Computer Vision and Image Understanding 229 (2023) 103658

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Siamese self-supervised learning for fine-grained visual classification

Ruyi Ji [a,b], Jiaying Li [c], Libo Zhang [a,b,d,*]

[a] Institute of Software Chinese Academy of Sciences, Beijing, 100190, China
[b] University of Chinese Academy of Sciences, Beijing, 100049, China
[c] Beijing Information Science and Technology University, Beijing, 100192, China
[d] Nanjing Institute of Software Technology, Beijing, 210000, China

ARTICLE INFO

Communicated by Nikos Paragios

MSC:
41A05
41A10
65D05
65D17

Keywords:
Fine-grained visual classification
Seamese network
Self-supervised learning

ABSTRACT

Fine-grained visual classification (FGVC) is challenging to capture subtle yet distinct visual cues due to large intra-class and small inter-class variances. To this end, we propose a new Siamese Self-supervised Learning method to perform alignment between different views of one image. Specifically, we employ the attention mechanism to explore the semantic parts of one image, and then generate different views by *crop* and *erase* strategy. Meanwhile, we adopt the Siamese network to perform the feature alignment across various views and capture the view-invariant feature in a self-supervised way. Finally, we introduce the center loss to explicitly ensure consistency between different views. Extensive experimental results show the proposed method performs on par with the state-of-the-art methods on three public benchmarks including CUB-200-2011, FGVC-Aircraft, and Stanford Cars.

1. Introduction

The FGVC aims to classify sub-categories under the same super-category, for example, different species of birds and dogs. As a upstream research, FGVC has facilitated a wide variety of applications in the downstream tasks such as person re-identification (Leng et al., 2020; Hong et al., 2021), instance segmentation (Zhang et al., 2021c) and emotion detection (Abdul-Mageed and Ungar, 2017). Nevertheless, it remains a highly challenging task to date and has attracted extensive research attention in computer vision field over the past few years.

As depicted in Fig. 1, FGVC is challenging due to two reasons: (1) high intra-class variances: the samples belonging to the same category usually present significant different appearance, (2) low inter-class variances: the samples belonging to different categories usually share similar appearance. Recently, considerable efforts have been made to alleviate these issues. A set of methods (Lin et al., 2015a; Chang et al., 2020; Zhuang et al., 2020; He et al., 2019) exemplify this research strand. In specific, B-CNN (Lin et al., 2015a) utilizes the bilinear pooling to encode features generated by parallel extractors to form richer representations. MC-Loss (Chang et al., 2020) takes advantage of two channel-wise modules to excavate the discriminability and diversity of features within a single network. However, the aforementioned methods fall short in ensuring the consistency of feature distribution in each category explicitly, resulting in degraded performance.

To address this issue, we propose a novel end-to-end Siamese Self-supervised Learning for the FGVC task in this study. The proposed method is constructed on the concept of Siamese architecture. Driven by the advantages of Siamese network, we apply it to encode view-invariant features and align features of different views of one image. Moreover, we impose the consistency constraint on features of different views to prevent them from contradicting each other. In short, the contributions of this paper are summarized as follows.

- To alleviate the issue of large intra-class and small inter-class variances, a new Siamese Self-supervised Learning (SSSL) method is proposed for the FGVC task, where we utilize Siamese architecture with shared parameters to encode the feature from different views and guide model to learn view-invariant features in a self-supervised learning way.
- To our best knowledge, we are the first to introduce self-supervised learning to ensure the consistency between different views of an image explicitly and the proposed model relies on the positive pairs for optimization, without the limitation of batch size.
- The proposed method performs favorably against the state-of-the-art methods on three challenging FGVC benchmarks including CUB-200-2011, FGVC-Aircraft and Stanford Cars. Comprehensive ablative studies are provided to shed light on the effectiveness of model design choices.

* Corresponding author.
E-mail address: libo@iscas.ac.cn (L. Zhang).

https://doi.org/10.1016/j.cviu.2023.103658
Received 19 September 2022; Received in revised form 31 January 2023; Accepted 15 February 2023
Available online 23 February 2023

2. Related work

2.1. Fine-grained visual categorization

It is vital to characterize the detailed visual cues for the fine-grained visual classification. Early, prior arts (Zhang et al., 2014, 2016; Lin et al., 2015b; Huang et al., 2016) directly guide models to capture discrepancies in subtle regions with the supervision of object- or part-level annotations. However, such annotation requires expertise and is labor-intensive, making these methods inapplicable for real scenarios. Later, developments over this task shift towards weakly supervised learning with image-level label. Roughly speaking, there are two primary research groups for the existing approaches. The first category focuses on exploring the potential of neural network to enrich feature representation. The representatives of this research line are methods (Lin et al., 2015a; Yu et al., 2018). Specifically, Lin et al. (2015a) popularize the bilinear pooling which assembles features from two extractors into a high-order representation that characterize pair-wise interaction information. Later, its follow-up approach (Yu et al., 2018) presents hierarchical bilinear pooling to aggregate features from the multiple scales. Nevertheless, this kind of learning discriminative representation strategy is hard to verify whether higher-order feature can pay enough attention to discriminative regions. The direct influence is that the learned representation fails to characterize the fine-grained object explicitly. To close this gap, the other category attends to exploring object- or part-based discriminative visual cues. Generally, the salient response of feature activation maps underpins this research strand. For example, MMAL-Net (Zhang et al., 2021b) locates the salient object based on activation maps and excavates diverse parts by a sliding window mechanism. Du et al. (2020) iteratively encode image patches at multiple scales to capture multi-granularity informative visual clues. Even though these models achieve decent improvements, they disregard rich correlation information between discriminative parts, resulting in the limited performance. Differing from them, we focus on learning views-invariant feature from different views of an image to form a robust representation and ensure class consistency explicitly.

2.2. Siamese neural network

A Siamese Neural Network (SNN) is a commonly used strategy to quantify the similarity between different instances. Since seminal work is proposed by Chopra et al. (2005) for face verification, Siamese neural network has been quite popular in the computer vision field and derives many excellent variants in a wide range of visual understanding tasks, such as object tracking (Shan et al., 2021) and image matching (Melekhov et al., 2016). Encouraged by their success, we construct network framework based on such architecture in our study. However, rather than the similarities between sample pairs, we emphasize how to derive the view-invariant information from the different views of one image, which renders the base of the proposed method.

