9. Confusion matrix of our method on the (a) CUB-200-2011 and (b) ISIA Food-200 datasets. Some instances of low recognition rate categories are annotated by red boxes.

icant differences. The visualization results of food recognition datasets in this figure show obvious intra-class clustering. This proves that the proposed KG module has a strong intra-class aggregation ability. Although the t-SNE method cannot prove the method's ability for the recognition task, it can be seen from the local structure that our method has a more vital ability to aggregate features within the class. This investigation confirms that feature representations will get into more separable clusters after injecting the category-related knowledge embedding.

In addition, we further show the confusion matrix of our method on the CUB-200-2011 and ISIA Food-200 datasets in Fig. 9, where the vertical axis shows the ground-truth classes, and the horizontal axis shows the predicted classes. Yellower colors indicate better performance. We can see that our method still does not provide perfect performance for some bird and food categories. We enlarge specific regions to highlight the misclassified results and show some samples with low recognition rates. As shown in Fig. 9 (a), some birds of the same meta-category are extremely difficult to distinguish, such as California Gull and Western Gull, Artic Tern and Common Tern. There are also images with similar poses that cannot be recognized well, such as Olive Sided Flycatcher and Wester Wood Pewee, requiring further study and exploration. From Fig. 9 (b) we can see that these food categories are very similar in visual appearances, such as Chow mein, Mie Goreng, and Fried noodles. Even humans cannot easily distinguish these food categories based on images. Some food categories have the same ingredients and different cooking techniques, which are difficult to distinguish, such as pork knuckle and the Schweinshaxe. Many types of food are difficult to classify based on images alone. A possible solution is to combine multiple media formats of information for the recognition task, such as ingredient lists and cooking processes. The transformer architecture has promising applications in multimedia, including the visual field and the NLP field, and provides the possibility for a unified framework.

5. Conclusion

Fine-grained image recognition is an interesting and fundamental topic. In this paper, we investigate the problem of fine-grained image recognition from the perspective of fragmented information integration. Furthermore, we present a transformer with peak suppression and knowledge guidance (TPSKG) for the fine-grained image recognition task. Our method learns the diverse fine-grained representations by the peak suppression module penalizing the most discriminative parts. It then learns the knowledge embedding including a large number of discriminative clues for different images of the same category, and injects them into fine-grained representations leading to significantly higher recognition performance. The proposed network can be trained end-to-end in one stage, requiring no bounding box/part annotations. Qualitative and quantitative evaluations on six public fine-grained datasets demonstrate that the proposed TPSKG can achieve competitive performance compared to the state-of-the-art approaches.

CRediT authorship contribution statement

Xinda Liu: Methodology, Investigation, Writing – original draft. **Lili Wang:** Supervision, Writing – review & editing. **Xiaoguang Han:** Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China through Projects 61932003 and 61772051, by National Key R&D plan 2019YFC1521102, by the Beijing Natural Science Foundation L182016, by the funding of Shenzhen Research Institute of Big Data (Shenzhen 518000).

