

# Part-Aware Fine-grained Object Categorization using Weakly Supervised Part Detection Network

Yabin Zhang, Kui Jia, and Zhixin Wang

**Abstract**—Fine-grained object categorization aims for distinguishing objects of subordinate categories that belong to the same entry-level object category. It is a rapidly developing subfield in multimedia content analysis. The task is challenging due to the facts that (1) training images with ground-truth labels are difficult to obtain, and (2) variations among different subordinate categories are subtle. It is well established that characterizing features of different subordinate categories are located on local parts of object instances. However, manually annotating object parts requires expertise, which is also difficult to generalize to new fine-grained categorization tasks. In this work, we propose a Weakly Supervised Part Detection Network (PartNet) that is able to detect discriminative local parts for the use of fine-grained categorization. A vanilla PartNet builds on top of a base subnetwork two parallel streams of upper network layers, which respectively compute scores of classification probabilities (over subordinate categories) and detection probabilities (over a specified number of discriminative part detectors) for local regions of interest (RoIs). The image-level prediction is obtained by aggregating element-wise products of these region-level probabilities, and meanwhile diverse part detectors can be learned in an end-to-end fashion under the image-level supervision. To generate a diverse set of RoIs as inputs of PartNet, we propose a simple Discretized Part Proposals module (DPP) that directly targets for proposing candidates of discriminative local parts, with no bridging via object-level proposals. Experiments on benchmark datasets of CUB-200-2011, Oxford Flower 102 and Oxford-IIIT Pet show the efficacy of our proposed method for both discriminative part detection and fine-grained categorization. In particular, we achieve the new state-of-the-art performance on CUB-200-2011 and Oxford-IIIT Pet datasets when ground-truth part annotations are not available.

**Index Terms**—Fine-grained object categorization, part proposal, weakly supervised learning

## I. INTRODUCTION

FINE-grained object categorization aims for distinguishing objects of subordinate categories that belong to the same entry-level object category, e.g., various species of birds [1], [2], [3], pets [4], [5], or flowers [6]. The difference between entry-level and fine-grained object categorization is shown in Figure 1. Owing to its importance in a wide variety of applications, e.g., multimedia information retrieval [7], [8], [9], e-commerce [10] and rich image captioning [11], [12], fine-grained object categorization has attracted widespread

attention from multimedia community. However, the fine-grained categorization tasks are challenging because the variations among different subordinate object categories are subtle, which are often overwhelmed by those caused by arbitrary poses, viewpoint change, and/or occlusion. It is also difficult to obtain and label a large number of training images of subordinate object categories. Consequently, the performance of fine-grained categorization lies behind that of generic object recognition.

|                                    |                                                                                    |                                                                                     |                                                                                     |                                                                                     |
|------------------------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Input                              |  |  |  |  |
| Entry-level Object Categorization  | Bird                                                                               | Bird                                                                                | Cat                                                                                 | Cat                                                                                 |
| Fine-grained Object Categorization | Brandt Cormorant                                                                   | Red Faced Cormorant                                                                 | Bengal                                                                              | Abyssinian                                                                          |

Fig. 1. Entry-level object categorization *versus* fine-grained object categorization. In entry-level object categorization, we only need to distinguish the first two images of “Bird” from the last two images of “Cat”. In fine-grained object categorization, the subcategories belonging to the same entry-level one should be further differentiated.

It is well known that characterizing features of different subordinate object categories are located at some local parts of object instances (e.g., the head and body of bird as illustrated in Figure 3). Correspondingly, many fine-grained categorization datasets provide ground-truth part annotations [2], [3]. Existing methods [13], [14], [15], [16], [17] use these part annotations to train detection models that can detect from input images the most discriminative parts for the use of fine-grained categorization. However, manually annotating object parts requires expertise, which is also difficult to generalize to fine-grained categorization tasks of new entry-level object categories. To get relief from manual part annotations, a number of recent methods [18], [19], [20], [21], [22], [23], [24], [25], [26] are proposed that aim for mining and leveraging discriminative local parts using image-level category labels only. Weakly supervised learning [20], [21], [22], [24], [25] and attention mechanism in deep networks [18], [19], [23], [26] are the two main workhorses to achieve such a goal. Given region proposals [27], weakly supervised learning based methods use a separate stage of region clustering [20], [21], [22] or region mining [24], [25] to learn part detectors, which is suboptimal for the final task of fine-grained categorization. Attention-based methods [18], [19], [26] overcome such a limitation by automatically identifying and using salient/discriminative pixels and regions in an end-to-end fashion. However, they

This work was supported in part by the National Natural Science Foundation of China (Grant No.: 61771201), and the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No.: 2017ZT07X183).

Y. Zhang, K. Jia, and Z. Wang are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China. E-mails: zhang.yabin@mail.scut.edu.cn, kuijia@scut.edu.cn, wang.zhixin@mail.scut.edu.cn.

seem to have the weakness that a diverse set of discriminative parts are difficult to obtain, which restricts their practical performance.

To address the above limitations, we propose in this work a novel fine-grained categorization architecture called Weakly Supervised Part Detection Network (PartNet). A vanilla PartNet builds on top of a base convolutional (conv) subnetwork two parallel streams of upper network layers: given proposed regions of interest (RoIs), the *classification stream* performs region-level differentiation over subordinate object categories and outputs classification probabilities; the *detection stream* learns a specified number of part detectors that assign association probabilities of these RoIs with each of the learned detectors; the final image-level prediction is obtained by aggregating element-wise products of region-level probabilities of the two streams. PartNet training uses image-level supervision that enables the detection stream to achieve end-to-end learning of diverse part detectors in a weakly supervised manner.

Our proposed PartNet requires proposals of RoIs as inputs of the classification and detection streams. Existing fine-grained categorization works [20], [21], [22], [24], [25] either directly use regions provided by off-the-shelf object proposal methods such as Selective Search (SS) [27], or segment regular sub-regions from object proposals. However, criteria of object proposal methods are designed for region completeness of object instances, with no mechanism of proposing discriminative local parts; segmenting regular sub-regions from object proposals is an indirect approach to discriminative part proposals. Inspired by the discretization of proposal space in Region Proposal Networks (RPN) [28], we introduce in this work a simple but effective Discretized Part Proposals module (DPP). Our part proposals are anchored at salient locations in individual spatial cells of feature maps, where activation values are of higher magnitude. Correspondingly, candidates of discriminative local parts can be proposed independently of spatial locations of (possibly false positive) object instances. Experiments on benchmark fine-grained object categorization datasets show the efficacy of the proposed method. We summarize major contributions of this work as follows.

- We introduce in this paper a novel fine-grained categorization architecture called PartNet (cf. Section III-A). By using parallel classification and detection streams that process RoI features and aggregating their region-level scores, the proposed PartNet achieves end-to-end learning of diverse part detectors in a weakly supervised manner.
- Existing region proposal methods focus on completeness of object-level regions, which is not directly relevant to proposing candidates of discriminative local parts. We introduce in this work a simple but effective DPP (cf. Section III-B), which supports the success of PartNet for fine-grained categorization and could also be useful to other tasks that rely on discriminative local features.
- We present a few variants of PartNet including (1) PartNet with the higher resolution of feature maps and (2) PartNet with orthogonal weight matrix in the classification stream (cf. Section III-C). Experiments on the benchmark datasets of CUB-200-2011, Oxford Flower 102 and Oxford-IIIT Pet show that our proposed PartNet and its

variants are effective for both discriminative part detection and fine-grained categorization. In particular, we achieve the new state-of-the-art performance on the CUB-200-2011 and Oxford-IIIT Pet datasets when ground-truth part annotations are not available (cf. Section IV).

## II. RELATED WORKS

In this section, we first present a brief review of fine-grained object categorization methods. We discuss how discriminative parts among fine-grained categories are essential for this task, with special focus on those methods that do not rely on ground-truth part annotations. We also discuss methods of object/part proposal and weakly supervised object detection, which are the techniques closely related to our proposed method.

### A. Part-Aware Fine-Grained Object Categorization

Since the introduction of fine-grained categorization tasks, researchers realize that extracting features from discriminative local parts is essential to the success of the task. For example, [29] sequentially searches discriminative parts by unifying heuristic function and successor function via a Long Short-Term Memory network (LSTM). The heuristic function evaluates the informativeness of the proposed bounding boxes and the successor function predicts the offsets to the discriminative proposals of the proposed boxes. All the detected image parts are fused for fine-grained recognition. Jointly optimizing the fine-grained classification loss and the Euclidean distances between the proposed part proposals and the ground-truth part proposals, state-of-the-art result is obtained on the benchmark CUB-200-2011 dataset [2]. To get relief from manual part annotations, recent efforts resort to weakly supervised learning [20], [21], [22], [24], [25] and/or attention mechanism in deep networks [18], [19], [23], [26], in order to either implicitly make use of information of salient parts [18], [26], or explicitly identify discriminative local parts based on image-level category labels only [19], [21], [20], [22], [23], [24], [25]. We briefly review some of these representative methods as follows.