2.3. Self-supervised learning

Without the requirement of annotation information, self-supervised learning is tasked with learning the discriminative representation from unlabeled data. Early works of self-supervised learning usually require numerous negative pairs to circumvent collapsed solutions. For instance, approaches (Xu et al., 2021b; Chen et al., 2020a,b, 2021) intuitively adopt a large batch size to cover numerous negative samples. To alleviate the limitation of excessive batch size, the work (Wu et al., 2018) preserves all instance features in a memory bank. Method (He et al., 2020) sets up a queue to contain samples from multiple mini-batches with an updating mechanism of enqueuing and dequeuing. Notably, contrastive learning (Hadsell et al., 2006) is introduced into the above approaches to pull the positive sample pairs and push the negative ones in the embedding space. The recent advance BYOL (Grill et al., 2020) uses the online network to predict the output of target network, and proves the feasibility of self-supervised learning without negative sample pairs. BYOL adopts two different views of one image as the inputs of online network and target network respectively, which is suitable for various image augmentation as well. By contrast, our model utilizes the self-supervised learning scheme to prevent the learned feature of different views of one image from contradicting each other.

Fig. 1. Samples of fine-grained visual categories. FGVC remains challenging due to the following two factors: (1) high intra-class variances: the birds belonging to the same category usually present significantly different appearances, such as illumination variations (the first column), clutter background (the second column), occlusion (the third column) and view-point changes (the fourth column); (2) low inter-class variances: the birds in different columns belong to different categories, but share similar appearance in the same rows.

2.4. Attention mechanism

Sharing the similar spirit with human visual system, attention mechanism is successfully adopted to highlight relevant parts and depress the irrelevant parts. As one of the astounding works, Hu et al. (2018) innovatively consider the attention mechanism from the channel perspective and assign the various weights based on the contribution of each channel. To characterize discriminative feature, the CBAM module (Woo et al., 2018) considers attention mechanisms from both spatial and channel perspectives. Analogous to the CBAM (Woo et al., 2018), Park et al. (2018) design a BAM module to build a hierarchical attention at bottlenecks. Recently, Wang et al. (2020a) present a self-attention importance-weighting module to assess the contributions of samples during the training stage. Different from the aforementioned methods, we rely on the attention mechanism to form the different semantic views of one image.

3. Method

In this section, we describe the rationale behind Siamese Self-supervised Learning for the FGVC task in detail. As shown in Fig. 2, the proposed network is mainly composed of the following three components: siamese encoder, self-supervised learning, and loss function. First, the siamese encoder is used to extract latent features from raw image. Then, we perform cropping and erasing operations on the high response areas to form the different semantic views. Meanwhile, we explicitly enforce consistency on category centers shared by different observations via center loss in a self-supervised learning manner. In the end, the shared fully-connected layer followed by a softmax layer is adopted to finalize the prediction of subordinate category.

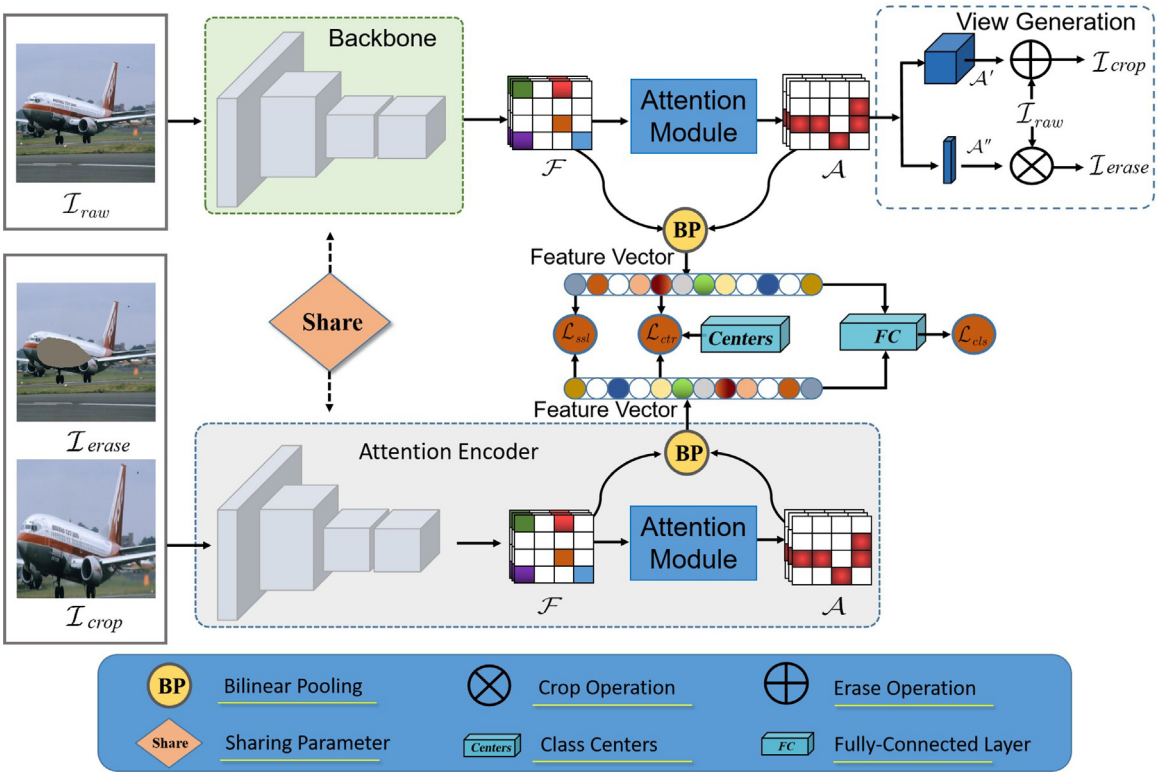

Fig. 2. Illustration of our Siamese Self-supervised Learning network architecture for the FGVC task. The proposed method is built on the concept of Siamese architecture to learn the class consistency of different views of attention guided feature representation. Wherein $\mathcal{F}$, $\mathcal{A}$, and BP represent extracted features, attention maps, and bilinear pooling, respectively.

3.1. Siamese encoder

In our design, the raw image and its two views share a siamese encoder which is formed by the two major components: backbone and attention module.

