Fine-Grained Object Classification via Self-Supervised Pose Alignment

Xuhui Yang[1], Yaowei Wang[1*], Ke Chen[2,1*], Yong Xu[1,2,3], Yonghong Tian[1]
[1]Peng Cheng Laboratory [2]South China University of Technology
[3]China Communication and Computer Network Laboratory of Guangdong
{yangxh, wangyw}@pcl.ac.cn, {chenk, yxu}@scut.edu.cn, tianyh@pcl.ac.cn

Abstract

Semantic patterns of fine-grained objects are determined by subtle appearance difference of local parts, which thus inspires a number of part-based methods. However, due to uncontrollable object poses in images, distinctive details carried by local regions can be spatially distributed or even self-occluded, leading to a large variation on object representation. For discounting pose variations, this paper proposes to learn a novel graph based object representation to reveal a global configuration of local parts for self-supervised pose alignment across classes, which is employed as an auxiliary feature regularization on a deep representation learning network. Moreover, a coarse-to-fine supervision together with the proposed pose-insensitive constraint on shallow-to-deep sub-networks encourages discriminative features in a curriculum learning manner. We evaluate our method on three popular fine-grained object classification benchmarks, consistently achieving the state-of-the-art performance. Source codes are available at https://github.com/yangxh11/P2P-Net.

1. Introduction

Semantic patterns in image-based object classification are determined by visual appearance and shape of object classes. The problem of classifying fine-grained objects is made more challenging due to the inherently subtle shape difference across the subordinate categories (*e.g.* bird breeds [34] and vehicle brands [18]). As a result, fine-grained classification relies on distinguished appearance details on local parts, which typically desires detection on those specific parts additionally using expensive part annotations [19, 38]. Learning a discriminative representation for fine-grained objects remains non-trivial in the context of deep learning. Generally, a good representation in fine-grained classification should not only be sensitive to the subtle detail changes that usually anchored on specific parts
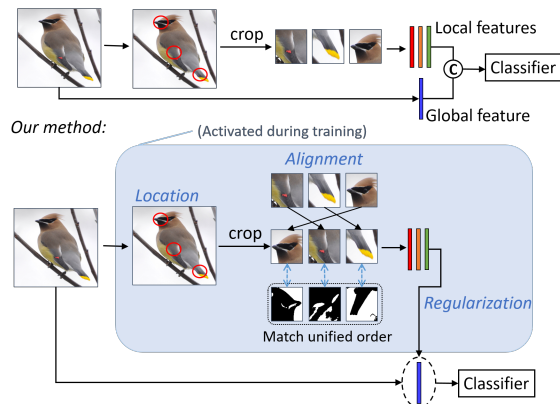
---

*Corresponding authors

Figure 1. Different from direct concatenation of local and global features in conventional part-based methods, the main idea of the proposed P2P-Net is to incorporate a pose-insensitive configuration of local distinctive details into object representation via utilizing self-supervised pose alignment as feature regularization to narrow intra-category variance and enlarge inter-category margin.

but also be invariant to the deformations of object parts and the changes of viewing angles.

On the one hand, a large number of deep approaches were devoted to the former challenge of capturing fine-grained details in local regions, which relied on detecting discriminative parts to extract local features to complement the global one [31, 41, 46]. On the other hand, another group of deep methods [17, 39, 45] avoided explicitly localizing object parts but tried to enhance the object representation's discrimination by applying bilinear pooling operation on feature vectors from two independent network streams. However, only a few works have paid attention to the challenge of feature inconsistency caused by the changes of object pose or viewing angle.

Similar to other pose estimation problems [1, 16], poses of fine-grained objects can approximately be described as a geometric configuration of discrete object parts. As a result, we argue that *reliable object parts localization* and *parts alignment* are essential to model object poses, which can be utilized to generate pose insensitive object representations

[33] so as to eliminate object pose variations. Evidently, explicitly regression on object poses as an auxiliary task is a straightforward solution, but in the context of fine-grained classification object poses typically suffers from inherent annotation ambiguities and a lack of sufficient samples. In view of this, we propose an unsupervised part-to-pose network (P2P-Net) which captures both part details and object poses to regularize the representation learning with no additional annotations of part position or pose information. Inspired by Feature Pyramid Networks (FPN) [20] that is originally designed for object detection, our P2P-Net distinguishes the confidently class-discriminative regions from region proposals in a weakly supervised learning manner. To incorporate the fine-grained appearance details on specific parts, the global object representation is regularized with local representations of the detected salient regions via minimizing distribution difference with the contrastive loss.

The major differences between the P2P-Net and the existing part-based methods, as shown in Fig. 1, lie in: 1) we align the parts in a self-supervised learning manner so as to discount the pose variations; 2) a feature regularization to strengthen representation discrimination is available only at training, *i.e.*, our method can discard the time consuming part branch (highlighted in light blue in Fig. 1) during testing. The effectiveness of the proposed self-supervised part-based pose alignment as feature regularization can be demonstrated in our experiments. Besides, we also propose curriculum supervision to capture discriminative details at coarse-to-fine scales for further improvement. The proposed P2P-Net consistently outperforms the state-of-the-art fine-grained classifiers on three popular benchmarks.

Our contributions can be summarized as follows:

- This paper proposes an end-to-end P2P-Net to incorporate discriminative features on confidential parts and then promotes the discrimination of object representations via pose-insensitive feature regularization.
- Technically, our P2P-Net designs an adaptive graph matching algorithm on confidential parts and achieves feature consistency against pose variations in an unsupervised learning manner.
- A generic curriculum supervision on image classification is introduced by designing label-smoothing based easy-to-hard supervision signals coupled with shallow-to-deep sub-networks.
- Experimental results on multiple public benchmarks demonstrate that the proposed method can achieve a new state-of-the-art performance on the problem of fine-grained image classification .