References

[1] K. Han, J. Guo, C. Zhang, M. Zhu, Attribute-aware attention model for fine-grained representation learning, International Conference on Multimedia, ACM (2018) 2040–2048.
[2] X. He, Y. Peng, Only learn one sample: Fine-grained visual categorization with one sample training, International Conference on Multimedia, ACM (2018) 1372–1380.
[3] J. Li, L. Zhu, Z. Huang, K. Lu, J. Zhao, I read, I saw, I tell: Texts assisted fine-grained visual classification, in: International Conference on Multimedia, ACM, 2018, pp. 663–671.
[4] B. Zhao, X. Wu, J. Feng, Q. Peng, S. Yan, Diversified visual attention networks for fine-grained object classification, IEEE Trans. Multimedia 19 (6) (2017) 1245–1256.
[5] V.M. Araújo, A.S. Britto Jr., L.S. Oliveira, A.L. Koerich, Two-view fine-grained classification of plant species, Neurocomputing 467 (2022) 427–441.
[6] L.A. Hendricks, S. Venugopalan, M. Rohrbach, R.J. Mooney, K. Saenko, T. Darrell, Deep compositional captioning: Describing novel object categories without paired training data, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 1–10.
[7] J. Bao, D. Chen, F. Wen, H. Li, G. Hua, CVAE-GAN: fine-grained image generation through asymmetric training, International Conference on Computer Vision, IEEE (2017) 2764–2773.
[8] O.M. Aodha, S. Su, Y. Chen, P. Perona, Y. Yue, Teaching categories to human learners with visual explanations, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 3820–3828.
[9] K. Pang, Y. Yang, T.M. Hospedales, T. Xiang, Y. Song, Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2020, pp. 10344–10352.
[10] W. Min, L. Liu, Z. Luo, S. Jiang, Ingredient-guided cascaded multi-attention network for food recognition, International Conference on Multimedia, ACM (2019) 1331–1339.
[11] W. Min, S. Jiang, J. Sang, H. Wang, X. Liu, L. Herranz, Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration, IEEE Trans. Multimedia 19 (5) (2017) 1100–1113.
[12] W. Min, S. Jiang, R.C. Jain, Food recommendation: Framework, existing solutions, and challenges, IEEE Trans. Multimedia 22 (10) (2020) 2659–2671.
[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.
[14] S. Huang, Z. Xu, D. Tao, Y. Zhang, Part-stacked CNN for fine-grained visual categorization, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 1173–1182.
[15] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: International Conference on Machine Learning, vol. 32, JMLR, 2014, pp. 647–655.
[16] W. Min, S. Jiang, L. Liu, Y. Rui, R.C. Jain, A survey on food computing, ACM Comput. Surveys 52 (5) (2019) 92:1–92:36.
[17] W. Ge, X. Lin, Y. Yu, Weakly supervised complementary parts models for fine-grained image classification from the bottom up, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 3034–3043.
[18] S. Jiang, W. Min, L. Liu, Z. Luo, Multi-scale multi-view deep feature aggregation for food recognition, IEEE Trans. Image Process. 29 (2020) 265–276.
[19] J. Wang, N. Li, Z. Luo, Z. Zhong, S. Li, High-order-interaction for weakly supervised fine-grained visual categorization, Neurocomputing 464 (2021) 27–36.
[20] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, L. Wang, Learning to navigate for fine-grained classification, in: European Conference on Computer Vision, vol. 11218, Springer, 2018, pp. 438–454.
[21] X. Liu, W. Min, S. Mei, L. Wang, S. Jiang, Plant disease recognition: A large-scale benchmark dataset and a visual region and loss reweighting approach, IEEE Trans. Image Process. 30 (2021) 2003–2015.
[22] C. Liu, H. Xie, Z. Zha, L. Ma, L. Yu, Y. Zhang, Filtration and distillation: Enhancing region attention for fine-grained visual categorization, in: The AAAI Conference on Artificial Intelligence, AAAI, 2020, pp. 11555–11562.
[23] Y. Peng, X. He, J. Zhao, Object-part attention model for fine-grained image classification, IEEE Trans. Image Process. 27 (3) (2018) 1487–1500.
[24] H. Zheng, J. Fu, Z. Zha, J. Luo, T. Mei, Learning rich part hierarchies with progressive attention networks for fine-grained image recognition, IEEE Trans. Image Process. 29 (2020) 476–488.
[25] Z. Wang, S. Wang, S. Yang, H. Li, J. Li, Z. Li, Weakly supervised fine-grained image classification via guassian mixture model oriented discriminative learning, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2020, pp. 9746–9755.
[26] X. Wang, S. Li, C. Chen, A. Hao, H. Qin, Depth quality-aware selective saliency fusion for RGB-D image salient object detection, Neurocomputing 432 (2021) 44–56.
[27] T. Lin, A. RoyChowdhury, S. Maji, Bilinear convolutional neural networks for fine-grained visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1309–1322.
[28] Y. Chen, Y. Bai, W. Zhang, T. Mei, Destruction and construction learning for fine-grained image recognition, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 5157–5166.