Backbone. It is well-known that fine-grained datasets are usually characterized by small volume. For the sake of accurate performance, we apply a truncated ResNet model (discarding the last fully-connected layers) as our backbone, which is pre-trained on the ILSVRC CLSLOC dataset (Russakovsky et al., 2015). Let an image $\mathcal{I}$ be fed into the CNN-based backbone to extract basic features $\mathcal{F} \in \mathbb{R}^{C\times H\times W}$, where C represents the number of channel, H and W indicate the height and width of feature maps respectively. It is noteworthy that the proposed method can work well on diverse pre-trained networks, such as VGG (Simonyan and Zisserman, 2015), ResNet (He et al., 2015), and Inception (Szegedy et al., 2016).

Attention Module. Next, we employ Bilinear Pooling (BP) strategy to explicitly capture the second-order statistics of the basic features. In detail, we first leverage the attention module to generate K attention maps $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}, \mathcal{A} \in \mathbb{R}^{K\times H\times W}$, which is mathematically calculated as follows,

$$\mathcal{A} = ReLU(BN(Conv(\mathcal{F})) \tag{1}$$

where $Conv(\cdot)$ indicates a convolution layer (kernel 1×1, output channel K, and stride 1), $BN(\cdot)$ (Ioffe and Szegedy, 2015) denotes the batch normalization operation and $ReLU(\cdot)$ (Glorot et al., 2011) refers to Rectified Linear Unit. To highlight high response regions on attention maps, we perform the broadcast element-wise multiplication between feature and individual attention map in $\mathcal{A}$, as formulated in Eq. (2),

$$\mathcal{F}_i' = \mathcal{F} \odot a_i \tag{2}$$

where $\odot$ denotes element-wise multiplication between feature maps and i-th attention map. Sequentially, we leverage the average global pooling followed by the concatenation operation to obtain the holistic feature f_r of the raw image.

$$f_r = \prod_{i=1}^{K}(GAP(\mathcal{F}_i')), \qquad f_r \in \mathbb{R}^{KC\times 1} \tag{3}$$

where $\prod$ denotes the concatenation operation and $GAP(\cdot)$ means the global average pooling operation.

3.2. Self-supervised learning

Multi-view is of vital importance in computer vision, which provides distinctive visual and semantic clues for recognition and understanding. For fine-grained visual classification, apart from the appearance, the multi-view also provide view consistency between views. Visual consistency is imposed by ensuring that visually similar views of the same image are encoded to have similar feature representation. For semantic consistency, we impose the constraint that the various views of an image should be predicted to have the same class. Such a schema enforce model to focus on discriminative parts of a fine-grained target, which effectively alleviate the issue of large intra-class and small inter-class variance. In the following, we shift our focus on how to generate different semantic views of an image based on the feature $\mathcal{F}$ and attention maps $\mathcal{A}$, and then present the optimization of self-supervised learning for the proposed method in detail.

View-cropped. As stated in method (Ding et al., 2019; Zhang et al., 2019), it can considerably enhance discriminative feature representation by removing irrelevant information on background. We randomly choose $M \in [1, K]$ feature maps from $\mathcal{A}$ to compose $\mathcal{A}' \in \mathbb{R}^{M\times H\times W}$. The explanation of this schema is that a desirable cropped view should ensure the semantic integrity of object and stable robustness of the model simultaneously. Moreover, we re-liberate importance feature maps in $\mathcal{A}'$. To be specific, we first pool feature maps into the form

Algorithm 1 Search Connected regions via a binary mask $\mathcal{M}_c$

Require: A binary mask $\mathcal{M}_c$;

```
Pick a pixel p ∈ M_c as the starting point;
while True do
    Leverage a flood-fill algorithm to label all the pixels in the connected region that covers the pixel p;
    if All the pixels traverse then
        Break;
    end if
    Search for the next unprocessed pixel as p;
end while
return Connectivity of the connected regions, and the according region size
```

of estimated probabilistic distribution α_i $(i = 1, 2, \ldots, M)$. Formally, α_i is mathematically calculated as,

$$\alpha_i = \frac{GAP(\mathcal{A}_i')}{\sum_{j=1}^{M} GAP(\mathcal{A}_j')}, \qquad \mathcal{A}_i' \in \mathcal{A}' \tag{4}$$

Then, the regularized feature map C is derived by a weighted sum of $\mathcal{A}_i'$, which is expressed as,

$$C = \sum_{i=1}^{M} \alpha_i \mathcal{A}_i' \tag{5}$$

After that, we bilinearly interpolate the calculated C to match the raw image resolution, *i.e.*, $C \rightarrow \mathcal{M}_c$. Based on an empirical threshold ε_1, we crop an attentive area as a view of the raw image. Therein, the value in $\mathcal{M}_c$ larger than $\varepsilon_1 * max\ (\mathcal{M}_c)$ is set to 1, otherwise assigned to 0. $\mathcal{M}_c$ can be formulated as,

$$\mathcal{M}_c(i,j) = \begin{cases} 1 & if\ \mathcal{M}_c(i,j) \geq \varepsilon_1 * max(\mathcal{M}_c), \\ 0 & otherwise. \end{cases} \tag{6}$$

Next, we utilize the Algorithm 1 to search the largest connected region $\hat{\mathcal{M}}_c$ from the $\mathcal{M}_c$. Thereby, a view is formed by cropping operation and resized to match the resolution of the raw image.

$$\mathcal{I}_{crop} = \mathcal{I} \odot \hat{\mathcal{M}}_c \tag{7}$$

In our design, we experimentally observe that the multiple involved feature maps ensure the integrity of decisive facade object while the random selection can enhance the robustness in training process.