2. Related Work

In the past decades, deep learning has made remarkable progresses in the application of fine-grained image classi-

fication [2, 3]. These methods can be categorized into two main groups – part-free and attention/part-based methods.

The former group of algorithms emerges a trend that a part of fine-grained recognition methods seek to boost the backbone's recognition capability by either weighting features using attention maps or employing augmented samples as training data. Lin *et al*. proposed B-CNN [21], which uses two-stream architectures to simultaneously model the location and appearance of details. Zheng *et al*. [44] first explored to spot salient peaks by clustering spatially-correlated channels, and then in [46] used a trilinear attention module to model the inter-channel relationships and then distorted the image via attention based pixel sampling. Ding *et al*. [6] performed a selective sparse sampling operation to discover diverse and fine-grained details. Chen *et al*. [4] first proposed an image splitting operation for fine-grained classification by partitioning the image into local patches, which are randomly shuffled to reconstruct a new image for training, and thus force the model to concern on local details rather than its global configuration. Du *et al*. [7] took one step further in applying a multi-scale jigsaw puzzle generator to capture cross-granularity information with a progressive training strategy. Inspired by these part-free methods, we propose a curriculum training scheme generic to all image classification tasks, which can improve model generalization by using easy-to-hard supervision signals coupled with shallow-to-deep sub-networks.

It is observed that capturing the subtle differences in discriminative parts plays an important role to distinguish similar categories. In light of this, a number of attention/part-based methods have been encouraged to detect and learn discriminative features according to the attention/score maps or region proposals [10, 13, 40]. Early works detected parts based on densely part annotations [19, 43] in a fully supervised learning style. Recent studies relaxed the desire of expensive part annotations and localized informative regions in a weakly-supervised way, *i.e*., only given category labels. In [9], a recurrent attention convolutional neural network is proposed to iteratively zoom in local discriminative regions and introduced an inter-scale ranking loss between multi-scale regions to reinforce feature learning. Wang *et al*. [37] proposed a discriminative region grouping sub-network to discover discriminative regions, while Yang *et al*. [41] introduced a weakly supervised feature pyramid network into fine-grained classification to select top distinguishable parts. Other works further investigated the spatial or context relations among object parts [12, 28, 37]. Those part-based methods are robust on locating discriminative parts, but usually fused features of parts and image with a simple concatenation without further exploiting the latent correlation across parts features [14, 42], and thereby inevitably encountered the feature alignment problem. Our approach follows the same way of weakly supervised part-
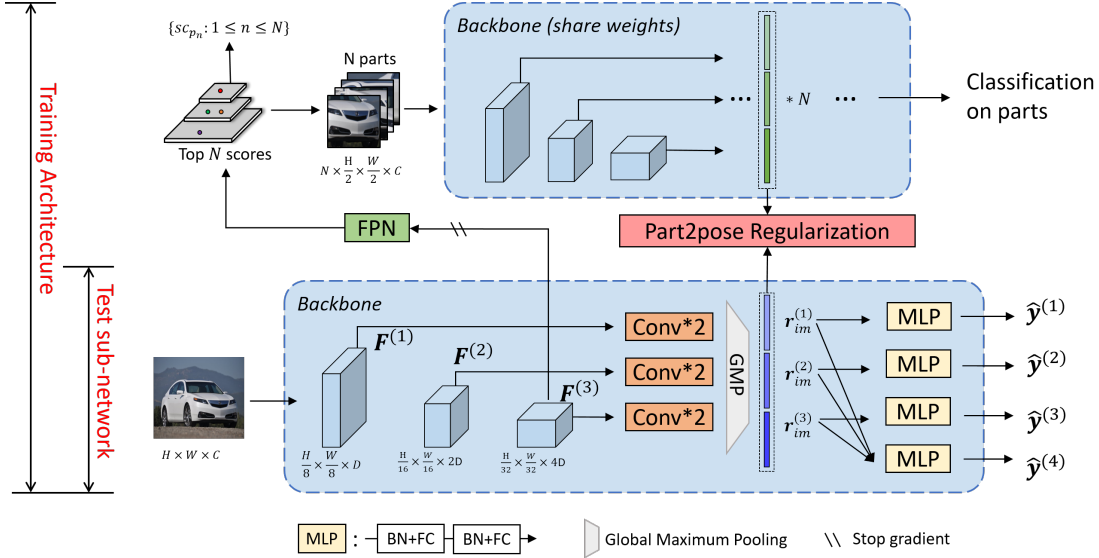
Figure 2. Pipeline of the proposed P2P-Net. Both global feature encoding backbone (bottom) and part2pose feature regularization (top) are utilized for training while only the bottom one is activated during testing. Network parameters of the feature encoders (highlighted in light blue) in both upper and bottom branches are shared. $\boldsymbol{F}^{(s)}$ denotes an intermediate feature map from convolutional layer at specific depth, where $1 \leqslant s \leqslant S$ and $S = 3$, while $\hat{\boldsymbol{y}}^{(s)}$ denotes classification predictions made on representations $\boldsymbol{r}_{im}^{(s)}$. Note that, $\hat{\boldsymbol{y}}^{(S+1)}$ is made on the concatenation of $\{\boldsymbol{r}_{im}^{(s)}\}$. For the detected parts, sc_{p_n} denotes the score of the corresponding part p_n, where $1 \leqslant n \leqslant N$. Detailed structure of the feature pyramid network (FPN) block and part2pose regularization block are shown in Fig. 3.

based approaches, but we go three steps further to learn pose insensitive representations: 1) aligning parts according to their correlations; 2) enhancing the discrimination of object representation with detected parts via feature regularization instead of feature concatenation; 3) encouraging discriminative features in a curriculum learning manner.