Based on off-the-shelf object proposal methods (e.g., SS [27]), multi-scale part proposals are generated in [21] at regular spatial grids of object proposals. These part proposals are then clustered from which useful ones are selected, in a weakly supervised manner, by measuring their importance scores for fine-grained categorization. Xiao *et al.* [20] also use image-level supervision and patch clustering to identify discriminative parts from patch proposals: a classifier of the entry-level category is first trained and used to filter out background patches; spectral clustering is then applied to the remaining patches to learn part detectors (e.g., cluster centers), which are further used to select discriminative parts from patch proposals; final classification is performed using features of the detected parts. Image-level supervision and object-part spatial constraint are applied to select the discriminative part proposals in [22], and then neural clustering clusters selected proposals into semantic parts: a pre-trained entry level classifier is fine-tuned on target data and used to filter out noisy patches; object-level bounding boxes are obtained

by class activation mapping (CAM) [30] and used to further refine the selected proposals; part detectors, which are obtained by performing clustering on the neurons of a middle layer in the classification model, cluster selected proposals into diverse semantic parts. Zhang *et al.* [24], [25] learn initial part detectors from distinctive region proposals by measuring activation outputs of network neurons; the detectors are refined via iterative alternation between new distinctive sample mining and part model retraining; neural activations are pooled into the final representation via a spatially weighted combination of Fisher Vectors coding, which considers the importance of each activation. In [19], multi-scale attention mechanism is employed into classification networks in order to guide deep feature learning to focus on discriminative (species-specific) regions, where starting from the full image, a hierarchy of three-level region scales are gradually attended, and their features are extracted for classification. Fine-grained categorization is obtained by integrating the information of three scale regions. In [18], a diversified LSTM based attention model is proposed that aims to learn a diverse set of discriminative region attentions, so that classification among fine-grained categories can rely more on features of these attended regions. In [23], multiple part attentions are generated by clustering, weighting from spatially-correlated convolutional channels. Part-level patches of each part and object-level images are taken as input to train individual part-CNN. The features of each part and object image of the part-CNNs are concatenated together for final classification. In [26], activation values of feature maps are defined as assignment strengths for surrogate parts, and the part-level features are generated within the Bag-of-Words framework. Multi-scale and multi-position part features are obtained with the scale pooling and sub-region partition schemes on the feature maps respectively. The final image prediction is the product of the global image prediction and the part-level prediction achieved by averaging the parts' features.

Attention-based methods have the nice property that salient/discriminative pixels and regions can be automatically learned and attended in an end-to-end fashion. However, they seem to have the weakness that a diverse set of discriminative parts are difficult to obtain.<sup>1</sup> For example, only one (but multi-scale) part is attended in [19]; consequently, other potentially discriminative parts are ignored in classification. On another hand, existing methods based on explicit region proposals [21], [20] use a separate stage of region clustering to obtain part detectors, which is suboptimal for the final task of fine-grained categorization. While our proposed PartNet also relies on explicit region proposals, we employ in the upper network parallel streams of classification and detection, which simultaneously achieve discriminative part detection and fine-grained categorization. The detection stream also enables learning of diverse part detectors. Superior results on the benchmark datasets [2], [6] show the efficacy of our proposed method.

<sup>1</sup>Even if automatic detection of salient regions is enabled by attention-based methods, it seems that explicit region proposals (e.g., via multi-scale proposals at regular spatial grids) always help. In fact, regions of varying sizes are cropped in [18] at different locations of the original image in order to provide more diversified attention canvas.

## B. Weakly Supervised Object Detection/Localization

Weakly supervised object detection/localization aims to learn object detectors using only image-level category labels, i.e., ground-truth object annotations (e.g., object bounding boxes) are not required. Simple extensions of such techniques could be useful for fine-grained categorization by learning part detectors in a weakly supervised manner. There are many weakly supervised object detection/localization methods proposed in the literature, among which CNN based methods show great promise recently [30], [31], [32], [33], [34]. This may be due to the fact that CNNs have remarkable localization ability despite being trained on image-level labels [30]. We particularly mention here a model of Weakly Supervised Deep Detection Networks (WSDDN) [31]. It introduces a two-stream network architecture where the classification stream differentiates each object proposal among different object categories, and the detection stream ranks for each category all the object proposals. Scores from the two streams are aggregated via element-wise product, which are finally used for image-level supervision. Our proposed PartNet is inspired by [31]. Instead of ranking object proposals for each category in the detection stream, we rely on part proposals and learn multiple part detectors that altogether contribute to the classification of fine-grained categories.

## C. Generation of Object/Part Proposals

Both our proposed PartNet and other part-aware fine-grained categorization methods [21], [20], [22] rely on proposals of local object regions/parts. Part proposals differ from the established object proposal techniques [27], [35] in that salient part locations and part boundaries are less clearly defined. Consequently, it is less obvious to extend existing object proposal techniques for a good part proposal method. Nevertheless, existing efforts either directly use object proposal methods for use of part proposals, e.g., SS [27] is used in [20], [22], or simply use sub-regions of object proposals as part proposals [21]. In this work, we propose a simple DPP that borrows the idea of spatial space discretization from RPN [28]. Our part proposals are anchored at discriminative locations of feature maps, which are obtained by training using image-level category labels only. Comparative studies with SS [27] show the efficacy of our proposed DPP.

## III. THE PROPOSED METHOD

In this section, we present in details our proposed PartNet, which is empowered by a simple but effective DPP module. We also introduce a few variants of PartNet that altogether contribute to an effective solution to part-aware fine-grained categorization.

### A. Weakly Supervised Part Detection Network

As discussed in Section II, it is well established that identification of discriminative local parts is essential for fine-grained categorization. In this work, we design a novel architecture called PartNet (cf. Figure 2 for an illustration), which explicitly learns part detectors using image-level category labels

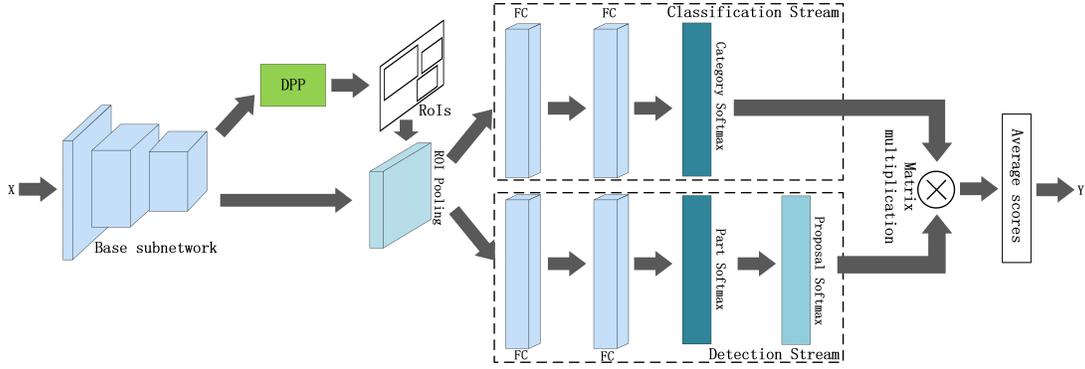


Fig. 2. Framework of the PartNet. The base subnetwork represents the convolutional layers that are pre-trained on ImageNet dataset firstly and then fine-tuned on the fine-grained training data. The DPP represents our proposed module of Discretized Part Proposals (cf. Section III-B) for generating RoIs. The classification stream differentiates region-level proposals over subordinate object categories, while the detection stream assigns association probabilities of those proposals with part detectors. Region-level probabilities of the two streams are combined with matrix multiplication. The image-level classification is obtained by averaging the classification probabilities of different detected parts. The different softmax layers are detailed in Section III-A.

only. Examples of our detected local parts from the fine-grained categorization datasets are also shown in Figure 3 and Figure 7.

A vanilla PartNet uses conv layers as its base subnetwork. Assume the final conv layer of the base subnetwork outputs  $N$  feature maps. Using our proposed DPP (cf. Section III-B), a number  $R$  of local regions of those feature maps are proposed that give RoIs on the input image. These RoIs are of varying sizes, and we use ROI pooling [36] to produce features of the fixed size  $m \times m$ , which, after vectorization, gives a feature vector  $\mathbf{f}_{RoI} \in \mathbb{R}^{Nm^2}$  for each proposed RoI. We use two parallel streams of fully connected (FC) layers on top of the base subnetwork to further process, in a batch mode, these RoI features  $\{\mathbf{f}_{RoI}\}$ . Assume there are  $C$  fine-grained object categories in the considered task. The *classification stream* performs differentiation of the proposed RoIs among these categories. The *detection stream* learns a specified number  $P$  of patterns of parts (i.e., part detectors) that can identify from the proposed RoIs the most effective ones for fine-grained categorization. The two streams output part-level scores of classification/detection probabilities, which are then aggregated and used for image-level training or inference. We present component-wise specifics of our proposed PartNet as follows.