View-erased. It has been proven that intentionally erasing high response region of an image can enforce neural network to focus on the rest of semantic parts. Inspired by prior works (Hu and Qi, 2019; Sun et al., 2020), we introduce this strategy to construct the other view from the raw image. In detail, we first randomly select one attention map from $\mathcal{A}$ as $\mathcal{A}'' \in \mathbb{R}^{H \times W}$ and re-scale $\mathcal{A}''$ as an erasing mask $\mathcal{M}_e$. Contrary to cropping operation, we set the value in $\mathcal{M}_e$ less than $\varepsilon_2 * max\ (\mathcal{M}_e)$ to 1, and assign others to 0, which is expressed as follows,

$$\mathcal{M}_e(i,j) = \begin{cases} 1 & if\ \mathcal{M}_e(i,j) \leq \varepsilon_2 * max(\mathcal{M}_e), \\ 0 & otherwise. \end{cases} \tag{8}$$

where ε_2 is an empirical threshold as well. Then, we perform the element-wise multiplication operation between the raw image $\mathcal{I}$ and $\mathcal{M}_e$ to form the view-erased.

$$\mathcal{I}_{erase} = \mathcal{I} \odot \mathcal{M}_e \tag{9}$$

Notably, different from the view-cropped, here we focus on removing one of the discriminative parts. Therefore, randomly choosing one feature map from $\mathcal{A}$ fits this goal well. Additionally, above strategy can serve as one kind of data augmentation at the training stage.

Optimization. After generating view-specific representations, we need to solve how to optimize the proposed model. As shown in Fig. 2, we leverage the siamese encoder to encode features from two views and define them as f_c and f_e respectively. To train the proposed method efficiently, similar to the BYOL (Grill et al., 2020), only positive pairs are involved in our self-supervised learning pipeline to avoid the shackles of overlarge batch size. We impose the ℓ_2 constraint on view-specific features to capture view-invariant representation and ensure class consistency, where the loss is formulated as follows,

$$\mathcal{L}_{ssl} = \| f_c - f_e \|_2 \tag{10}$$

where $\| \cdot \|_2$ represents ℓ_2 loss and $f_{i \in \{c,e\}} \in \mathbb{R}^{KC \times 1}$.

3.3. Loss function

Center loss. For each category, we first define the class centers as $C \in \mathbb{R}^{KC \times N}$ whose initialization obeys uniform distribution, N is the number of subcategories. In forward propagation process, the class center is updated by moving average in Eq. (11). It is clear that center distribution is involved with feature representation of category from preceding epochs.

$$c' \leftarrow c + \bigtriangledown (f_r - c) \tag{11}$$

where $c \in C$ and $c \in \mathbb{R}^{KC \times 1}$. Here c and c' refer to a specific class center before and after iterations respectively. $\bigtriangledown$ is the step for updating the center. And we use the ℓ_2 loss to impose the constraint between f_r and C, which is formulated as below,

$$\mathcal{L}_{ctr} = \| f_r - C \|_2 \tag{12}$$

Classification loss. Rather than aforementioned constraints, we adopt the cross-entropy loss among category predictions and the corresponding ground-truth label to optimize the proposed model.

$$\mathcal{L}_{cls} = 1/3 \sum_{i \in \{r,c,e\}} \mathcal{L}_{ce}(y_i, y^*) \tag{13}$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss between the predictions $y_{i \in \{r,c,e\}}$, computed by the raw image and different views, and the corresponding ground-truth label y^*,

Hence, we unify all above loss constraints together to guide model towards the better performance at the training stage. The overall loss of the proposed model can be written as following,

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \mathcal{L}_{ctr} + \beta \mathcal{L}_{ssl} \tag{14}$$

where α, β are hyper-parameters to balance three loss items and set to be 0.1 in our experiment empirically.

Notably, we assemble high response of all attention maps to assure enough discriminative information at test stage. To be exact, we directly average K attention maps to gain a holistic attention map, for the purpose of streamlining the operation efficiently. Next, we locate and crop the salient object from the raw image to finalize the prediction.

4. Experiments

4.1. Implementation details

We utilize PyTorch (Paszke et al., 2017) as default deep learning framework to implement the proposed model and the entire model is trained on a workstation with a 2.20 GHz Intel processor, 1 Tesla V100 GPU and 32 GB memory. Following common protocol in FGVC task, we adopt a truncated model pre-trained on ILSVRC CLSLOC (Russakovsky et al., 2015) as backbone. The input image is resized to be 448 × 448 with color jittering. Random rotation and random horizontal flip are applied for data augmentation. All the above settings are standard in the literature. Stochastic Gradient Descent (SGD) algorithm, with the momentum of 0.9, epoch number of 160, weight decay of 0.00001, batch size of 12, is applied to optimize the proposed model in an end-to-end manner. The initial learning rate is set to 0.001, which is multiplied by 0.9 in every 2 epochs by fixed step size decay learning rate schedule. Unless otherwise specified, Acc@1 stands for the top-1 classification accuracy in performance reports. All the source codes of the proposed method will be made publicly available after the paper is accepted.

Table 1
Comparison results on CUB-200–2011 dataset.

Methods	Backbone	Acc@1(%)
MC Loss (Chang et al., 2020)	VGG-16	78.7
RA-CNN (Fu et al., 2017)	VGG-19	85.3
Improved B-CNN (Lin and Maji, 2017)	VGG-16	85.8
MAMC (Sun et al., 2018)	ResNet-50	86.2
MA-CNN (Zheng et al., 2017)	VGG-19	86.5
DFL-CNN (Wang et al.)	VGG-16	86.7
AP-CNN (Ding et al., 2021)	VGG-19	86.7
PC (Dubey et al., 2018)	DenseNet-161	86.9
ACNet (Ji et al., 2020)	VGG-16	87.8
CIN (Gao et al., 2020)	ResNet-101	88.1
SnapMix (Huang et al., 2020)	ResNet-101	88.4
Grad-CAM (Xu et al., 2021a)	ResNet-50	88.5
ELoPE (Hanselmann and Ney, 2020)	ResNet-101	88.5
S3N (Ding et al., 2019)	ResNet-50	88.5
DB (Sun et al., 2020)	ResNet-50	88.6
MPN-COV (Li et al., 2018)	ResNet-101	88.7
PMG (Du et al., 2020)	VGG-16	88.8
DF-GMM (Wang et al., 2020c)	ResNet-50	88.8
PCA-Net (Zhang et al., 2021a)	ResNet-101	88.9
FDL (Liu et al., 2020)	DenseNet-161	89.1
Mix+ (Li et al., 2020)	ResNet-50	89.2
WS_DAN (Hu and Qi, 2019)	Inception-V3	89.4
TBMSL-Net (Zhang et al., 2021b)	ResNet-50	89.6
API-Net (Zhuang et al., 2020)	DenseNet-161	90.0
Stacked LSTM (Ge et al., 2019)	GoogleNet	90.4
Baseline	ResNet-101	85.7
Ours	ResNet-101	**90.6**

Table 2
Comparison results on FGVC-Aircraft dataset.