3. Methodology

For the problem of fine-grained image classification with $\mathbb{X}$ and $\mathbb{Y}$ denoting the input and output space, given L training samples $\{\boldsymbol{I}, y\}^L$, where $\boldsymbol{I} \in \mathbb{X}$ and $y \in \mathbb{Y}$ denote one visual observation (*i.e.* image) and its corresponding category label, the goal is to learn a mapping function $\Phi : \mathbb{X} \to \mathbb{Y}$ that correctly classifies images into one of the $|\mathbb{Y}| = K$ categories. In this section, we present the whole architecture of our P2P-Net, as shown in Fig. 2, which is composed of three parts: 1) curriculum supervision for complementary representation learning (see Sec. 3.1); 2) contrastive feature regularization for strengthening object representation's discrimination (see Sec. 3.2); 3) unsupervised graph matching method for part alignment (see Sec. 3.3). The testing sub-network is only a part of the whole architecture, which greatly reduces the computational complexity during inference since only the bottom backbone is activated.

3.1. Curriculum Supervision on Backbone Network

In the problem of fine-grained image classification, a typical ResNet [11] (*e.g.*, ResNet34 or ResNet50) is

adopted as backbone in this paper. As Fig. 2 shows, several intermediate feature maps of the backbone at different depths are fed into different convolutional blocks respectively. They are noted as $\boldsymbol{F}^{(s)}$, where $s \in \{1, ..., S\}$ stands for different stages and s is proportional to the sub-network depths. Each block is followed by a global maximum pooling (GMP) layer to obtain image representations $\{\boldsymbol{r}_{im}^{(s)}\}$. At the final layer we stack several independent multi-layer perceptrons (MLP) for classification separately. As a result, there are in total $S + 1$ predictions for one image, which will be aggregated for the final prediction, whose detail is introduced in Sec. 3.4. Note that, the above design of training sub-networks is generic and can readily be applied to existing classification models.

Beyond employing a conventional ResNet as backbone, we introduce a curriculum training scheme encouraged by [7], which is based on the observation that deeper network has stronger learning capability and tends to correctly distinguishes more challenging samples. To improve model generalization and encourage diversity of representations from specific layers, we introduce a curriculum training strategy by employing a soft supervision method based on label smoothing [26]. The cross-granularity representations can then be utilized either independently or jointly to make better predictions. We modify the one-hot vector as follows:

$$\boldsymbol{y}_\alpha[t] = \begin{cases} \alpha, & t = y \\ \frac{1-\alpha}{K}, & t \neq y \end{cases}, \tag{1}$$

where α denotes a smoothing factor between 0 and 1, t represents the element index of the label vector $\boldsymbol{y} \in \mathbb{R}^K$. Note that, α controls the magnitude of the ground truth class in the new target $\boldsymbol{y}_\alpha$, which can thus be used to construct easy-to-hard curriculum targets coupled with shallow-to-deep feature encoders. As a result, predictions $\{\hat{\boldsymbol{y}}^{(s)}\}$ on representations from different layers or their combination are supervised with different labels during training, whose loss function can be written in the following form:

$$
\begin{aligned}
\ell_{sce}(\hat{\boldsymbol{y}}^{(s)}, y, \alpha^{(s)}) &= \ell_{ce}(\hat{\boldsymbol{y}}^{(s)}, \boldsymbol{y}_{\alpha^{(s)}}) \\
&= \sum_{t=0}^{K-1} -\boldsymbol{y}_{\alpha^{(s)}}[t] \log(\hat{\boldsymbol{y}}^{(s)}[t]),
\end{aligned}
\tag{2}
$$

where $\ell_{sce}(\cdot)$ denotes a smooth cross-entropy loss which takes one more smoothing factor than the normal cross-entropy loss [27]. Since $S + 1$ predictions are available in our P2P-Net, the overall classification loss of each image sample can be written as follows:

$$
L_{im}^{cls} = \sum_{s=1}^{S+1} \ell_{sce}(\hat{\boldsymbol{y}}^{(s)}, y, \alpha^{(s)}).
\tag{3}
$$

Generally, deeper sub-networks have more capacity to learn a mapping from more challenging samples. Therefore, we gradually increase $\alpha^{(s)}$ from a value greater than $\frac{1}{K}$ to 1. As s getting larger, the supervision label $\boldsymbol{y}_{\alpha^{(s)}}$ approaches more closer to the one-hot target code, which indicates that the corresponding sub-network should make more confident prediction with deeper network architecture.

3.2. Contrastive Feature Regularization

Locating discriminative parts is an effective way to mitigate the challenge of inter-class similarity. Typically, in existing works [36, 44], representations of parts are concatenated or fused to generate a new representation (to complement the global object representation) for classification. In this paper, we treat the parts localization problem as an object detection task and propose a feature regularization on representations between local parts and global object to enforce incorporating fine-grained details from distinctive parts into image representation.