1) *The classification stream*: As shown in Figure 2, we use two consecutive FC layers (with ReLUs) to differentiate each of the RoI feature vectors  $\{\mathbf{f}_{RoI}^i\}_{i=1}^R$  into fine-grained categories. Since some of the proposed RoIs are on the background, which are in fact common in different fine-grained categories, we introduce an additional output neuron in the second FC layer that corresponds to the *background category*. The second FC layer thus outputs a matrix  $\mathbf{X}_{cls}$  of the size  $(C+1) \times R$ . A softmax operator, termed as “category softmax”, is then followed to make  $\mathbf{X}_{cls}$  as a score matrix  $\mathbf{S}_{cls} \in \mathbb{R}^{(C+1) \times R}$  of classification probabilities. Elements of  $\mathbf{S}_{cls}$  are computed as

$$s_{cls}^{ij} = \frac{e^{x_{cls}^{ij}}}{\sum_{c=1}^{C+1} e^{x_{cls}^{cj}}}, \quad (1)$$

where  $x_{cls}^{ij}$  is an entry of  $\mathbf{X}_{cls}$ , and  $i$  and  $j$  index the categories and RoI features respectively.

2) *The detection stream*: The detection stream aims for learning a specified number  $P$  of part detectors that detect from (the proposed RoIs of) the input image local parts that are most useful/discriminative for fine-grained categorization. To this end, we use two consecutive FC layers (with ReLUs) to process RoI features  $\{\mathbf{f}_{RoI}^i\}_{i=1}^R$ . To model those local parts that are either on the background or less discriminative among fine-grained categories, we use  $P+1$  output neurons in the second FC layer. Outputs of the second FC layer are denoted as  $\mathbf{X}_{det} \in \mathbb{R}^{(P+1) \times R}$ . FC layers themselves barely give the detection stream the ability to learn distinctive and semantically meaningful part detectors. We use a softmax operator, termed as “part softmax”, immediately following the second FC layer, which gives the output matrix  $\tilde{\mathbf{S}}_{det} \in \mathbb{R}^{(P+1) \times R}$ . Elements of  $\tilde{\mathbf{S}}_{det}$  are computed as

$$\tilde{s}_{det}^{ij} = \frac{e^{x_{det}^{ij}}}{\sum_{p=1}^{P+1} e^{x_{det}^{pj}}}, \quad (2)$$

where  $x_{det}^{ij}$  is an entry of  $\mathbf{X}_{det}$ , and  $i$  and  $j$  index the part detectors and RoI features respectively. While there are no ground-truth part annotations available, learning part detectors is made possible by the use of part softmax (cf. Eq. (2)): in the forward pass, each proposed RoI is associated with one of the  $P+1$  output neurons of the second FC layer by scaling up the corresponding score toward the value of 1 while suppressing others; this is reinforced in the backward pass and consequently, patterns of discriminative parts are learned as parameters of FC layers in a weakly supervised and locally optimal manner. To give an intuition on what we have learned for part detectors, we illustrate in Figure 3 examples of the proposed RoIs for an input image, where scores in each column are the ones computed from (2) for each RoI, and rows of different colors correspond to individual part detectors, including the background one. Figure 3-(a) shows that these RoIs are differentiated and associated with different part detectors, and those associated with the same one

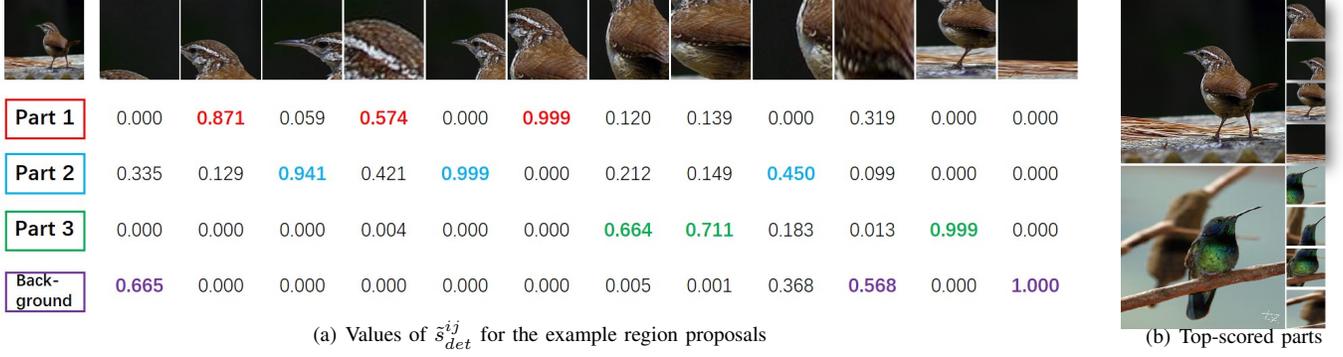


Fig. 3. (a) Visualization of the detection scores  $\tilde{s}_{det}^{ij}$  of equation (2) when applying PartNet to the CUB-200-2011 dataset, where score precision is rounded to the level of  $10^{-3}$ . The first row shows the input image (left) and example region proposals (right) generated by the DPP module. The second, third, fourth, and bottom rows respectively present the scores of three part detectors and the background detector for each proposal. (b) Visualization of input images and their respective detected (top-scored) local parts. Please refer to Figure 7 for more examples of detected local parts.

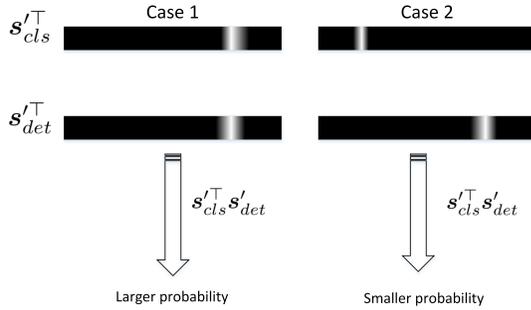


Fig. 4. In case 1, the RoIs that have larger values in  $s'_{det}$  also have larger values in  $s'_{cls}$ , and the classification probability is larger. Otherwise, the classification probability is smaller in case 2. In order to achieve accurate classification, the RoIs, that have larger values in the right category of  $S'_{cls}$ , should consistently have larger values in  $S'_{det}$ .

have the similar visual appearance, suggesting that individual part detectors are trained to characterize patterns of local distinctiveness. Figure 3-(b) also shows that when applying these learned part detectors to images of different categories, they detect local regions that have the potential of fine-grained discrimination.

Then, we use a second softmax operator, termed as ‘‘proposal softmax’’, on  $\tilde{S}_{det}$  to rank their associations with each of the  $P$  part detectors. This produces  $S_{det} \in \mathbb{R}^{(P+1) \times R}$  whose elements are compute as

$$s_{det}^{ij} = \frac{e^{\tilde{s}^{ij}}}{\sum_{r=1}^R e^{\tilde{s}^{ir}}}. \quad (3)$$

The second softmax also serves as a normalization layer that normalizes RoI scores associated with each part detector (i.e., each row of  $\tilde{S}_{det}$ ), so that the resulting  $S_{det}$  can be better used for score aggregation with those of the classification stream, as explained shortly. In this work, we by default set the number  $P$  of part detectors as 3. We also investigate the effects of different values of  $P$  on fine-grained categorization (cf. Section IV-B).