Methods	Backbone	Acc@1(%)
Improved B-CNN (Lin and Maji, 2017)	VGG-16	88.5
MA-CNN (Zheng et al., 2017)	VGG-19	89.8
MC Loss (Chang et al., 2020)	VGG-16	91.0
FDL (Liu et al., 2020)	DenseNet-161	91.3
NTS (Yang et al., 2018)	ResNet-50	91.4
MPN-COV (Li et al., 2018)	ResNet-101	91.4
ACNet (Ji et al., 2020)	VGG-16	91.5
DFL-CNN (Wang et al.)	VGG-16	92.0
RA-CNN (Fu et al., 2017)	VGG-19	92.5
PMG (Du et al., 2020)	VGG-16	92.7
S3N (Ding et al., 2019)	ResNet-50	92.8
PCA-Net (Zhang et al., 2021a)	ResNet-101	92.8
EfficientNet-B7 (Tan and Le, 2019)	EfficientNet-B7	92.9
PC (Dubey et al., 2018)	DenseNet-161	92.9
WS_DAN (Hu and Qi, 2019)	Inception-V3	93.0
Mix+ (Li et al., 2020)	ResNet-50	93.1
GCL (Wang et al., 2020b)	ResNet-50	93.2
CIN (Gao et al., 2020)	ResNet-101	93.3
ELoPE (Hanselmann and Ney, 2020)	ResNet-101	93.5
DB (Sun et al., 2020)	ResNet-50	93.5
Multi Granularity (Chang et al., 2021)	ResNet-50	93.6
SnapMix (Huang et al., 2020)	ResNet-101	93.7
DF-GMM (Wang et al., 2020c)	ResNet-50	93.8
API-Net (Zhuang et al., 2020)	DenseNet-161	93.9
CAL (Rao et al., 2021)	ResNet-101	94.5
TBMSL-Net (Zhang et al., 2021b)	ResNet-50	94.7
Baseline	ResNet-101	91.4
Ours	ResNet-101	**94.8**

4.2. Fine-grained visual classification datasets

CUB-200-2011. The CUB-200-2011 is a bird species dataset, which is competitive and widely-used for fine-grained image classification. The dataset contains 11,788 images of 200 different bird subcategories, which consists of 5,994 images for training and 5,794 images for testing. There are roughly 30 train and test images for each subcategory.

Stanford Cars. The Stanford Cars dataset has 16,185 images from 196 classes, officially split into 8,144 training and 8,041 testing images for 196 categories. Each category's number is roughly 50–50 split and the sub-category can be determined by cars' brand, model, and year.

Table 3
Comparison results on Stanford Cars dataset.

Methods	Backbone	Acc@1(%)
MPN-COV (Li et al., 2018)	ResNet-101	93.3
TASN (Zheng et al., 2019)	ResNet-50	93.8
NTS (Yang et al., 2018)	ResNet-50	93.9
SEF (Luo et al., 2020)	ResNet-50	94.0
GCL (Wang et al., 2020b)	ResNet-50	94.0
PMG (Du et al., 2020)	VGG-16	94.3
CIN (Gao et al., 2020)	ResNet-101	94.5
WS_DAN (Hu and Qi, 2019)	Inception-V3	94.5
Cross-X (Luo et al., 2019)	ResNet-50	94.6
AP-CNN (Ding et al., 2021)	VGG-19	94.6
S3N (Ding et al., 2019)	ResNet-50	94.7
DF-GMM (Wang et al., 2020c)	ResNet-50	94.8
AutoAugment (Cubuk et al., 2018)	Inception-V3	94.8
Mix+ (Li et al., 2020)	ResNet-50	94.9
ELoPE (Hanselmann and Ney, 2020)	ResNet-101	95.0
TBMSL-Net (Zhang et al., 2021b)	ResNet-50	95.0
Multi Granularity (Chang et al., 2021)	ResNet-50	95.1
API-Net (Zhuang et al., 2020)	DenseNet-161	95.3
CAL (Rao et al., 2021)	ResNet-101	95.5
Baseline	ResNet-101	92.8
Ours	ResNet-101	**95.5**

FGVC-Aircraft. The FGVC-Aircraft dataset contains 10,000 images for 100 categories, which are divided into 6,667 training and 3,333 testing images. And the train/test set split ratio is around 2 : 1. Most images in this dataset are airplanes. And the dataset is organized in a four-level hierarchy, from finer to coarser: Model, Variant, Family, Manufacturer.

4.3. Comparison with state-of-the-art methods

CUB-200-2011. Table 1 shows our method can expressly outperform FDL (Liu et al., 2020) and WS_DAN (Hu and Qi, 2019), even if they apply more complicated backbone. It can be observed that we exceed two methods relying on data augmentation (WS_DAN (Hu and Qi, 2019) and SnapMix (Huang et al., 2020)) by 1.2% and 2.2% top-1 accuracy, respectively. Compared to Stacked LSTM (Ge et al., 2019), which relies on additional object detection and instance segmentation to capture complementary information, we still show obvious advantage over it, *i.e.*, 90.6% *vs.* 90.4%, demonstrate the effectiveness of the proposed method.

FGVC-Aircraft. As Table 2 reports, our method performs best on FGVC-Aircraft dataset. Compared to the methods like (PCA-Net (Zhang et al., 2021a) and API-Net (Zhuang et al., 2020)) construct image pairs elaborately to capture contrastive interaction information, we significantly outperform them by absolute improvements of 2.0% and 0.9% in terms of top-1 accuracy. Our method exceeds Improved B-CNN (Lin and Maji, 2017) by a large margin, *i.e.*, 88.5% *vs.* 94.8%. Moreover, due to the dynamic generation of various view of an image, we can relieve prohibitive computation burden compared with TBMSL-Net (Zhang et al., 2021b) which rigidly chooses 7 local parts based on the summation of each window.

Stanford Cars. As listed in Table 3, our approach achieves a very competitive result again. Compared to previous method CIN (Gao et al., 2020), our method achieves absolute improvement of 1.0% in terms of Top-1 accuracy. The proposed method performs better than AP-CNN (Ding et al., 2021), even it adopts top-down pathway to fuse information of different levels. We believe that the performance gap between the proposed method and those like AP-CNN (Ding et al., 2021) demonstrates that our model has strong ability to capture the discriminative cues for the FGVC task.

