Interpretable and Accurate Fine-grained Recognition via Region Grouping

Zixuan Huang[1] Yin Li[2,1]
[1]Department of Computer Sciences, [2]Department of Biostatistics and Medical Informatics
University of Wisconsin–Madison
{zhuang356, yin.li}@wisc.edu

Abstract

We present an interpretable deep model for fine-grained visual recognition. At the core of our method lies the integration of region-based part discovery and attribution within a deep neural network. Our model is trained using image-level object labels, and provides an interpretation of its results via the segmentation of object parts and the identification of their contributions towards classification. To facilitate the learning of object parts without direct supervision, we explore a simple prior of the occurrence of object parts. We demonstrate that this prior, when combined with our region-based part discovery and attribution, leads to an interpretable model that remains highly accurate. Our model is evaluated on major fine-grained recognition datasets, including CUB-200 [56], CelebA [36] and iNaturalist [55]. Our results compare favorably to state-of-the-art methods on classification tasks, and our method outperforms previous approaches on the localization of object parts. Our project website can be found at https://www.biostat.wisc.edu/~yli/cvpr2020-interp/.

1. Introduction

Deep models are tremendously successful for visual recognition, yet their results are oftentimes hard to explain. Consider the examples in Fig. 1. Why does a deep model recognize the bird as "Yellow-headed Blackbird" or consider the person "Smiling"? While the interpretation of a model can happen at multiple facets, we believe that at least one way of explaining the model is to segment meaningful regions of object parts (e.g., the eyes, mouth, cheek, forehead and neck of a face), and further identify their contributions towards the decision (e.g., the mouth region is more discriminative for smiling). How can we design an interpretable deep model that learns to discover object parts and estimates their importance for visual recognition?

It turns out that part discovery, i.e., learning object parts without explicit supervision of part annotations, is by itself a challenging problem. As a baby step, we focus on the task of fine-grained recognition, where the parts belonging to the same super category share common visual patterns. For example, most tails of birds have a similar shape. Our key observation is that features from a convolutional network can be used to group pixels into a set of visually coherent regions [28, 25], from which a subset of discriminative segments can be selected for recognition [33, 32, 11]. *With only object labels as the guidance*, we hope that the grouping will help to find visually distinct parts, and the selection process will identify their contributions for classification.

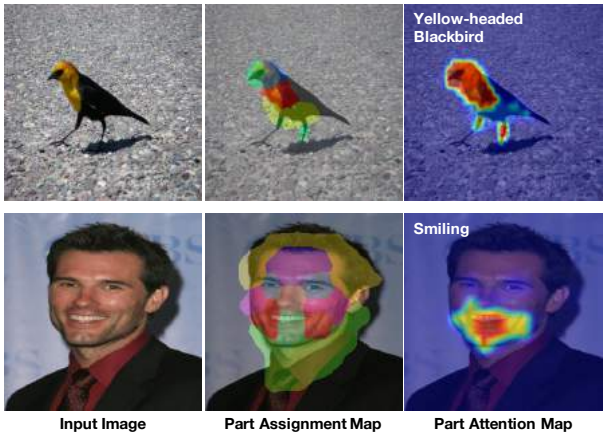

Figure 1. Why does a deep model recognize the bird as *"Yellow-headed Blackbird"* or consider the person *"Smiling"*? We present an interpretable deep model for fine-grained recognition. Given an input image (left), our model is able to segment object parts (middle) and identify their contributions (right) for the decision. *Results are from our model trained using only image-level labels.*

A major challenge for our region-based part discovery is that there is no explicit supervisory signal to define part regions. Therefore, prior knowledge about object parts must be incorporated to facilitate the learning. A core innovation of our work is the exploration of a simple prior about object parts: given a single image, the occurrence of a part follows a U-shaped distribution. For example, the head of a bird is likely to occur in most bird images while the legs of a bird might only appear in some images. Surprisingly, we demonstrate that this simple prior, when combined with our

region-based part discovery, leads to the identification of meaningful object parts. More importantly, the resulting interpretable deep model remains highly accurate. Several recent methods have been developed for discovering parts in fine-grained classification, yet none of them considered the prior we use.

To this end, we present our interpretable deep model for fine-grained classification. Specifically, our model learns a dictionary of object parts, based on which a 2D feature map can be grouped into "part" segments. This is done by comparing pixel features to part representations in a learned dictionary. Moreover, region-based features are pooled from the result segments, followed by an attention mechanism to select a subset of segments for classification. Importantly, during training, we enforce a U-shaped prior distribution for the occurrence of each part. This is done by minimizing the Earth Mover's Distance between our prior and the empirical distribution of part occurrence. During training, our model is only supervised by object labels with our proposed regularization term. During testing, our model jointly outputs the segments of object parts, the importance of the segmented parts, and the predicted label. The interpretation of our model is thus granted by the part segmentation and the contribution of each part for classification.