Weakly-Supervised Part Localization – As shown in Fig. 3, the backbone's last feature block is followed by a feature pyramid network (FPN), which generates a pyramid of score maps of different spatial sizes, *e.g.*, 14×14, 7×7 and 4×4, by following the same setting as [41]. Each score element in the maps corresponds to a predefined image patch with a fixed size. Note that, these patches having a unique size can be overlapped with others. Moreover, elements in a larger map (*e.g.*, 14×14) correspond to smaller image patches compared to the ones in a smaller map (*e.g.*, 7×7). Under the assumption that the scores indicate the
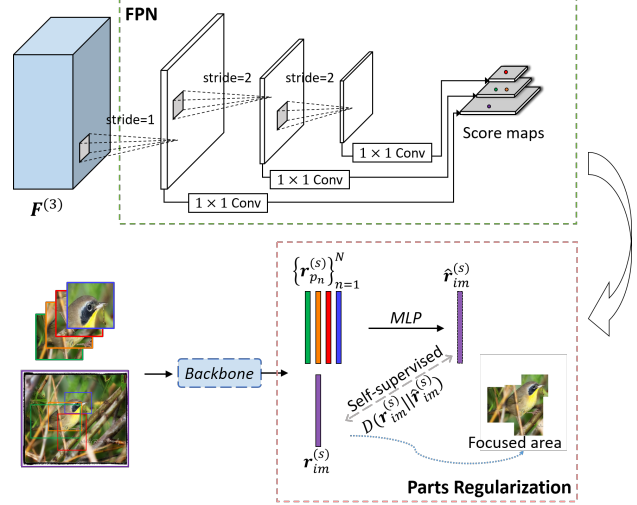

Figure 3. Illustration of the FPN and the part-based contrastive feature regularization. As shown in Fig. 2, the input of FPN is $F^{(3)}$ from the backbone while the outputs are score maps of different scales. We apply the non-maximum suppression (NMS) to the score maps and select the top N discriminative parts, features from which are then used to regularize the object representation learning in a self-supervised contrastive learning manner.

discrimination degree of image patches, when selecting the top N parts of highest scores, we apply the non-maximum suppression (NMS) to eliminate parts with large Intersection over Union (IoU).

Classification and Ranking Loss on Detected Parts – According to the top N scores' indexes, we crop the corresponding image patches from the input image, which contain distinctive details of local object parts. For computational efficiency, these N parts are then resized into 224×224 (half the spatial size of the original image) and fed into the backbone with shared weights. The curriculum supervision described in Sec. 3.1 on the whole image is repeated for each resized patches (parts). We define a part's feature as $\boldsymbol{r}_{p_n} = [\boldsymbol{r}_{p_n}^{(1)}; \boldsymbol{r}_{p_n}^{(2)}; \ldots; \boldsymbol{r}_{p_n}^{(S)}]$, where $1 \leqslant n \leqslant N$. Similar to Eq. (3), the classification loss on the n-th part can be depicted as

$$
L_{p_n} = \sum_{s=1}^{S+1} \ell_{sce}(\hat{\boldsymbol{y}}_{p_n}^{(s)}, y, \alpha^{(s)}),
\tag{4}
$$

and the total classification loss on all parts is

$$
L_{parts}^{cls} = \sum_{n=1}^{N} L_{p_n}.
\tag{5}
$$

Moreover, given top N parts' classification losses $\{L_{p_n}\}^N$ and the corresponding scores $\{sc_{p_n}\}^N$, we further argue that the prediction losses should be consistent with the confidence scores of image parts. Given part indexes n

and n', if $L_{p_n} < L_{p_{n'}}$, then p_n should has a greater score than $p_{n'}$, so $sc_{p_n} > sc_{p_{n'}}$ is preferred in such a situation. As a result, we introduce an extra ranking loss based on a summation of conditional hinge losses [29]:

$$
\begin{aligned}
L^{rank} &= \sum_{n=1}^{N} \sum_{n'=1}^{N} \ell_{hg}(L_{p_n}, L_{p_{n'}}) * c_{nn'} \\
&= \sum_{n=1}^{N} \sum_{n'=1}^{N} max(0, L_{p_n} - L_{p_{n'}} + \delta) * c_{nn'},
\end{aligned} \tag{6}
$$

with the indicator $c_{nn'}$ is defined as

$$
c_{nn'} = \begin{cases} 1, & sc_{p_n} > sc_{p_{n'}} \\ 0, & sc_{p_n} \leqslant sc_{p_{n'}} \end{cases}, \tag{7}
$$

where sc_{p_n} denotes the score of part p_n and margin hyper-parameter δ is fixed as 1 in our paper. The loss encourages sc_{p_n} to be greater than $sc_{p_{n'}}$ if L_{p_n} has a relative smaller value than $L_{p_{n'}}$. Ideally, by minimizing the ranking loss, the scores $\{sc_{p_n}\}$ and the parts classification losses $\{L_{p_n}\}$ should change in an opposite way. As a result, such a design can improve reliability of discriminative part detection.

Contrastive Loss for Feature Regularization – Instead of making classification via concatenating local parts' and global image's features, we propose a feature regularization to constraint the object representation learning in a contrastive learning manner. Given image representation $r_{im} = [r_{im}^{(1)}; r_{im}^{(2)}; ...; r_{im}^{(S)}]$ and part representation $r_{p_n} = [r_{p_n}^{(1)}; r_{p_n}^{(2)}; ...; r_{p_n}^{(S)}]$, we regularize each stage's representation with a contrastive loss as follows:

$$
L^{reg} = \sum_{s=1}^{S} \ell_{kl}(r_{im}^{(s)}, \phi([r_{p_1}^{(s)}; r_{p_2}^{(s)}; ...; r_{p_N}^{(s)}])), \tag{8}
$$

where ℓ_{kl} is the Kullback-Leibler divergence function, $\phi(\cdot)$ is an approximation function to be optimized, which we adopt a 2-Layer MLP to model it in the experiments. Such a regularization loss can enforce object representation learning branch to focus on discriminative details from specific local regions, which is visually shown in Fig. 3. Through this way we can further filter out redundant misleading information in each $r_{im}^{(s)}$ so as to improve discrimination of object representation.