3) *Aggregation of classification and detection scores for image-level supervision/inference*: The classification and detection streams output score matrices  $S_{cls}$  and  $S_{det}$  respectively for all proposals. To use them for image-level supervision or inference, we first remove from  $S_{cls}$  the last row that represents the probabilities of RoIs’ belonging to the background category, and also remove from  $S_{det}$  the last row that contains scores of RoIs associated with the background/irrelevant part detector, resulting in reduced matrices  $S'_{cls} \in \mathbb{R}^{C \times R}$  and  $S'_{det} \in \mathbb{R}^{P \times R}$  respectively. Suppose an input image is of the  $c^{th}$  fine-grained category. We denote the  $c^{th}$  row of  $S'_{cls}$  as  $s'^{\top}_{cls} \in \mathbb{R}^R$  that contains the probabilities that the  $R$  RoIs are classified as the  $c^{th}$  category. We similarly denote the  $p^{th}$  row of  $S'_{det}$  as  $s'^{\top}_{det} \in \mathbb{R}^R$  that contains the probabilities that the  $R$  RoIs are detected as instances of the  $p^{th}$  discriminative parts. Discriminative part detection requires that RoIs that have larger values in  $s'_{det}$  (i.e., the detected instances of the  $p^{th}$  part) should consistently have larger values in  $s'_{cls}$  (cf. Figure 4 for an illustration). We thus choose to use  $s'^{\top}_{cls} s'^{\top}_{det}$  as a measure of part-level classification confidence. Write compactly in a matrix form we have  $S'_{cls} S'^{\top}_{det} \in \mathbb{R}^{C \times P}$ , each row of which contains the probabilities that the detected discriminative parts are of a certain fine-grained category. We then average the part-level probabilities to form the image-level classification representation  $\mathbf{y} \in \mathbb{R}^C$  of the input image as

$$\mathbf{y} = \frac{1}{P} S'_{cls} S'^{\top}_{det} \mathbf{1}_P, \quad (4)$$

where  $\mathbf{1}_P$  denotes a  $P$ -dimensional vector with all values of 1. Note that as mentioned in Section III-A2, the proposal softmax (cf. Eq. (3)) in the detection stream serves as a normalization layer that ensures each entry value of  $S'_{cls} S'^{\top}_{det}$  is in the range  $[0, 1]$ . Consequently, the computed  $\mathbf{y}$  in Eq. (4) can be considered as image-level classification probabilities.

We use the result of Eq. (4) as the inference of a PartNet for an input image. To train the PartNet, assume a set of  $M$  training images are given, each of which has its one-hot vector form of ground-truth category label as  $\mathbf{g} \in \mathbb{R}^C$ . Denote parameters of the PartNet collectively as a vector  $\theta$ , we use

the following loss of binary cross-entropy to train the network

$$\frac{\lambda}{2} \|\theta\|_2^2 - \sum_{i=1}^M \sum_{j=1}^C g_{ij} \log y_{ij}(\theta) + (1 - g_{ij}) \log(1 - y_{ij}(\theta)), \quad (5)$$

where  $g_i$  and  $y_i$  are respectively the ground truth label and inference result for the  $i^{\text{th}}$  training sample, and  $g_{ij}$  and  $y_{ij}$  are their  $j^{\text{th}}$  entries. We optimize Eq. (5) using Stochastic Gradient Descent (SGD) with momentum.

### B. Discretized Part Proposals in Spatial Cells of Feature Maps

The PartNet presented in Section III-A needs proposals of RoIs that specify local regions of input images for classification and detection streams to work on. Existing part proposal methods [21], [22], [20] either directly use regions provided by object proposal method [27], or use their regular sub-regions. However, object proposal methods use criteria that focus on region completeness of object instances and are not effective by design for proposing candidates of discriminative parts. Segmenting regular sub-regions from object proposals can help, but it is not a direct approach to discriminative part proposals. In this work, we propose a simple DPP method towards this goal. Our method is inspired by the discretization of proposal space in RPN [28]; but we do not have a training process since ground-truth part annotations are not available.

It is well known that CNNs have a remarkable localization ability despite being trained using image-level labels [30], and ideally the discriminative parts should locate at positions of feature maps that have larger feature values. A similar idea is also adopted in [26] where the values of feature maps are defined as assignment strengths for surrogate parts. We thus opt to generate part proposals anchored at these positions directly. More specifically, given feature maps of the size  $C \times W \times H$  that have  $C$  channels, we calculate a histogram vector  $\mathbf{h} \in \mathbb{R}^{WH}$  that counts for each of the  $W \times H$  spatial locations the occurrence that channel-wise peak value is located at the current position, and use the obtained  $\mathbf{h}$  to identify discriminative spatial locations. The location of the peak value for each channel is also used in [23]. Counts in the histogram  $\mathbf{h}$  measure the degrees of discrimination for different spatial locations, and part proposals are anchored at those with more counts.

To make part proposals spatially spread over the feature maps, we regularly divide the  $W \times H$  spatial locations into  $S \times S$  non-overlapping cells (e.g.,  $S = 4$ ), which produces the corresponding sub-vectors from the histogram vector  $\mathbf{h}$ . We use spatial locations corresponding to the max count of each histogram sub-vector as our anchors of part proposals. For each anchor position, we define  $K$  anchor boxes of varying sizes and aspect ratios [28]. We by default set  $K = 28$  in our experiments, and Table I gives its box sizes and aspect ratios. The influence of using different  $K$  values is also investigated in Section IV-C.

### C. Other Variants

In this section, we present two variants of PartNet in order to boost the performance on fine-grained categorization tasks.

TABLE I  
THE SPECIFIED SIZES AND ASPECT RATIOS WHEN WE USE  $K = 28$   
ANCHOR BOXES FOR EACH ANCHOR POSITION ON THE FEATURE MAPS.

|               |                |                |                |                |                 |                 |                 |                 |                 |                 |
|---------------|----------------|----------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Anchor sizes  | 3 <sup>2</sup> | 5 <sup>2</sup> | 7 <sup>2</sup> | 9 <sup>2</sup> | 11 <sup>2</sup> | 13 <sup>2</sup> | 15 <sup>2</sup> | 17 <sup>2</sup> | 19 <sup>2</sup> | 21 <sup>2</sup> |
| Aspect ratios | 1:1            |                |                |                | 1:1             | 1:2             | 2:1             |                 |                 |                 |

**Higher resolution of feature maps** Variations among fine-grained categories are often subtle, regional, and imaged in finer details. However, finer details could disappear when feature maps are of lower resolution. To avoid this issue, we present a variant of the vanilla PartNet as follows by modifying the base subnetwork structure. For the models that downsample feature maps via stride-2 conv layers (e.g., ResNet [37]), its last layer of the classifier is removed firstly, then we replace its last stride-2 conv layer (i.e., conv5\_1 in ResNet-34) with a stride-1 one, and modify the subsequent conv layers via 2-dilated conv layers [38]. For the models that downsample feature maps via stride-2 max pooling layers (e.g., VGGNet [39]), the last stride-2 max pooling layer and the subsequent layers are removed. By this way, the resolution of the base subnetwork feature maps is doubled.

**Orthogonal weight matrix in the classification stream** Orthogonal weight matrices are observed to be helpful to propagate information in deep networks [40]. In this work, we present a variant of the vanilla PartNet that applies the technique of Singular Value Bounding (SVB) [40] to the second FC layer in the classification stream. We expect this variant to produce more discriminative scores of classification probabilities among different fine-grained categories.

### D. Final Prediction

The proposed PartNet achieves fine-grained categorization by aggregating regional discrimination of detected individual parts. However, each of the individual parts may independently contribute to fine-grained categorization their own discrimination. The input image may also provide complementary holistic features. To utilize all these part-level and image-level discriminative information, we adopt a region zooming strategy as in [23], [19], [14], [22].

Specifically, given a trained PartNet, the  $P$  part detectors of the detection stream respectively rank the  $R$  region proposals generated by DPP, resulting in a score matrix  $\mathbf{S}'_{det} \in \mathbb{R}^{P \times R}$  (cf. Section III-A3). Intuitively, if features of a proposal  $i$  match pattern of a part  $j$ , then the  $(j, i)$  entry of  $\mathbf{S}'_{det}$  would have a larger value, otherwise it would have a smaller one. We thus select for each of the  $P$  part detector  $M$  region proposals of top scores (e.g.,  $M = 50$ ), and use the selected regions to fine-tune the image-level model, resulting in  $P$  part-level models. During testing, the top-1 region proposal for each part detector is selected and zoomed as the input of the corresponding part-level model. Our final prediction is made by averaging the classification probability of PartNet with those of its associated image-level and the  $P$  part-level models. We term such an ensemble model as PartNet-Full.

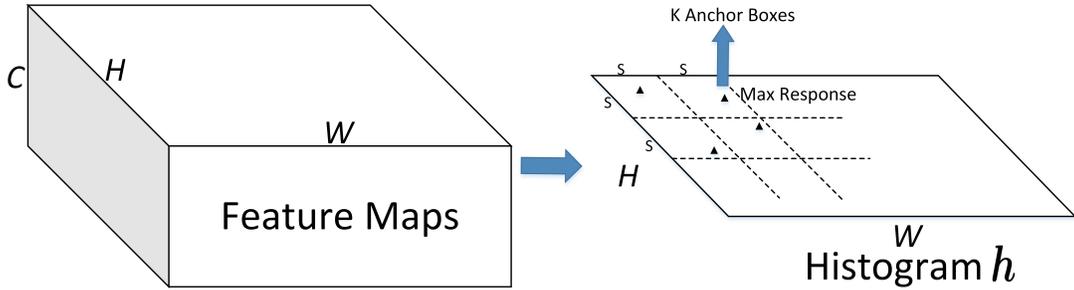


Fig. 5. The framework of the DPP module. The DPP takes as input the feature maps that are generated by the last convolutional layer. A histogram, which is obtained by counting the occurrence of channel-wise peak value for each of the spatial locations, is firstly generated to measure the degrees of discrimination for different spatial locations. All the spatial locations are divided into  $S \times S$  non-overlapping cells, and  $K$  anchor boxes of varying sizes and aspect ratios are proposed anchored at the location corresponding to the max count of the histogram for each cell.