To evaluate our model, we conduct extensive experiments using three fine-grained recognition datasets for both interpretability and accuracy. To quantify interpretability, we compare the output region segments from our model to the annotated object parts. For accuracy, we report standard metrics for fine-grained classification. On smaller scale datasets, such as CUB-200 [36] and CelebA [56], our model is shown to find parts of the birds and faces with low localization error, while at the same time compares favorably to state-of-the-art methods in terms of accuracy. On the more challenging iNaturalist dataset [55], our model improves the accuracy of a strong baseline network (ResNet101) by **5.7%**, reduces the object localization error, and demonstrates promising qualitative results for part discovery.

2. Related Work

There has been an emerging interest in explaining deep models. Our work focused on developing interpretable deep models for fine-grained classification, following the paradigm of region-based recognition. We briefly survey relevant literature on interpretable deep learning, part-based fine-grained classification, and recent deep models for region segmentation and region-based recognition.

Visualizing and Understanding Deep Networks. Several recent efforts have been developed to visualize and understand a trained deep network. Many of these post-hoc approaches [39, 17, 61, 15, 70, 51] focused on developing visualization tools for the activation maps and/or the filter weights within trained networks. Other works sought to identify the discriminative regions in an input image given a pre-trained network [54, 48, 51, 18, 61, 43, 4, 71, 38]. Beyond qualitative results, Bau et al. [4] proposed a quantitative benchmark that compares the activation of a network's units to human-annotated concept masks. An alternative direction is to learn a simple model, such as a linear classifier [45] or decision tree [13], to mimic the behavior of a trained network, thereby providing an explanation of the target model's outputs. Our work shares the same motivation for interpreting deep models, yet we integrate interpretation into the learning of the model. Similar to [4], we also use human-annotated object parts to quantify the interpretability of our network.

Interpretable Deep Models. Interpretability can be built with a deep model. Many recent works developed deep models that are interpretable by their design. For instance, Zhang et al. [66] designed a regularization method that encourages each filter in high-level convolutional layers to focus on a specific object part. Brendel et al. [6] proposed BagNet that takes small image patches as input, followed by a bag-of-feature (BoF) representation for whole image classification. BagNet can naturally attribute the decision to local regions, and thus help to explain the decision process. Alvarez-Melis and Jaakkola [40] proposed to assign relevance scores to the basis of global image features. Alternatively, new network architectures can be designed for interpretable models. For example, Capsule Networks [47] substitute the commonly used scalar activation with vectors, where the latter is believed to represent entities such as an object or an object part. A relevant idea is further extended in [53] upon conventional CNNs by enforcing sparse connection from convolutional units to the final prediction.

The most relevant work is from Chen et al. [8]. They proposed to learn prototypes of object parts within the network. The decision of the model thus depends on the identification of the prototypes found in the input image. Similar to their work, our model also seeks to explicitly encode the concepts of object parts. However, our work is different from [8] in two key aspects: (1) we adopt region grouping to provide explanation grounded to image segments; and (2) the learning of our model is regularized by a strong prior of the occurrence of object parts.

Part Discovery for Fine-grained Recognition. Identifying discriminative object parts is important for fine-grained classification [50, 49, 58, 67]. For example, bounding box or landmark annotations can be used to learn object parts for fine-grained classification [24, 34, 41, 62, 64]. To avoid costly annotation of object parts, several recent works focused on unsupervised or weakly-supervised part learning using deep models. Xiao et al. [58] performed spectral clustering on convolutional filters to find representative filters for parts. Wang et al. [57] proposed to learn a bank of convolutional filters that captures class-specific object

parts. Moreover, attention models have also been explored extensively for learning parts. Liu et al. [35] made use of reinforcement learning to select region proposals for fine-grained classification. Zheng et al. [68] grouped feature channels for finding parts and their attention, where channels sharing similar activation patterns were considered as a part candidate.

Similar to previous works, our work also seeks to find parts and to identify their importance for fine-grained classification. However, our work differs from previous works by considering an explicit regularization of the occurrence of object parts. Moreover, we also consider a large scale dataset (iNaturalist [55]) for part discovery. We will compare to previous methods on both recognition accuracy and part localization error in our experiments.

Weakly-supervised Segmentation of Object Parts. Our work is also connected to previous works on weakly-supervised or unsupervised segmentation of object parts. Zhang et al. [65] extracted activation from a pre-trained CNN to represent object parts in a graph. Their learning of parts is supervised by a few part annotations. Collins et al. [12] performed a non-negative matrix factorization over the activation of a pre-trained CNN, where each component defines a segment of the image. Jampani et al. [28] proposed an iterative deep model for learning superpixel segments. More recently, Hung et al. [25] presented a deep model that incorporates strong priors, such as spatial coherence, rotation invariance, semantic consistency and saliency, for unsupervised learning of object parts. Our work is inspired by [25], where we also explore novel regularization for learning to segment object parts. However, we consider weakly supervised part segmentation in the context of fine-grained classification. Moreover, we explore a very different prior of part occurrence.