3.3. Graph Matching for Part Alignment

Although the top N discriminative parts have been located through the Sec. 3.2, it is found out that the configuration of these parts is not consistent, *i.e.*, they are not strictly aligned (See Sec. 4.4). As we only use category labels for supervision to discover discriminative parts, the exact semantic class of parts (*e.g.*, head, body or tail) remains unknown. Therefore, simply concatenating the spotted parts'
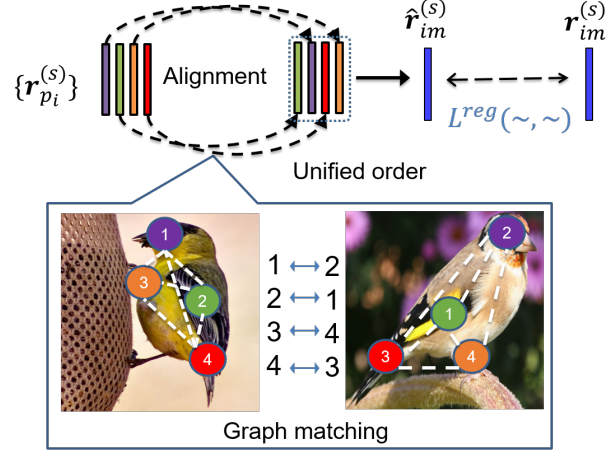

Figure 4. Graph matching for part alignment. For illustrative purpose, only one stage's representations are shown in the figure. Located parts are regarded as ordered nodes according the discrimination scores. We re-order the parts by applying graph matching on the parts correlation matrix according to its global configuration, whose features are then employed to regularized the learning of global image representation.

features in an arbitrary order for regularization may raise the feature inconsistency problem. Fortunately, for any specific object category, its discriminative appearance usually appears on a limited size of local parts. Consequently, we propose an unsupervised graph matching method to sort the found parts in a unified order based on a basic assumption that the correlation between top N parts is similar across images.

The concept of our graph matching is illustrated in Fig. 4. Intuitively, it is supposed that top N parts of the bird image on the left-hand side is ordered as $<head, wing, foot, tail>$ (*i.e.* $<1, 2, 3, 4>$) according to their scores $\{sc_{p_n}\}^4$, while the unified order maintained by our method may be $<wing, head, tail, foot>$. The goal of our graph matching algorithm is to resort the order of detected object parts from $<1, 2, 3, 4>$ to $<2, 1, 4, 3>$ in the unified order for feature consistency. In this way, resorted parts features discounting pose variations are used to constrain object representation by contrastive learning, which is formulated in Eq. (8).

Technically, since we can not identify object's distinct parts directly, we maintain a unified correlation matrix to model latent relations in-between parts. The entry of the correlation matrix is given as follows:

$$
M_{ij} = <r_{p_i}, r_{p_j}>, \tag{9}
$$

where M_{ij} denotes the relation score between part p_i and part p_j. Given a new image sample, we compute its parts correlation matrix, denoted as M', of each possible permutation of parts. Then the M' has the largest matching degree with the reference matrix M is considered as the best

alignment. The formulation can be simplified as

$$\hat{M} = \underset{M'}{\operatorname{argmax}} \, vec(M')^T vec(M), \qquad (10)$$

where the matching degree can be simply measured by summarizing the element-wise product of vectorization of two matrices.

This is exactly a graph matching problem considering edges (relations) similarity but not nodes (parts) similarity. In this paper, as parts also contain distinct classes information, it is not suitable to add nodes (parts) similarity even if they belong to the same type (*e.g.*, head). Note that we do not consider the mismatch situation in the experiments. The permutation with the max matching degree is selected as the correct order and the algorithm returns the resorted parts representations. Computation about graph matching is efficient because the size of parts $N \leqslant 5$ in this paper although there are $N!$ permutations in total. Finally, for self-correction, an online-updating scheme is adopted to refine the parts centers and the reference matrix M. The parts centers are updated by weighting the new sample's resorted features and the old samples' parts features, which follows the basic rule that older samples are of smaller weights.

3.4. Training and Inference

Given the aforementioned key components, our network can be effectively trained in an end-to-end manner. Specifically, the network loss is given as a summation of aforementioned losses with only one trade-off parameter β:

$$L = L_{im}^{cls} + L_{parts}^{cls} + L^{rank} + \beta \cdot L^{reg}, \qquad (11)$$

where L_{im}^{cls}, L_{parts}^{cls}, L^{rank}, and $\cdot L^{reg}$ are the classification loss on image, the classification loss on discriminative parts, the ranking loss on keeping consistency of scores for parts, and the regularization loss of image representation, respectively. In testing, for the stable generalization performance in an ensemble learning, we combine multiple prediction outputs with equal weights as follows

$$\hat{y}^{(final)} = \sum_{s=1}^{S+1} \hat{y}^{(s)}, \qquad (12)$$

where the maximum entry in the vector $\hat{y}^{(final)}$ corresponds to the class prediction. More importantly, most of the proposed components are only activated during training. In other words, during testing, only the backbone branch (*i.e.*, the bottom one in Fig. 2) with curriculum learning is utilized to make predictions. In summary, computational cost of our P2P-Net is slightly greater than its original backbone (*e.g.* ResNet) at inference.