#### IV. EXPERIMENTS

In this section, we conduct fine-grained categorization experiments on the benchmark datasets of CUB-200-2011 [2], Oxford Flower 102 [6] and Oxford-IIIT Pet [5]. We present ablation studies to investigate the component-wise effectiveness of our proposed PartNet, its variants, and the DPP scheme, and also compare with the state of the art. We implement the proposed method on PyTorch and provide the codes at <https://github.com/YBZh/PartNet>.

##### A. Datasets and implementation details

**CUB-200-2011** [2] The Caltech-UCSD Birds 200-2011 dataset is the most widely-used dataset for fine-grained categorization and contains 200 species of birds. It includes 5,994 images for training and 5,794 images for testing. For each image, one bounding box annotation and 15 keypoint annotations are given. We do not use these bounding box or keypoint annotations in our experiments.

**Oxford Flower 102** [6] Oxford Flower 102 contains 102 categories of flowers. There are 1,020 images for training, 1,020 images for validation, and 6,149 images for testing. We do not use the image segmentations provided in this dataset, and instead, we only use the category labels in our experiments.

**Oxford-IIIT Pet** [5] Oxford-IIIT Pet contains 37 pet subcategories, among which 12 are cat subcategories and 25 are dog subcategories. There are 3,680 images for training and 3,669 images for testing. We do not use the pixel level segmentation provided in this dataset, and instead, we only use the category labels in our experiments.

**Baselines and implementation details** We first present the baseline models that we use to compare with our proposed methods. Given a 34-layer ResNet [37] that is pre-trained on the ImageNet [41], we modify its final FC layer of the classifier to make the number of its output neurons the same as that of the target fine-grained categories. This baseline model is termed as ResNet-34. To fairly compare the baseline with variants of our proposed PartNet, we also introduce an additional baseline of Dilated ResNet-34, which is obtained by modifying ResNet-34 as the way described in Section III-C. The ResNet-34 and Dilated ResNet-34 based models

are used in our ablation studies. To investigate the efficacy of our proposed contributions when comparing with many of the existing methods, we also construct our methods on the VGGNet [39]. The VGGNet, where batch normalization [42] is used to improve network training, is pre-trained on the training images of ImageNet dataset [41] firstly and then we modify its structure as the way described in Section III-C. Then a FC layer is followed as the fine-grained categories classifier. Those baseline models are fine-tuned on the target fine-grained categorization datasets and are termed as image-level models.

Our proposed PartNet (and its variants) are constructed based on the above image-level models. Taking the ResNet-34 as an example, we build up the base subnetwork of PartNet by removing its layer of global average pooling and also the subsequent (final) layer of classifier; we then use an RoI pooling layer [36] whose inputs are formed by the part proposals generated by DPP, together with output feature maps of the base subnetwork; following the RoI pooling layer, a parallel pair of detection and classification streams are used that respectively produce scores of detection and classification probabilities; these scores are finally aggregated and used for image-level training or inference (cf. Section III-A). Figure 2 gives an illustration.

To train the above models, we use SGD with momentum: we set the weight decay as  $1e-4$  and momentum as 0.9; we train each model for 160 epochs with a batch size of 128; for parameters that are initialized from pre-trained models, we use a learning rate of  $1e-3$ ; for other parameters, we use an initial learning rate of  $1e-1$ , which drops by a factor of 10 respectively after 80 and 120 epochs. For inputs of PartNet and image-level models, we pre-process each image by resizing its shorter size to 448 while keeping the aspect ratio unchanged. Then we crop a random  $448 \times 448$  region for the use of training (we also use the horizontal flip version of the cropped  $448 \times 448$  region for data augmentation) and a central  $448 \times 448$  region for the use of testing. The part-level models are obtained by fine-tuning the image-level model with the detected part proposals, which are rescaled to the size of  $448 \times 448$  as inputs. All our experiments are based on the above training settings.

Based on models constructed from VGGNet, we report the time to train the models and to label a new sample on Tesla M40 GPUs. Taking the CUB-200-2011 dataset as an example, it takes 19.4, 66.8 and 89.8 GPU hours to train the image-level model, PartNet, and each part-level model respectively. Thus, the whole training time of our method is 355.6 GPU hours, and the model training can be finished in 176 GPU hours considering that the three part-level models can be trained in parallel. It takes 32 ms, 71 ms and 32 ms to label a new sample by image-level model, PartNet, and each part-level model respectively. Thus the prediction of a new sample by the PartNet-Full can be finished in 103 ms considering that predictions of the image-level model and PartNet can be made in parallel, and the same applies to predictions of the three part-level models. The number of parameters of PartNet (58.63 M) is about 41% of those of the original VGGNet model (about 144 M), confirming the efficiency of our proposed method.

### B. Ablation Studies on the Detection Stream

The detection stream is the key component in PartNet. We evaluate its effectiveness on the CUB-200-2011 dataset using a PartNet constructed from Dilated ResNet-34.

The detection stream detects discriminative local parts essentially by learning to assign varying weights to different region proposals. To evaluate its effectiveness, we remove the detection stream of PartNet and correspondingly set scores of detection probabilities for different region proposals as being equal (i.e., setting elements of  $S'_{det}$  in Eq. (4) as  $\frac{1}{R}$ ). We term such a model as Degenerate PartNet. Table II compares results of PartNet, Degenerate PartNet, and also the baseline Dilated ResNet-34, where region proposals are generated by SS [27]. Dilated ResNet-34 performs fine-grained categorization directly on the image level, and its result is worse than that of Degenerate PartNet, showing the benefit of performing fine-grained categorization on the region level. This is consistent with observations in [20], [22]. By learning and assigning varying weights to different region proposals, our proposed PartNet further improves the result. Discriminative local parts can also be detected from region proposals by ranking these weights, which will be presented shortly.

TABLE II  
COMPARATIVE EXPERIMENTS ON THE CUB-200-2011 DATASET [2]  
WITH OR WITHOUT THE DETECTION STREAM IN THE PARTNET, WHICH  
IS CONSTRUCTED FROM THE BASELINE OF DILATED RESNET-34.

| Method             | Proposal Method | Accuracy (%) |
|--------------------|-----------------|--------------|
| Dilated Resnet-34  | NA              | 82.02        |
| Degenerate PartNet | SS              | 82.84        |
| PartNet            | SS              | <b>83.53</b> |

For the detection stream of PartNet, we need to specify the number  $P$  of output neurons of the second FC layer, which is also the specified number of part detectors. To investigate how different values of  $P$  influence classification performance, we conduct experiments on the CUB-200-2011 dataset by setting  $P = 1, 3, 5$ , and 10. Results in Table III show that classification accuracy slightly improves as more part detectors

are used, but at the price of increased computation cost. In our experiments, we set  $P = 3$  for a balance between accuracy and efficiency.

TABLE III  
CLASSIFICATION PERFORMANCE ON THE CUB-200-2011 DATASET [2]  
WHEN USING DIFFERENT NUMBERS OF PART DETECTORS (I.E.,  $P$   
VALUES IN EQ. (2)) IN THE DETECTION STREAM OF PARTNET. THE  
PARTNET IS CONSTRUCTED FROM DILATED RESNET-34.

| No. of Part Detectors | 1     | 3     | 5     | 10           |
|-----------------------|-------|-------|-------|--------------|
| Accuracy (%)          | 83.51 | 83.53 | 83.62 | <b>83.63</b> |

### C. Ablation Studies on DPP

We investigate our proposed DPP method by conducting experiments on the CUB-200-2011 dataset using a PartNet constructed from Dilated ResNet-34.

The number of proposals generated by DPP for each image may influence the performance of PartNet. To investigate, we first generate a number  $K = 28$  of boxes for each spatial cell of feature maps (cf. Table I in Section III-B for how sizes and aspect ratios of the boxes are specified); we then rank the 28 boxes associated with each cell according to their sizes/areas, and uniformly sample 3, 7, and 14 ones out of them respectively. This creates scenarios of generating  $K = 3, 7, 14, 28$  boxes per spatial cell for our proposed DPP. Results in Table IV show that classification accuracies slightly improve as more proposals are used. In our experiments, we by default set  $K = 28$  for each spatial cell of feature maps.