[29] G. Sun, H. Cholakkal, S. Khan, F.S. Khan, L. Shao, Fine-grained recognition: Accounting for subtle differences between similar classes, in: The AAAI Conference on Artificial Intelligence, AAAI, 2020, pp. 12047–12054.
[30] P. Zhuang, Y. Wang, Y. Qiao, Learning attentive pairwise interaction for fine-grained classification, in: The AAAI Conference on Artificial Intelligence, AAAI, 2020, pp. 13130–13137.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Conference on Neural Information Processing Systems (2017) 5998–6008.
[32] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Conference of the Association for Computational Linguistics, Association for Computational Linguistics, 2019.
[33] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Conference on Neural Information Processing Systems, 2020.
[34] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, Vol. 139, PMLR, 2021, pp. 10347–10357.
[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Vol. 12346, Springer, 2020, pp. 213–229.
[36] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: deformable transformers for end-to-end object detection, CoRR.
[37] L. Ye, M. Rochan, Z. Liu, Y. Wang, Cross-modal self-attention network for referring image segmentation, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 10502–10511.
[38] F. Yang, H. Yang, J. Fu, H. Lu, B. Guo, Learning texture transformer network for image super-resolution, Conference on Computer Vision and Pattern Recognition (2020) 5791–5800.
[39] J. Cordonnier, A. Loukas, M. Jaggi, On the relationship between self-attention and convolutional layers, in: International Conference on Learning Representations, 2020.
[40] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, I. Sutskever, Generative pretraining from pixels, in: International Conference on Machine Learning, Vol. 119, PMLR, 2020, pp. 1691–1703.
[41] S. Abnar, W.H. Zuidema, Quantifying attention flow in transformers, in: Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 4190–4197.
[42] J. Snell, K. Swersky, R.S. Zemel, Prototypical networks for few-shot learning, Conference on Neural Information Processing Systems (2017) 4077–4087.
[43] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset, Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011).
[44] G.V. Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, S.J. Belongie, Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 595–604.
[45] M. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Indian Conference on Computer Vision, Graphics & Image Processing, IEEE, 2008, pp. 722–729.
[46] A. Khosla, N. Jayadevaprakash, B. Yao, L. Fei-Fei, Novel dataset for fine-grained image categorization, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2011.
[47] W. Min, L. Liu, Z. Wang, Z. Luo, X. Wei, X. Wei, S. Jiang, ISIA Food-500: A dataset for large-scale food recognition via stacked global-local attention network, in: International Conference on Multimedia, ACM, 2020.
[48] A. Dubey, O. Gupta, R. Raskar, N. Naik, Maximum-entropy fine grained classification, Conference on Neural Information Processing Systems (2018) 635–645.
[49] W. Luo, X. Yang, X. Mo, Y. Lu, L. Davis, J. Li, J. Yang, S. Lim, Cross-x learning for fine-grained visual categorization, in: International Conference on Computer Vision, IEEE, 2019, pp. 8241–8250.
[50] T. Kim, K. Hong, H. Byun, The feature generator of hard negative samples for fine-grained image recognition, Neurocomputing 439 (2021) 374–382.
[51] S. Wang, Z. Wang, H. Li, W. Ouyang, Category-specific semantic coherency learning for fine-grained image recognition, International Conference on Multimedia, ACM (2020) 174–183.
[52] Z. Wang, S. Wang, P. Zhang, H. Li, W. Zhong, J. Li, Weakly supervised fine-grained image classification via correlation-guided discriminative learning, in: International Conference on Multimedia, ACM, 2019, pp. 1851–1860.