4. Experiments

4.1. Datasets and Settings

We evaluate the proposed method on widely used benchmarks for fine-grained visual classification, including Caltech-UCSD Birds (CUB) [34], the Stanford Cars (CAR) [18] and the FGVC Aircraft (AIR) [25]. The **CUB** is a dataset with 11,788 images from 200 bird species. Data split is fixed with 5,994 images for training and 5,794 images for testing. The **CAR** dataset consists of 16,185 images of 196 classes of cars, whose data is divided into 8,144 training images and 8,041 testing images, where each class has been split roughly in a 50-50 split. The **AIR** benchmark contains 10,000 images of 100 aircraft variants, among which only 3,333 images for testing. We follow the same image and part size as recent work [41], *i.e.*, raw images are first resized into 550×550 followed by random horizontal flipping and cropping (center cropping used in testing) into 448×448. In the experiments, only class labels are available without any additional prior knowledge.

We evaluate the P2P-Net on the ResNet34 and the ResNet50 backbones, with weights pre-trained on the ImageNet dataset [5]. Models are trained for 300 epochs with a mini-batch size of 16. Learning rate follows the cosine annealing schedule [23] with initial learning rate setting as 0.002. Particularly, the learning rate of the backbone is set as one-tenth as other layers to make training more stable. The number of parts (N) and intermediate feature maps (S) are 4 and 3 as default values. The loss weight β in Eq. (11) is empirically set as 0.1. For the smoothing factors, we simply set $\{\alpha^{(s)}\}$ as $\{0.7, 0.8, 0.9, 1.0\}$ in an ascending order for the fact that larger α indicates higher confidence of class predictions.

4.2. Comparison with State-of-the-arts

Experimental results about comparative evaluation with recent fine-grained classifiers on the CUB_200_2011, the Stanford Cars, and the FGVC-Aircraft are illustrated in Tab. 1. As the table shows, with the same backbone ResNet50, the proposed P2P-Net outperforms the state-of-the-art methods on three widely-used benchmarks by more than 0.4% at average (+0.6% on CUB, +0.3% on CAR, and +0.4% on AIR), whose performance gain can be credited to our proposed components. What's more, our method also can achieve competitive performance to existing methods even with a shallow backbone – ResNet34.

4.3. Ablation Studies

We conduct ablation studies to validate the effectiveness on the key components of our method, including curriculum supervision, part-to-pose feature regularization, and unsupervised part alignment respectively. The baseline method consists of a backbone ResNet50 and a MLP. Based on the

Method	Backbone	Accuracy (%)		
		CUB	CAR	AIR
B-CNN [21]	VGG	84.1	91.3	84.1
RA-CNN [9]	VGG19	85.3	92.5	88.2
MA-CNN [44]		86.5	92.8	89.9
FCAN [22]		84.7	93.1	-
MAMC [32]		86.3	93.0	-
DFL-CNN [35]		87.4	93.1	91.7
NTS-Net [41]		87.5	93.9	91.4
DCL [4]		87.8	94.5	93.0
TASN [46]		87.9	93.8	-
Cross-X [24]	ResNet50	87.7	94.6	92.6
S3N [6]		88.5	94.7	92.8
LIO [47]		88.0	94.5	92.7
BNT [15]		88.1	94.6	92.4
ASD [31]		88.6	94.9	93.5
DF-GMM [36]		88.8	94.8	93.8
PMG [8]		89.6	95.1	93.4
API-Net [48]	ResNet101	88.6	94.9	93.4
API-Net [48]	DenseNet-161	90.0	95.3	93.9
P2P-Net (ours)	ResNet34	89.5	94.9	92.6
P2P-Net (ours)	ResNet50	**90.2**	**95.4**	**94.2**

Table 1. Comparison with the state-of-the-art methods.

Method	Accuracy (%)		
	CUB	CAR	AIR
(a) Baseline	85.5	92.7	90.3
(b) Baseline+CS	88.4	94.9	93.8
(c) Baseline+FR (w/o UPA)	89.0	94.8	92.0
(d) Baseline+FR (w/ UPA)	89.0	95.0	92.5
(e) Baseline+FC	88.4	94.7	93.8
(f) Baseline+CS+FR (w/o UPA)	90.0	95.0	93.9
(g) Baseline+CS+FR (w/ UPA)	**90.2**	**95.4**	**94.2**

Table 2. Ablation study on the proposed components with the baseline ResNet50. CS: curriculum supervision, FR: feature regularization, UPA: unsupervised part alignment, FC: feature concatenation.

baseline, we evaluated different combinations of the proposed components, whose results are presented in Tab. 2.

Curriculum Supervision (CS) – In comparison with (a) and (b), the CS module leads to large performance gain varying from +2.1% to +3.5% in classification accuracy on three datasets. Such results can be explained in two aspects: 1) by making predictions at different depths on the network, complementary information of multi-granularity can be fused to capture discriminative object features; 2) by connecting more layers to the output, parameters in shallower layers become easier to optimize. We further conduct contrastive experiments on the CUB between using $\{\alpha^{(s)}\}$ and using one-hot labels (no label smoothing). The accuracy of different predictions of the former one

Method	RMSE		
	CUB	CAR	AIR
Baseline	0.501	0.354	0.399
Baseline+FC	0.268	0.213	0.354
Baseline+FR (w/ UPA)	**0.213**	**0.179**	**0.263**

Table 3. RMSE of learned representations of different methods.

is $\{81.7, 87.3, 85.8, 87.8, 88.4\}$, while the latter one's is $\{82.1, 87.0, 85.6, 87.3, 88.2\}$. Such a result proves the proposed smoothing scheme's effectiveness.