TABLE IV  
EFFECT OF DIFFERENT NUMBERS OF PROPOSALS WHEN USING OUR  
PROPOSED DPP FOR FINE-GRAINED CATEGORIZATION. EXPERIMENTS  
ARE CONDUCTED ON THE CUB-200-2011 DATASET [2] USING A  
PARTNET CONSTRUCTED FROM DILATED RESNET-34.

| No. of Proposals per Cell | $K = 3$ | $K = 7$ | $K = 14$ | $K = 28$     |
|---------------------------|---------|---------|----------|--------------|
| Accuracy (%)              | 84.31   | 84.41   | 84.36    | <b>84.43</b> |

We also compare with other region proposal methods used in recent fine-grained categorization works [14], [20], including SS [27] and an improved version of SS termed Filtered SS. Filtered SS removes noisy proposals that are irrelevant to the objects of interest in an image (e.g., those on the background) by an object-level attention model [20], and it thus enjoys an unfair advantage over both SS and our proposed DPP. For both SS and Filtered SS, we use the same number of region proposals as our DPP does: when these methods produce more proposals, we rank them in terms of areas of proposed regions, and then uniformly sample the same number of region proposals; in some rare case that these methods produce fewer proposals, we also duplicate some ones. Results in Table V show that our DPP method outperforms both SS and Filtered SS, confirming that candidates of discriminative local parts can be sampled directly at salient positions of feature maps, with no need to be bridged via object-level proposals.

### D. Ablation Studies on Variants of PartNet

Our first PartNet variant is analysed with ResNet-34 and Dilated ResNet-34 models. The Dilated ResNet-34 model

TABLE V  
CLASSIFICATION ACCURACIES (%) OF PARTNET ON THE CUB-200-2011 DATASET [2] WHEN USING DIFFERENT METHODS TO GENERATE REGION PROPOSALS.

| Method       | SS [27] | Filtered SS [20] | DPP          |
|--------------|---------|------------------|--------------|
| Accuracy (%) | 83.53   | 84.00            | <b>84.43</b> |

TABLE VI  
EFFECT OF RESOLUTION OF FEATURE MAPS TO FINE-GRAINED CATEGORIZATION TASKS. THE DILATED RESNET-34 PRODUCES DOUBLED FEATURE MAP RESOLUTION OVER THAT OF RESNET-34. EXPERIMENTS ARE CONDUCTED ON THE CUB-200-2011 DATASET [2].

| ImageNet Pre-training | Method                                     | Acc. (%)     |
|-----------------------|--------------------------------------------|--------------|
| No                    | ResNet-34                                  | 54.44        |
| No                    | Dilated ResNet-34                          | <b>60.95</b> |
| Yes                   | ResNet-34                                  | 81.78        |
| Yes                   | Dilated ResNet-34                          | <b>82.02</b> |
| Yes                   | PartNet constructed from ResNet-34         | 82.98        |
| Yes                   | PartNet constructed from Dilated ResNet-34 | <b>84.36</b> |

uses 2-dilated convolution [38] to double the resolution of feature maps without affecting the size of the receptive field. To investigate the effect of feature map resolution itself for fine-grained categorization, we first use the baseline model of ResNet-34, which is either pre-trained on the ImageNet or trained from scratch on the CUB-200-2011 dataset. Since ResNet-34 produces feature maps whose resolution is only half of that of feature maps produced by Dilated ResNet-34, our DPP method cannot generate  $K = 28$  per-cell proposals for ResNet-34. We thus set  $K = 14$  in this comparative experiment. Results in Table VI show that for both of the considered training settings (i.e., with or without ImageNet pre-training), higher resolution of feature maps contributes to better classification accuracies, showing its usefulness in fine-grained categorization by preserving finer details of appearance features. When applying dilated convolution to our proposed PartNet (constructed from ResNet-34), performance gets a clear boost as well.

Our second PartNet variant enforces weight matrix of the second FC layer in the classification stream to be orthogonal, by using the SVB technique proposed in [40]. To investigate its effectiveness, we again conduct experiments on the CUB-200-2011 dataset using PartNet constructed from Dilated ResNet-34. Results in Table VII show that this variant achieves improved classification performance. Note that results of PartNet and PartNet-Full reported in Sections IV-E and IV-F are based on this variant.

TABLE VII  
RESULTS OF PARTNET ON THE CUB-200-2011 DATASET [2] WITH OR WITHOUT USING WEIGHT ORTHOGONALIZATION FOR THE SECOND FC LAYER OF THE CLASSIFICATION STREAM.

| Weight Orthogonalization | Method  | Accuracy (%) |
|--------------------------|---------|--------------|
| No                       | PartNet | 84.43        |
| Yes                      | PartNet | <b>84.73</b> |

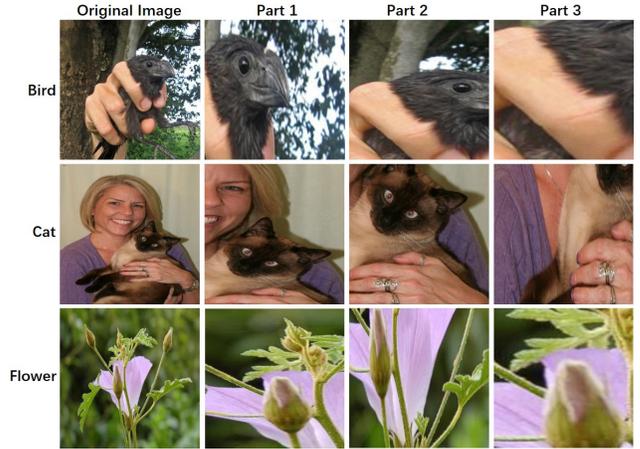


Fig. 6. Some failure results of part detection. Images of “Bird”, “Cat”, and “Flower” are from CUB-200-2011 [2], Oxford-IIIT Pet [5], and Oxford Flower 102 [6] datasets, respectively.

#### E. Ensemble of PartNet with Its Associated Image- and Part-level Models

We introduce in Section IV-A that PartNet is constructed from an image-level base model, and multiple part-level models can also be obtained by fine-tuning the image-level model with the region proposals respectively detected by the learned part detectors of PartNet. PartNet contributes to fine-grained categorization by aggregating local discriminative evidence provided by part detectors. Complementary to PartNet, image- and part-level models may respectively provide their own discrimination by emphasizing either the holistic image or each of the individual parts. It is arguably beneficial to use an ensemble of these models to further boost classification performance. Empirical success of similar model ensemble is also presented in [22], [23], [19].

For model ensemble in this section, we use the VGGNet based PartNet with the two variants introduced in Section III-C. Table VIII shows results of individual models and various model combinations of the ensemble. We can observe that:

- Classification accuracies of the individual part-level models are relatively low, which may be attributed to the relatively less information contained in individual parts, and also the occasional failure of part detection. Figure 6 shows two main reasons (i.e., complex background and heavy occlusion) causing the failure of part detection. However, averaging the predictions of three part-level models boosts the accuracy by a large margin, proving that the detected three part-level regions are complementary with each other.
- Averaging the classification probabilities of image- and part-level models achieves large performance improvements (e.g., 3.92%, 1.35%, and 1.44% on the datasets of CUB-200-2011, Oxford Flower 102, and Oxford-IIIT Pet respectively) over the results of using the image-level models alone, justifying the complementarity of holistic image and individual parts.
- Combining PartNet, image-level, and part-level models

TABLE VIII

CLASSIFICATION ACCURACIES (%) OF INDIVIDUAL MODELS AND VARIOUS MODEL COMBINATIONS OF PARTNET WITH ITS ASSOCIATED IMAGE- AND PART-LEVEL MODELS ON THE CUB-200-2011 [2], OXFORD FLOWER 102 [6] AND OXFORD-IIIT PET [5] DATASETS. THE VGGNET IS USED TO EXTRACT FEATURES IN ALL THE EXPERIMENTS IN THIS TABLE AND THE SYMBOL "+" MEANS AVERAGING THE CLASSIFICATION PROBABILITIES OF CORRESPONDING MODELS.

| Method                                                       | CUB-200-2011 [2] | Oxford Flower 102 [6] | Oxford-IIIT Pet [5] |
|--------------------------------------------------------------|------------------|-----------------------|---------------------|
| Image-level                                                  | 82.19            | 95.12                 | 92.07               |
| PartNet                                                      | 85.11            | 95.95                 | 92.56               |
| Part 1                                                       | 78.51            | 92.79                 | 76.97               |
| Part 2                                                       | 75.68            | 93.30                 | 83.78               |
| Part 3                                                       | 77.55            | 91.12                 | 81.71               |
| Part 1 + 2 + 3                                               | 83.64            | 95.82                 | 88.39               |
| Part 1 + 2 + 3 + Image-level                                 | 86.11            | 96.47                 | 93.51               |
| Image-level + PartNet                                        | 83.98            | 95.62                 | 92.80               |
| Part 1 + 2 + 3 + PartNet                                     | 86.19            | 96.43                 | 92.53               |
| Our PartNet-Full<br>(Part 1 + 2 + 3 + PartNet + Image-level) | <b>86.90</b>     | <b>96.70</b>          | <b>95.37</b>        |