Table 4
Ablation study on self-supervised learning, center loss and bilinear pooling CUB-200–2011 dataset. Here SSL denotes the self-supervised learning, CL means the center loss and BL refers to the bilinear pooling.

SSL	CL	BL	Acc@1(%)
✗	✗	✗	85.7
✗	✓	✓	89.3
✓	✗	✓	89.7
✓	✓	✗	89.6
✓	✓	✓	90.6

Table 5
Ablation study on the effect of views on CUB-200–2011 dataset.

Variants	Acc.@1
variant w/o crop and erase	86.3
variant w/ crop	88.5
variant w/ erase	88.8
variant w/ crop and erase	89.6

Table 6
Ablation study on the effect of α on CUB-200–2011 dataset.

value of α	0.01	0.1	0.3	0.5	0.7	1	2
Acc.@1	89.7	90.6	90.3	90.2	90.0	89.8	89.5

Table 7
Ablation study on the effect of β on CUB-200–2011 dataset.

value of β	0.01	0.1	0.3	0.5	0.7	1	2
Acc.@1	89.6	90.6	90.5	90.5	90.2	90.1	89.7

Qualitative visualization. For an intuitive understanding and comparison, we visualize the attention maps yielded by the baseline ResNet-101 (He et al., 2015), approach (Hu and Qi, 2019) and ours in Fig. 3. One can see that there is a dilemma for existing methods to eliminate large intra-class and small inter-class variances in the feature space. It is clearly observed that discriminative semantic parts pose challenges on baseline ResNet-101 and high responses even diffuse into the background for approach (Hu and Qi, 2019). By comparison, our method can locate the subtle discriminative parts more accurately and attend more decisive facade parts such as the wings, tail, and head of an airplane; doors, headlights, and glasses of a car; beak, breasts, and wings of a bird.

4.4. Ablation studies

To pose insight into our design choices comprehensively, we perform the ablation studies on important components or hyper-parameters in our model with ResNet-101 backbone on CUB-200-2011 dataset.

The effect of self-supervised learning. For better performance, self-supervised learning is adopted to ensure the consistency of view-specific features in our method. To investigate the effectiveness of self-supervised learning in our design, we construct a variant without the self-supervised learning scheme, namely, there is no guarantee to explicitly ensure the consistency between different views. As can be seen from Table 4, such strategy can bring an absolute increment of 1.3% in terms of classification accuracy (see the 2-nd row and the 5-th row). To further analyze the effect of semantic views, we perform additional experiments as demonstrated in Table 5, where "variant w/o crop and erase" denotes that we adopt commonly used data augmentation strategy (random erase) to generate a new view, "variant w/ crop" indicates we only generate a *crop* view by the guidance of attention mechanism, "variant w/ erase" means that we only generate a *erase* view by the guidance of attention mechanism, "variant w/ crop and erase" demonstrate that we generate two view just as described in Section 3. As Table 5 shows, the performance of "variant w/o crop and erase" drops a lot compared with "variant w/ crop and erase". And "variant w/ crop" and "ariant w/ erase" validate the effectiveness of such two views separately. We argue that the best performance, *i.e.*, "variant w/ crop and erase" mainly comes from the consistency between two views, which allows the proposed method to perform the alignment of view-specific features within the subcategory and alleviates the issue of large intra-class and inter-class variances to a certain extent.

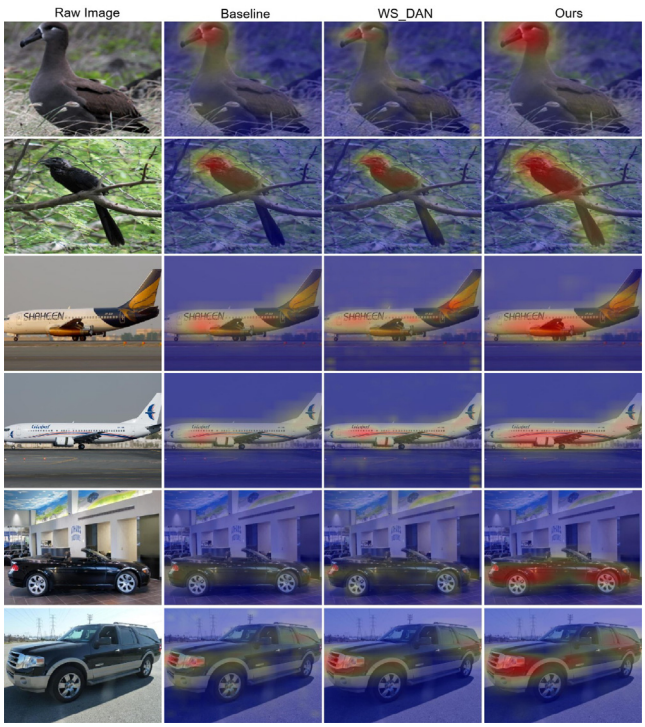

Fig. 3. The visual comparison between baseline, WS_DAN and ours.

The impact of center loss. The performance comparison between the proposed model and the variant without center loss is reported in last row and the third row of Table 4. When unpaired with the center loss, there is a sharp drop in classification performance, *i.e.*, 90, 6% *vs.* 89.7%. We conjecture that the center loss assists with the proposed model to learn intrinsic discriminative features shared by different views. Thereby, the feature similarities between samples in the class are further enhanced.

The influence of the bilinear pooling. To verify the efficacy of the bilinear pooling strategy in our method, we construct a variant without the bilinear pooling, which means we directly perform global average pooling on the basic feature $\mathcal{F}$ and finalize the subcategory prediction. As Table 4 shows, when coupled with the bilinear pooling strategy, the classification accuracy achieves an absolute improvement of 1% performance. We believe that bilinear pooling strategy allows the proposed model to encode higher-order statistic information from the basic feature, which facilitates model to achieve the accuracy results.