Feature Regularization (FR) – Comparing variants (c-d) with (a), the FR module provides significant improvement on performance, while it shows smaller gain compared to the CS component on the AIR. Since a global contour is an important cue for classifying big objects like aircrafts, when using local distinctive information to regularize the object representation learning, the result may not be better. To verify our motivation, we report comparative evaluation with an additional competitor (e), whose network structure is almost the same as methods (c-d). The mere difference lies in that, in (e) we use concatenation of features of local parts and a global image for classification, rather than the proposed feature regularization. As expected, (e) outperforms (c-d) on the AIR but is slightly worse on other datasets. Our explanation is that the global contour information of aircraft classes is well preserved when using feature concatenation, instead of regularizing image feature with relatively small parts. It is worth emphasizing again that, using FR is more computationally efficient owing to discarding the part-to-pose regularization branch during testing.

Unsupervised Parts Alignment (UPA) – Lastly, the method integrating FR with UPA outperforms its competitor without using UPA (See methods (c-d) and (f-g) in the table). These results demonstrate that the effectiveness of our unsupervised graph matching for part alignment, which can be a supplement to the part-based feature regularization. Besides, we also measure the compactness of image representations. In details, for each class, we compute root mean square error (RMSE) of learned representations and reported the average in Tab. 3, where P2P-Net (*i.e.*, the bottom one) achieves the least value. The second method in the table can be viewed as a typical feature concatenation of the conventional part-based methods. The result shows that the object representations learned with FR and UPA are more concentrated around their class centers in feature space. As we know that, objects of the same class are of varied poses, the result also demonstrates that our method are able to align detected discriminative parts to narrow the intra-class variance caused by pose changes.

Size of Parts – Result of using different parts number (N) on the CUB are shown in Tab. 4. Larger values may not bring a significant improvement, so $N = 4$ is preferred.

Parts number (N)	2	3	4	5	6
CUB	88.1	89.9	**90.2**	**90.2**	90.0

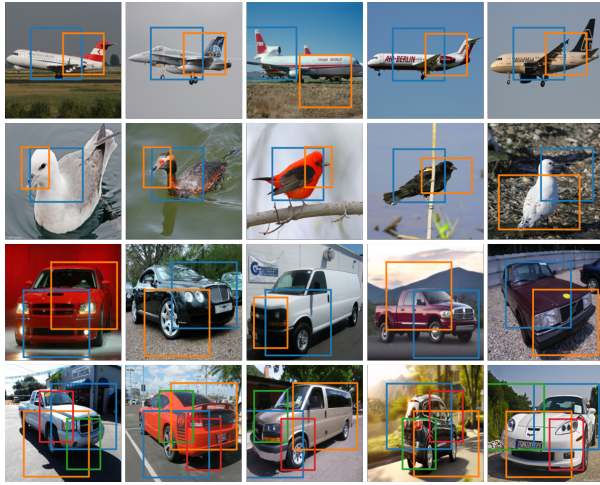Table 4. Influence of the number of discriminative parts.

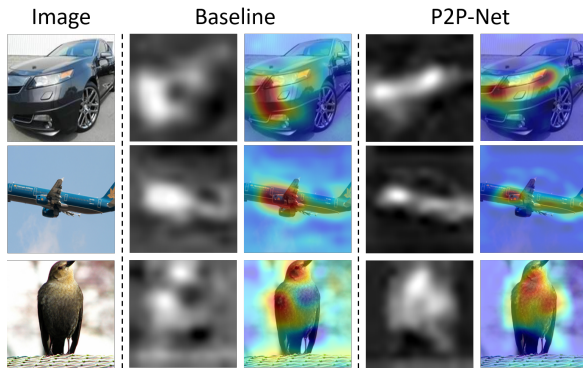Figure 5. Discriminative parts detected by our P2P-Net.

Figure 6. Class activation maps of some test samples.

4.4. Visualizations

Parts Location – The discriminative parts located by our P2P-Net are visualized in Fig. 5. Only the top 2 parts' bounding boxes are shown from the 1st to the 3rd row, while top 4 salient parts are highlighted in the 4th row. As shown in the figure, the top 2 discriminative parts share visual similarity on localized parts, although they are detected in weakly-supervised learning manner. Specifically, for aircraft samples, salient regions tend to locate in the body and tail; for bird breeds, they usually concern on birds' head and body; and for cars, the front and body of vehicles contain discriminative details. This phenomenon is an important prerequisite for the proposed graph matching method.

Class Activation Map – We also apply Grad-CAM [30] to the last convolutional layer to show intuitive visualization. Both gray scale and merged color activation images are il-
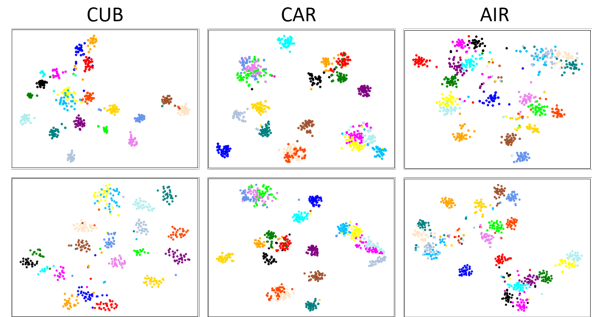

Figure 7. A T-SNE plot of learned representations on three datasets. First row: the baseline model; second row: baseline+FR (w/ UPA).

lustrated in Fig. 6. Compared to the baseline, our P2P-Net has less activation on the background and is more concentrated on the discriminative regions of objects.Evidently, it is verified that our P2P-Net is able to extract discriminative information from distinctive regions and eliminate the impact of noisy backgrounds.