TABLE IX

CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE CUB-200-2011 [2] DATASET.

| Method                         | Training annotation | Test annotation | CNN Features | Accuracy (%) |
|--------------------------------|---------------------|-----------------|--------------|--------------|
| VGG-BGLm [43]                  | BBox                | BBox            | VGGNet       | 80.40        |
| PG Alignment [44]              | BBox                | -               | VGGNet       | 82.00        |
| Coarse-to-Fine [45]            | BBox                | -               | VGGNet       | 82.50        |
| PG Alignment [44]              | BBox                | BBox            | VGGNet       | 82.80        |
| Coarse-to-Fine [45]            | BBox                | BBox            | VGGNet       | 82.90        |
| PBC [46]                       | BBox                | -               | GoogleNet    | 83.30        |
| FCAN [47]                      | BBox                | BBox            | ResNet-50    | 84.70        |
| Part-based RCNN [14]           | BBox + Parts        | -               | AlexNet      | 73.90        |
| PBC [46]                       | BBox + Parts        | BBox            | GoogleNet    | 83.70        |
| DPS-CNN [48]                   | Parts               | -               | GoogleNet    | 85.12        |
| SPDA [16]                      | BBox + parts        | BBox            | VGGNet       | 85.14        |
| Zhang et al. [15]              | Parts               | -               | VGGNet       | 85.92        |
| HSnet [29]                     | Parts               | -               | GoogleNet    | <b>87.50</b> |
| Two-level Attention [20]       | -                   | -               | AlexNet      | 69.70        |
| VGG-BGLm [43]                  | -                   | -               | VGGNet       | 75.90        |
| DVAN [18]                      | -                   | -               | VGGNet       | 79.00        |
| Zhang <i>et al.</i> [21]       | -                   | -               | VGGNet       | 79.34        |
| NAC [49]                       | -                   | -               | VGGNet       | 81.01        |
| STN [50]                       | -                   | -               | GoogleNet    | 84.10        |
| Bilinear-CNN [51]              | -                   | -               | VGGNet       | 84.10        |
| FCAN [47]                      | -                   | -               | ResNet-50    | 84.30        |
| PDFS [24]                      | -                   | -               | VGGNet       | 84.54        |
| PNA [25]                       | -                   | -               | VGGNet       | 84.70        |
| RA-CNN [19]                    | -                   | -               | VGGNet       | 85.30        |
| MA-CNN (2 parts + object) [23] | -                   | -               | VGGNet       | 85.40        |
| OPAM [22]                      | -                   | -               | VGGNet       | 85.83        |
| DT-RAM [52]                    | -                   | -               | ResNet-50    | 86.00        |
| MA-CNN (4 parts + object) [23] | -                   | -               | VGGNet       | 86.50        |
| Our PartNet-Full               | -                   | -               | VGGNet       | <b>86.90</b> |
| Our PartNet-Full               | -                   | -               | ResNet-34    | <b>87.30</b> |

further boosts classification performance on the three datasets, certifying the effectiveness of our proposed method.

#### F. Comparison with State-of-the-art Methods

We compare our PartNet with the state-of-the-art methods on the benchmark datasets of CUB-200-2011 [2], Oxford Flower 102 [6] and Oxford-IIIT Pet [5]. Table IX presents comparison results on the CUB-200-2011 dataset. The types of annotation used in the training and test stages of each method are also listed in the table, where "CNN Features" indicates which (base) network is used to extract features in each method.

TABLE X

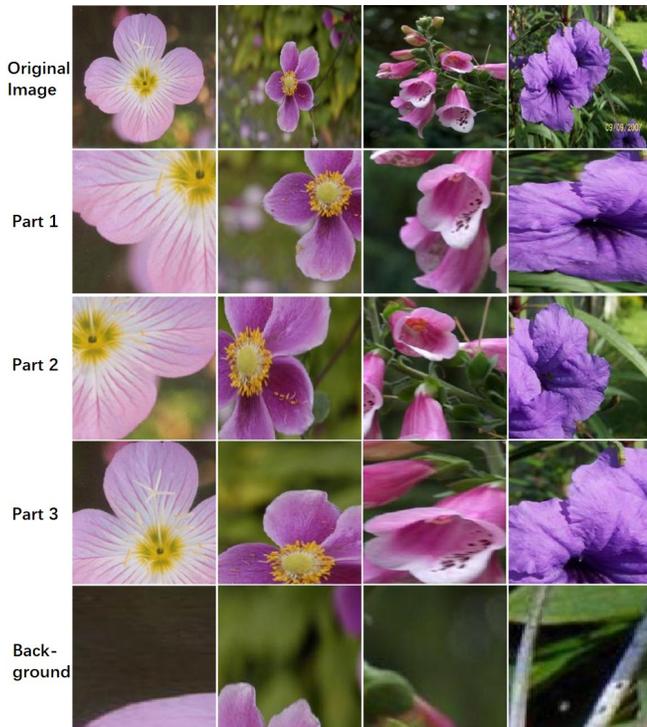
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE OXFORD FLOWER 102 [6] DATASET.

| Method           | CNN Features | Accuracy (%) |
|------------------|--------------|--------------|
| MPP [53]         | AlexNet      | 91.28        |
| Magnet [54]      | GoogleNet    | 91.40        |
| BoSP [26]        | VGGNet       | 94.02        |
| NAC [49]         | VGGNet       | 95.34        |
| PBC [46]         | GoogleNet    | 96.10        |
| OPAM [22]        | VGGNet       | <b>97.10</b> |
| Our PartNet-Full | VGGNet       | <b>96.70</b> |

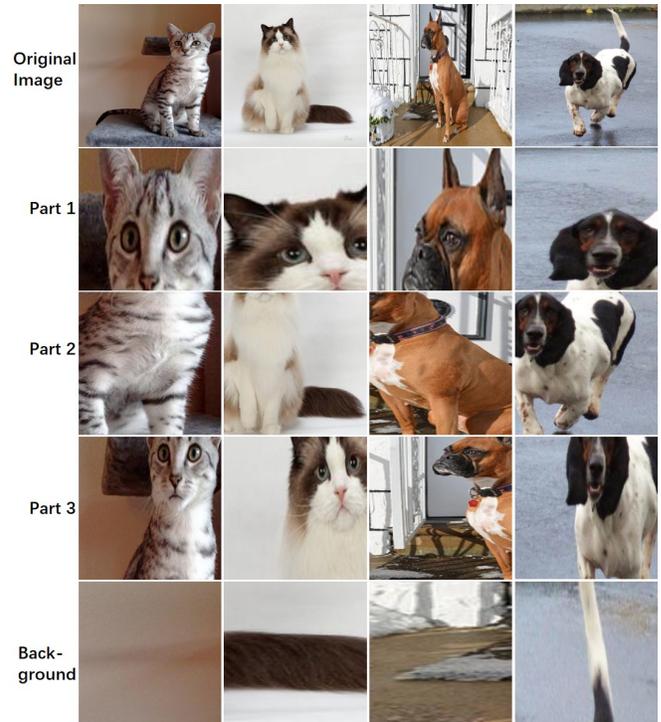
When constructing the PartNet using the VGGNet (with the two variants introduced in Section III-C), our PartNet-



(a) Birds



(b) Flowers



(c) Pets

Fig. 7. Visualization of the detected parts on datasets of CUB-200-2011 [2], Oxford Flower 102 [6] and Oxford-IIIT Pet [5]. The first row denotes the original images, and the second, third, and fourth rows denote the parts detected by the three part detectors respectively. The last row denotes the background or less discriminative proposals detected by the background detector. Results are obtained by the PartNet constructed from the VGGNet with the two variants introduced in Section III-C. The images in (a) Birds, (b) Flowers and (c) Pets are from the test data of CUB-200-2011 [2], Oxford Flower 102 [6] and Oxford-IIIT Pet [5] datasets respectively.

Full (i.e., the ensemble model described in Section IV-E) obtains the new state-of-the-art result on the CUB-200-2011 dataset when neither object nor part annotations are used. Furthermore, our method outperforms most of the existing

ones that need part or object annotations, such as [16], [15], [48], [14]. When constructing the PartNet using the base network of Dilated ResNet-34, our PartNet-Full obtains an even better result on the CUB-200-2011 dataset.