The influence of α an β. To determine desirable values of the empirical thresholds α and β, we conduct experiments with the different thresholds and summarize results in Tables 6 and 7. Experiments in the above tables show that along with the values α and β varying from 0.01 to 2, the classification accuracy significantly fluctuates. And we show that the proposed model is saturated with the values of $\alpha = 0.1$ and

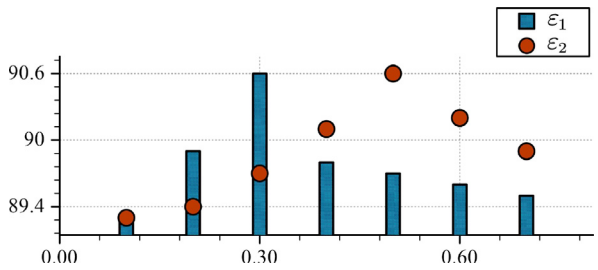

Fig. 4. The influence of the empirical thresholds ε_1 and ε_2 on CUB-200-2011 dataset.

$\beta = 0.1$. Hence, the values of α and β are empirically set to 0.1 in our experimental settings by default for better classification performance.

The analysis of ε_1 and ε_2. For the desire values of the empirical thresholds ε_1 and ε_2, we investigate numerous variants with different values and visualize experiment results in Fig. 4. We show that there is an obvious fluctuation on the classification accuracy as ε_1 and ε_2 change. And we can see that the proposed model is saturated with the value of $\varepsilon_2 = 0.5$ and $\varepsilon_1 = 0.3$. Hence, for accurate performance, we set ε_1 and ε_2 to be 0.3 and 0.5 empirically in our default setting.

5. Conclusion

In this paper, we propose a Siamese Self-supervised Learning for the fine-grained visual classification task. Specifically, different semantic views of an image are generated by the strategy of cropping and erasing under the guidance of attention mechanism and aligned to learn view-invariant representation by the Siamese architecture with shared parameters. The whole network is optimized by stochastic gradient decent algorithm in an end-to-end manner. Extensive experimental results conducted on CUB-200-2011, FGVC-Aircraft, and Stanford Cars demonstrate that the proposed method achieves favorable performance against the state-of-art methods. Comprehensive ablation studies and qualitative visualization further verify the efficacy of important components in the proposed method. The promising results provided by our model pave the way for better classification models for the FGVC task in the future. We hope the above findings shed light on the promising directions for the fine-grained visual classification task.

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.

Data availability

Data will be made available on request.

Acknowledgment

All authors approved the version of the manuscript to be published. This work was supported by the Key Research Program of Frontier Sciences, CAS, Grant No. ZDBS-LY-JSC038. Libo Zhang was supported by Youth Innovation Promotion Association, CAS (2020111), and Outstanding Youth Scientist Project of ISCAS.

References

Abdul-Mageed, M., Ungar, L.H., 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In: Barzilay, R., Kan, M. (Eds.), ACL. pp. 718–728.
Chang, D., Ding, Y., Xie, J., Bhunia, A.K., Li, X., Ma, Z., Wu, M., Guo, J., Song, Y., 2020. The Devil is in the channels: Mutual-channel loss for fine-grained image classification. IEEE Trans. Image Process. 29, 4683–4695.
Chang, D., Pang, K., Zheng, Y., Ma, Z., Song, Y., Guo, J., 2021. Your "Flamingo" is My "Bird": Fine-grained, or not. In: CVPR. pp. 11476–11485.
Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E., 2020a. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. ICML 2020, 13-18 July 2020, Virtual Event, In: Proceedings of Machine Learning Research, vol. 119, PMLR, pp. 1597–1607.
Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E., 2020b. Big self-supervised models are strong semi-supervised learners. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020. NeurIPS 2020, December 6-12, 2020, Virtual.
Chen, X., Xie, S., He, K., 2021. An empirical study of training self-supervised vision transformers, CoRR abs/2104.02057.
Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. In: CVPR. pp. 539–546.
Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., Le, Q.V., 2018. AutoAugment: Learning augmentation policies from data, CoRR abs/1805.09501.
Ding, Y., Ma, Z., Wen, S., Xie, J., Chang, D., Si, Z., Wu, M., Ling, H., 2021. AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. IEEE Trans. Image Process. 30, 2826–2836.
Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., Jiao, J., 2019. Selective sparse sampling for fine-grained image recognition. In: ICCV. pp. 6598–6607.
Du, R., Chang, D., Bhunia, A.K., Xie, J., Ma, Z., Song, Y., Guo, J., 2020. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In: ECCV. In: Lecture Notes in Computer Science, vol. 12365, pp. 153–168.
Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., Naik, N., 2018. Pairwise confusion for fine-grained visual classification. In: ECCV. In: Lecture Notes in Computer Science, vol. 11216, pp. 71–88.
Fu, J., Zheng, H., Mei, T., 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: CVPR. pp. 4476–4484.
Gao, Y., Han, X., Wang, X., Huang, W., Scott, M.R., 2020. Channel interaction networks for fine-grained image categorization. In: AAAI. pp. 10818–10825.
Ge, W., Lin, X., Yu, Y., 2019. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In: CVPR. pp. 3034–3043.
Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. In: AISTATS. In: JMLR Proceedings, vol. 15, JMLR.org, pp. 315–323.
Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent - A new approach to self-supervised learning. In: NeurIPS.
Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: CVPR. IEEE Computer Society, pp. 1735–1742.
Hanselmann, H., Ney, H., 2020. ELoPE: Fine-grained visual classification with efficient localization, pooling and embedding. In: WACV. IEEE, pp. 1236–1245.
He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B., 2020. Momentum contrast for unsupervised visual representation learning. In: CVPR. Computer Vision Foundation / IEEE, pp. 9726–9735.
He, X., Peng, Y., Zhao, J., 2019. Fast fine-grained image classification via weakly supervised discriminative localization. IEEE Trans. Circuits Syst. Video Technol. 29 (5), 1394–1407.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition, CoRR abs/1512.03385.
Hong, P., Wu, T., Wu, A., Han, X., Zheng, W., 2021. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In: CVPR. Computer Vision Foundation / IEEE, pp. 10513–10522.
Hu, T., Qi, H., 2019. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification, CoRR abs/1901.09891.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141.
Huang, S., Wang, X., Tao, D., 2020. SnapMix: Semantically proportional mixing for augmenting fine-grained data, CoRR abs/2012.04846.
Huang, S., Xu, Z., Tao, D., Zhang, Y., 2016. Part-stacked CNN for fine-grained visual categorization. In: CVPR. pp. 1173–1182.
Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. In: JMLR Workshop and Conference Proceedings, vol. 37, pp. 448–456.
Ji, R., Wen, L., Zhang, L., Du, D., Wu, Y., Zhao, C., Liu, X., Huang, F., 2020. Attention convolutional binary neural tree for fine-grained visual categorization. In: CVPR. pp. 10465–10474.
Leng, Q., Ye, M., Tian, Q., 2020. A survey of open-world person re-identification. IEEE Trans. Circuits Syst. Video Technol. 30 (4), 1092–1108.