[53] J. Ji, Y. Guo, Z. Yang, T. Zhang, X. Lu, Multi-level dictionary learning for fine-grained images categorization with attention model, Neurocomputing 453 (2021) 403–412.
[54] K. Song, X. Wei, X. Shu, R. Song, J. Lu, Bi-modal progressive mask attention for fine-grained recognition, IEEE Trans. Image Process. 29 (2020) 7006–7018.
[55] Y. Gao, X. Han, X. Wang, W. Huang, M.R. Scott, Channel interaction networks for fine-grained image categorization, in: The AAAI Conference on Artificial Intelligence, AAAI, 2020, pp. 10818–10825.
[56] R. Du, D. Chang, A.K. Bhunia, J. Xie, Z. Ma, Y. Song, J. Guo, Fine-grained visual classification via progressive multi-granularity training of jigsaw patches, in: European Conference on Computer Vision, Vol. 12365, Springer, 2020, pp. 153–168.
[57] Y. Zhao, K. Yan, F. Huang, J. Li, Graph-based high-order relation discovery for fine-grained recognition, Conference on Computer Vision and Pattern Recognition (2021) 15079–15088.
[58] S. Huang, X. Wang, D. Tao, Snapmix: Semantically proportional mixing for augmenting fine-grained data, in: The AAAI Conference on Artificial Intelligence, AAAI, 2021, pp. 1628–1636.
[59] Y. Chen, J. Song, M. Song, Hierarchical gate network for fine-grained visual recognition, Neurocomputing.
[60] H. Liu, J. Li, D. Li, J. See, W. Lin, Learning scale-consistent attention part network for fine-grained image recognition, IEEE Trans. Multimedia (2021), 1–1.
[61] Z. Lin, J. Jia, F. Huang, W. Gao, A coarse-to-fine capsule network for fine-grained image categorization, Neurocomputing 456 (2021) 200–219.
[62] Y. Zhang, Y. Sun, N. Wang, Z. Gao, F. Chen, C. Wang, J. Tang, MSEC: multi-scale erasure and confusion for fine-grained image classification, Neurocomputing 449 (2021) 1–14.
[63] Y. Cui, Y. Song, C. Sun, A. Howard, S.J. Belongie, Large scale fine-grained categorization and domain-specific transfer learning, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 4109–4118.
[64] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, N. Naik, Pairwise confusion for fine-grained visual classification, in: European Conference on Computer Vision, Vol. 11216, Springer, 2018, pp. 71–88.
[65] C. Huang, H. Li, Y. Xie, Q. Wu, B. Luo, PBC: polygon-based classifier for fine-grained categorization, IEEE Trans. Multimedia 19 (4) (2017) 673–684.
[66] S. Zagoruyko, N. Komodakis, Wide residual networks, British Machine Vision Conference, BMVA (2016).
[67] T. Hu, H. Qi, See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification, CoRR abs/1901.09891.
[68] B. Zoph, V. Vasudevan, J. Shlens, Q.V. Le, Learning transferable architectures for scalable image recognition, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 8697–8710.

Xinda Liu received the B.E. degree from the China University of Mining and Technology, Beijing, China, in 2013, the M.E. degree from Ningxia University, Yinchuan, China, in 2016, and is currently working toward the Ph.D. degree at the China State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include machine learning and image processing.

Lili Wang received the PhD degree from the Beihang University, Beijing, China. She is a professor with the School of Computer Science and Engineering, Beihang University, and a researcher with the State Key Laboratory of Virtual Reality Technology and Systems. Her research direction is in VR, AR and Computer Graphics including real-time rendering, realistic rendering, global illumination, soft shadow, and texture synthesis.

Xiaoguang Han received his B.Sc. degree in mathematics in 2009 from Nanjing University of Aeronautics and Astronatics and his M.Sc. degree in applied mathematics in 2011 from Zhejiang University. He obtained his Ph.D. degree in 2017 from the University of Hong Kong. He is currently an assistant professor at Shenzhen Research Institute of Big Data, the Chinese University of Hong Kong (Shenzhen). His research mainly focuses on computer vision, computer graphics, and 3D deep learning.