TABLE XI  
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE  
OXFORD-IIIT PET [5] DATASET.

| Method                   | CNN Features | Accuracy (%) |
|--------------------------|--------------|--------------|
| NAC [49]                 | VGGNet       | 91.60        |
| Two-level Attention [20] | VGGNet       | 92.51        |
| OPAM [22]                | VGGNet       | <b>93.81</b> |
| Our PartNet-Full         | VGGNet       | <b>95.37</b> |

Note that the state-of-the-art method HSnet [29] uses the ground-truth part annotations in the training stage, making it less relevant to compare directly with our proposed method. Our PartNet-Full combines multi-level models for final prediction by simply averaging the classification probabilities of these models. In contrast, the MA-CNN [23] trains a classifier based on the concatenated features of multi-level models, and OPAM [22] learns a weight for each model with the computationally expensive k-fold cross-validation method, yet their results are still worse than ours.

We present our result on the Oxford Flower 102 dataset in Table X. Our PartNet-Full obtains the result that is comparable with state-of-the-art method [22].

We also present our result on the Oxford-IIIT Pet dataset in Table XI. Our PartNet-Full obtains the new state-of-the-art result, justifying the efficacy of our PartNet.

### G. Part Detection Visualization

In Figure 7, we visualize the detected discriminative parts by the VGGNet based PartNet (with the two variants introduced in Section III-C), where images are from test data of the CUB-200-2011 [2], Oxford Flower 102 [6] and Oxford-IIIT Pet datasets. We observe in Figure 7 that our detected local parts have physical meanings:

- For the bird dataset, the first two parts (Part 1 and Part 2) are on local regions of bird head, with the second one being a slightly zoomed-in version of the first one, and the third part (Part 3) is on local regions of bird body (back and/or abdomen).
- For the flower dataset, the three parts are roughly on local regions of a flower or some of its petals, regardless of how many flowers are contained in each of the images.
- For the pet dataset, the first part (Part 1) and the third part (Part 3) are on local regions of pet head, with the first one being a slightly zoomed-in version of the third one, and the second part (Part 2) is on local regions of pet body.

These detected local parts arguably provide semantically discriminative information for fine-grained categorization. Figure 7 also shows that the background detector gathers region proposals that are on the image background and are thus less relevant to the task of interest. The influence of background proposals for image category prediction can thus be decreased by removing the background detector before combining the two streams (cf. Section III-A3).

## V. CONCLUSIONS

In this paper, we propose a novel Weakly Supervised Part Detection Network (PartNet) for part-aware fine-grained

object categorization. Our PartNet contains two streams: the classification stream classifies part-level region proposals over subordinate categories; the detection stream selects discriminative proposals for the use of fine-grained object categorization. The image-level classification is obtained by the combination of region-level probabilities of the two streams, and meanwhile diverse part detectors can be learned in an end-to-end fashion under the image-level supervision. To prepare part-level region proposals for the PartNet, we design a simple Discretized Part Proposals method that utilizes the localization information in the feature maps directly. Experiments on the benchmark datasets of CUB-200-2011 [2], Oxford Flower 102 [6] and Oxford-IIIT Pet [5] demonstrate the efficacy of our proposed PartNet on fine-grained categorization and salient part detection. Especially our approach obtains the new state-of-the-art result on the CUB-200-2011 and Oxford-IIIT Pet datasets when ground-truth part annotations are not available. We believe that such methods, that only need image categorization level supervision are important for new fine-grained categorization tasks.



**Yabin Zhang** received the B.E. degree in School of Electronic and Information Engineering from South China University of Technology, Guangzhou, China, in 2017, where he is currently pursuing the master's degree. His current research interests include computer vision and deep learning, especially the deep transfer learning.



**Kui Jia** received the B.E. degree from Northwestern Polytechnic University, Xian, China, in 2001, the M.E. degree from the National University of Singapore, Singapore, in 2004, and the Ph.D. degree in computer science from the Queen Mary University of London, London, U.K., in 2007.

He was with the Shenzhen Institute of Advanced Technology of the Chinese Academy of Sciences, Shenzhen, China, Chinese University of Hong Kong, Hong Kong, the Institute of Advanced Studies, University of Illinois at Urbana-Champaign, Champaign, IL, USA, and the University of Macau, Macau, China. He is currently a Professor with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China. His recent research focuses on theoretical deep learning and its applications in vision and robotic problems, including deep learning of 3D data and deep transfer learning.



**Zhixin Wang** received the B.E. degree in School of Electronic and Information Engineering from South China University of Technology, Guangzhou, China, in 2017, where he is currently pursuing the master's degree. His recent research focuses on computer vision and deep learning, especially the object detection.

## REFERENCES

- [1] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," California Institute of Technology, Tech. Rep. CNS-TR-2010-001, 2010.
- [2] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.
- [3] T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2011–2018.
- [4] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, vol. 2, 2011.
- [5] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [6] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*. IEEE, 2008, pp. 722–729.
- [7] L. Xie, J. Wang, B. Zhang, and Q. Tian, "Fine-grained image search," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 636–647, 2015.
- [8] X. Wang, T. Zhang, D. R. Tretter, and Q. Lin, "Personal clothing retrieval on photo collections by color and attributes," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 2035–2045, 2013.
- [9] L. Zhu, J. Shen, H. Jin, L. Xie, and R. Zheng, "Landmark classification with hierarchical multi-modal exemplar feature," *IEEE Transactions on Multimedia*, vol. 17, no. 7, pp. 981–993, 2015.
- [10] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3343–3351.
- [11] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4565–4574.
- [12] L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1–10.
- [13] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked cnn for fine-grained visual categorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1173–1182.
- [14] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *European conference on computer vision*. Springer, 2014, pp. 834–849.
- [15] N. Zhang, E. Shelhamer, Y. Gao, and T. Darrell, "Fine-grained pose prediction, normalization, and recognition," *arXiv preprint arXiv:1511.07063*, 2015.
- [16] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas, "Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1143–1152.
- [17] C. Huang, Z. He, G. Cao, and W. Cao, "Task-driven progressive part localization for fine-grained object recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2372–2383, 2016.
- [18] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1245–1256, 2017.
- [19] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Conf. on Computer Vision and Pattern Recognition*, 2017.
- [20] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. 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*, 2015, pp. 842–850.
- [21] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, J. Lu, V.-A. Nguyen, and M. N. Do, "Weakly supervised fine-grained categorization with part-based image representation," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1713–1725, 2016.
- [22] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1487–1500, 2018.
- [23] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Int. Conf. on Computer Vision*, 2017.
- [24] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1134–1142.
- [25] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking neural activations for fine-grained recognition," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2736–2750, 2017.
- [26] Y. Guo, Y. Liu, S. Lao, E. M. Bakker, L. Bai, and M. S. Lew, "Bag of surrogate parts feature for visual recognition," *IEEE Transactions on Multimedia*, 2017.
- [27] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.
- [28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
- [29] M. Lam, B. Mahasseni, and S. Todorovic, "Fine-grained recognition as hsnet search for informative image parts," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 6497–6506.
- [30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.
- [31] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2846–2854.
- [32] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "Contextlocnet: Context-aware deep network models for weakly supervised localization," in *European Conference on Computer Vision*. Springer, 2016, pp. 350–365.
- [33] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," *CoRR*, vol. abs/1704.00138, 2017. [Online]. Available: <http://arxiv.org/abs/1704.00138>
- [34] X. Zhang, H. Xiong, W. Lin, and Q. Tian, "Weak to strong detector learning for simultaneous classification and localization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.
- [35] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*. Springer, 2014, pp. 391–405.
- [36] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [38] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.
- [39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [40] K. Jia, D. Tao, S. Gao, and X. Xu, "Improving training of deep neural networks via singular value bounding," in *Conf Comp Vis Pattern Recognit*, vol. 2017, 2017, pp. 3994–4002.
- [41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.
- [42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [43] F. Zhou and Y. Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1124–1133.
- [44] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5546–5555.
- [45] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian, "Coarse-to-fine description for fine-grained visual categorization," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4858–4872, 2016.
- [46] C. Huang, H. Li, Y. Xie, Q. Wu, and B. Luo, "Pbc: polygon-based classifier for fine-grained categorization," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 673–684, 2017.

- [47] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin, "Fully convolutional attention networks for fine-grained recognition," *arXiv preprint arXiv:1603.06765*, 2016.
- [48] S. Huang and D. Tao, "Real time fine-grained categorization with accuracy and interpretability," *arXiv preprint arXiv:1610.00824*, 2016.
- [49] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1143–1151.
- [50] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.
- [51] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.
- [52] Z. Li, Y. Yang, X. Liu, S. Wen, and W. Xu, "Dynamic computational time for visual attention," *arXiv preprint arXiv:1703.10332*, 2017.
- [53] D. Yoo, S. Park, J.-Y. Lee, and I. So Kweon, "Multi-scale pyramid pooling for deep convolutional representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 71–80.
- [54] O. Rippel, M. Paluri, P. Dollar, and L. Bourdev, "Metric learning with adaptive density discrimination," *arXiv preprint arXiv:1511.05939*, 2015.