Feature Visualization – In Fig. 7, we draw the t-SNE scatter plot from the learned high dimensional features. On each dataset, we randomly selected twenty classes for visualization. As the t-SNE plot shows, after applying parts' feature regularization, image representations exhibit higher intra-classes variations and favor for enlarging inter-class margins, especially on the CUB and AIR datasets.

5. Conclusion

This paper proposes a novel feature regularization scheme about learning pose-insensitive discriminative representations for fine-grained classification. Our P2P-Net not only utilizes the located discriminative parts to promote discrimination of image representations, but also introduces graph matching for part alignment for robustness against pose variations. Besides, a curriculum supervision strategy is verified to further boost the performance without paying any price of extra annotations. Extensive experiments on well-known benchmarks against the state-of-the-art methods and ablation analysis have demonstrated the effectiveness of our method. In addition, the contributed scheme might be under adversarial attacks, causing total failure of the whole perception system, which encourages researchers and safety engineers to mitigate these risks.

Acknowledgements

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019. 1

[2] Dongliang Chang, Yifeng Ding, Jiyang Xie, Ayan Kumar Bhunia, Xiaoxu Li, Zhanyu Ma, Ming Wu, Jun Guo, and Yi-Zhe Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 29:4683–4695, 2020. 2

[3] Tianshui Chen, Wenxi Wu, Yuefang Gao, Le Dong, Xiaonan Luo, and Liang Lin. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 2023–2031, 2018. 2

[4] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019. 2, 7

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 6

[6] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6599–6608, 2019. 2, 7

[7] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, pages 153–168. Springer, 2020. 2, 3

[8] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, pages 153–168. Springer, 2020. 7

[9] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4438–4446, 2017. 2, 7

[10] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3034–3043, 2019. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[12] Xiangteng He and Yuxin Peng. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 2

[13] Xiangteng He, Yuxin Peng, and Junjie Zhao. Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization. *International Journal of Computer Vision*, 127(9):1235–1255, 2019. 2

[14] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint arXiv:1901.09891*, 2019. 2

[15] Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, and Feiyue Huang. Attention convolutional binary neural tree for fine-grained visual categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10468–10477, 2020. 7

[16] Lei Ke, Shichao Li, Yanan Sun, Yu-Wing Tai, and Chi-Keung Tang. Gsnet: Joint vehicle pose and shape reconstruction with geometrical and scene-aware supervision. In *European Conference on Computer Vision*, pages 515–532. Springer, 2020. 1

[17] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 365–374, 2017. 1

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 1, 6

[19] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1666–1674, 2015. 1, 2

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 2

[21] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015. 2, 7

[22] Xiao Liu, Tian Xia, Jian Wang, and Yuanqing Lin. Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. *CoRR*, abs/1603.06765, 2016. 7

[23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[24] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S Davis, Jun Li, Jian Yang, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8242–8251, 2019. 7

[25] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[26] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019. 3

[27] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 4

[28] Lei Qi, Xiaoqiang Lu, and Xuelong Li. Exploiting spatial relation for fine-grained image classification. *Pattern Recognition*, 91:47–55, 2019. 2

[29] Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004. 5

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 8

[31] Guolei Sun, Hisham Cholakkal, Salman Khan, Fahad Khan, and Ling Shao. Fine-grained recognition: Accounting for subtle differences between similar classes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12047–12054, 2020. 1, 7

[32] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision*, pages 805–821, 2018. 7

[33] Luming Tang, Davis Wertheimer, and Bharath Hariharan. Revisiting pose-normalization for fine-grained few-shot recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14352–14361, 2020. 2

[34] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1, 6

[35] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4148–4157, 2018. 7

[36] Zhihui Wang, Shijie Wang, Shuhui Yang, Haojie Li, Jianjun Li, and Zezhou Li. Weakly supervised fine-grained image classification via guassian mixture model oriented discriminative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9749–9758, 2020. 4, 7

[37] Zhihui Wang, Shijie Wang, Pengbo Zhang, Haojie Li, Wei Zhong, and Jianjun Li. Weakly supervised fine-grained image classification via correlation-guided discriminative learning. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1851–1860, 2019. 2

[38] Xiu-Shen Wei, Chen-Wei Xie, Jianxin Wu, and Chunhua Shen. Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*, 76:704–714, 2018. 1

[39] Lin Wu, Yang Wang, Xue Li, and Junbin Gao. Deep attention-based spatially recursive networks for fine-grained visual recognition. *IEEE Transactions on Cybernetics*, 49(5):1791–1802, 2018. 1

[40] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015. 2

[41] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision*, pages 420–435, 2018. 1, 2, 4, 6, 7

[42] Lianbo Zhang, Shaoli Huang, Wei Liu, and Dacheng Tao. Learning a mixture of granularity-specific experts for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8331–8340, 2019. 2

[43] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014. 2

[44] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5209–5217, 2017. 2, 4, 7

[45] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning deep bilinear transformation for fine-grained image representation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 4277–4286. Curran Associates, Inc., 2019. 1

[46] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5012–5021, 2019. 1, 2, 7

[47] Mohan Zhou, Yalong Bai, Wei Zhang, Tiejun Zhao, and Tao Mei. Look-into-object: Self-supervised structure modeling for object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11774–11783, 2020. 7

[48] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13130–13137, 2020. 7