Li, P., Xie, J., Wang, Q., Gao, Z., 2018. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In: CVPR. pp. 947–955.
Li, H., Zhang, X., Tian, Q., Xiong, H., 2020. Attribute mix: Semantic data augmentation for fine grained recognition. In: VCIP. pp. 243–246.
Lin, T., Maji, S., 2017. Improved bilinear pooling with CNNs. In: BMVC.
Lin, T., RoyChowdhury, A., Maji, S., 2015a. Bilinear CNN models for fine-grained visual recognition. In: ICCV. pp. 1449–1457.
Lin, D., Shen, X., Lu, C., Jia, J., 2015b. Deep LAC: Deep localization, alignment and classification for fine-grained recognition. In: CVPR. pp. 1666–1674.
Liu, C., Xie, H., Zha, Z., Ma, L., Yu, L., Zhang, Y., 2020. Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In: AAAI. pp. 11555–11562.
Luo, W., Yang, X., Mo, X., Lu, Y., Davis, L., Li, J., Yang, J., Lim, S., 2019. Cross-X learning for fine-grained visual categorization. In: ICCV. pp. 8241–8250.
Luo, W., Zhang, H., Li, J., Wei, X., 2020. Learning semantically enhanced feature for fine-grained image classification. IEEE Signal Process. Lett. 27, 1545–1549.
Melekhov, I., Kannala, J., Rahtu, E., 2016. Siamese network features for image matching. In: ICPR. pp. 378–383.
Park, J., Woo, S., Lee, J., Kweon, I.S., 2018. BAM: Bottleneck attention module. In: BMVC. p. 147.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch.
Rao, Y., Chen, G., Lu, J., Zhou, J., 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification, CoRR abs/2108.08728.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.
Shan, Y., Zhou, X., Liu, S., Zhang, Y., Huang, K., 2021. SiamFPN: A deep learning method for accurate and real-time maritime ship tracking. IEEE Trans. Circuits Syst. Video Technol. 31 (1), 315–325.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (Eds.), ICLR.
Sun, G., Cholakkal, H., Khan, S., Khan, F.S., Shao, L., 2020. Fine-grained recognition: Accounting for subtle differences between similar classes. In: AAAI. pp. 12047–12054.
Sun, M., Yuan, Y., Zhou, F., Ding, E., 2018. Multi-attention multi-class constraint for fine-grained image recognition. In: ECCV. In: Lecture Notes in Computer Science, vol. 11220, pp. 834–850.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826.
Tan, M., Le, Q.V., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: ICML. In: Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114.
Wang, Y., Morariu, V.I., Davis, L.S., Learning a discriminative filter bank within a CNN for fine-grained recognition. In: CVPR. pp. 4148–4157.
Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y., 2020a. Suppressing uncertainties for large-scale facial expression recognition. In: CVPR. pp. 6896–6905.
Wang, Z., Wang, S., Li, H., Dou, Z., Li, J., 2020b. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. In: AAAI. pp. 12289–12296.
Wang, Z., Wang, S., Yang, S., Li, H., Li, J., Li, Z., 2020c. Weakly supervised fine-grained image classification via guassian mixture model oriented discriminative learning. In: CVPR. pp. 9746–9755.
Woo, S., Park, J., Lee, J., Kweon, I.S., 2018. CBAM: Convolutional block attention module. In: ECCV. pp. 3–19.
Wu, Z., Xiong, Y., Yu, S.X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance-level discrimination, CoRR abs/1805.01978.
Xu, S., Chang, D., Xie, J., Ma, Z., 2021a. Grad-CAM guided channel-spatial attention module for fine-grained visual classification, CoRR abs/2101.09666.
Xu, P., Song, Z., Yin, Q., Song, Y.Z., Wang, L., 2021b. Deep self-supervised representation learning for free-hand sketch. IEEE Trans. Circuits Syst. Video Technol. 31 (4), 1503–1513.
Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L., 2018. Learning to navigate for fine-grained classification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), ECCV. In: Lecture Notes in Computer Science, vol. 11218, pp. 438–454.
Yu, C., Zhao, X., Zheng, Q., Zhang, P., You, X., 2018. Hierarchical bilinear pooling for fine-grained visual recognition. In: ECCV. In: Lecture Notes in Computer Science, vol. 11220, Springer, pp. 595–610.
Zhang, T., Chang, D., Ma, Z., Guo, J., 2021a. Progressive co-attention network for fine-grained visual classification, CoRR abs/2101.08527.
Zhang, N., Donahue, J., Girshick, R.B., Darrell, T., 2014. Part-based R-CNNs for fine-grained category detection. In: ECCV. In: Lecture Notes in Computer Science, vol. 8689, pp. 834–849.
Zhang, L., Huang, S., Liu, W., Tao, D., 2019. Learning a mixture of granularity-specific experts for fine-grained categorization. In: ICCV. pp. 8330–8339.
Zhang, F., Li, M., Zhai, G., Liu, Y., 2021b. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In: MMM. In: Lecture Notes in Computer Science, vol. 12572, pp. 136–147.
Zhang, G., Lu, X., Tan, J., Li, J., Zhang, Z., Li, Q., Hu, X., 2021c. RefineMask: Towards high-quality instance segmentation with fine-grained features. In: CVPR. pp. 6861–6869.
Zhang, H., Xu, T., Elhoseiny, M., Huang, X., Zhang, S., Elgammal, A.M., Metaxas, D.N., 2016. SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition. In: CVPR. pp. 1143–1152.
Zheng, H., Fu, J., Mei, T., Luo, J., 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In: ICCV. pp. 5219–5227.
Zheng, H., Fu, J., Zha, Z., Luo, J., 2019. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In: CVPR. pp. 5012–5021.
Zhuang, P., Wang, Y., Qiao, Y., 2020. Learning attentive pairwise interaction for fine-grained classification. In: AAAI. pp. 13130–13137.