Region-based Recognition. Finally, our model combines segmentation and classification into a deep model, thus links to the efforts of region-based recognition [19, 59, 31, 1], or compositional learning [52]. There has been a recent development of designing deep models for region-based recognition. For example, Li et al. [33] proposed to group CNN features into a region graph, followed by a graph convolutional network for visual recognition. Similar ideas were also explored by Chen et al. [11]. More recently, Li et al. [32] presented a deep model that jointly refines the grouping and the labeling of the regions, using expectation-maximization. Moreover, Arslan [3] proposed to build a graph neural network using pre-defined regions for brain image classification. Our model uses a similar idea to [33, 11, 32] for grouping CNN features. However, none of these previous works focused on the quality of grouping, and thereby they can not be directly used for interpretation.

3. Method

Consider a set of N 2D image feature maps $\boldsymbol{X}_{1:N} = \{\boldsymbol{X}_n\}$ and their categorical labels $y_{1:N} = \{y_n\}$, where $\boldsymbol{X}_n \in \mathbb{R}^{D\times H\times W}$ is D-dimensional features on 2D image plane $H \times W$ from a convolutional network and $y_n \in [1, ..., c]$ is the image-level label of fine-grained categories. The goal of our model is to learn a part dictionary $\boldsymbol{D} \in \mathbb{R}^{D\times K}$ and a decision function $\hat{y} = \phi(\boldsymbol{X}_n, \boldsymbol{D}; \boldsymbol{\theta})$ for fine-grained classification. Specifically, $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_K]$ and each column vector $\boldsymbol{d}_k$ represents an object part concept. And $\boldsymbol{\theta}$ are the parameters of $\phi(\cdot)$. $\phi(\cdot)$ thus takes both the feature maps $\boldsymbol{X}_n$ and the part dictionary $\boldsymbol{D}$ to predict the labels y_n. We now present an overview of our model, as illustrated in Fig. 2. Without loss of clarity, we sometimes drop the subscript n.

Specifically, we assume the function $\phi(\boldsymbol{X}, \boldsymbol{D}; \boldsymbol{\theta})$ can be further decomposed into three parts.

- **Part Segmentation**. A soft part assignment map $\boldsymbol{Q} \in \mathbb{R}^{K\times H\times W}$ is created by comparing the feature map $\boldsymbol{X}$ to the part dictionary $\boldsymbol{D}$. This is achieved by using a grouping function $g(\cdot)$ such that $\boldsymbol{Q} = g(\boldsymbol{X}, \boldsymbol{D}; \boldsymbol{\theta}_g)$.
- **Region Feature Extraction and Attribution**. Based on the assignment map $\boldsymbol{Q}$ and part dictionary $\boldsymbol{D}$, region features $\boldsymbol{Z} \in \mathbb{R}^{D\times K}$ are pooled from the feature maps $\boldsymbol{X}$. We further compute an attention vector $\boldsymbol{a} \in \mathbb{R}^K$, where each element provides the importance score for a part segment. Formally, $[\boldsymbol{Z}, \boldsymbol{a}] = f(\boldsymbol{X}, \boldsymbol{Q}, \boldsymbol{D}; \boldsymbol{\theta}_f)$.
- **Attention Based Classification**. The region features $\boldsymbol{Z}$ are re-weighted by region attention $\boldsymbol{a}$, followed by a linear classifier for the decision of y. This is realized by $h(\cdot)$ where $\hat{y} = h(\boldsymbol{Z}, \boldsymbol{a}; \boldsymbol{\theta}_c)$.

Regularization of Part Occurrence. Before we describe our design of $g(\cdot)$, $f(\cdot)$ and $h(\cdot)$, let us look at the major challenge of learning. Since the only supervisory signal is y, it is challenging to make sure that the dictionary $\boldsymbol{D}$ can capture meaningful object parts. Our key assumption is that we can regularize the learning by enforcing a prior distribution for the occurrence of each part $\boldsymbol{d}_k$ within a set of image features $\boldsymbol{X}_{1:N}$. Specifically, given $\boldsymbol{X}_{1:N}$, with slightly abusing of notation, we denote $p(\boldsymbol{d}_k|\boldsymbol{X}_{1:N})$ as the conditional probability of part $\boldsymbol{d}_k$ occurring in the set $\boldsymbol{X}_{1:N}$. We assume that $p(\boldsymbol{d}_k|\boldsymbol{X}_{1:N})$ follows a U-shaped distribution $\hat{p}(\boldsymbol{d}_k)$ that acts like a probabilistic binary switch, where we can control the probability of "on" and "off". For example, on the CUB-200 birds dataset, all bird parts are presented in most of the bird images, such that the switch is almost always on. In contrast, on the more challenging iNaturalist dataset, an object part is only activated for a certain number of categories, and thus the switch is likely to activate only for some of the images.

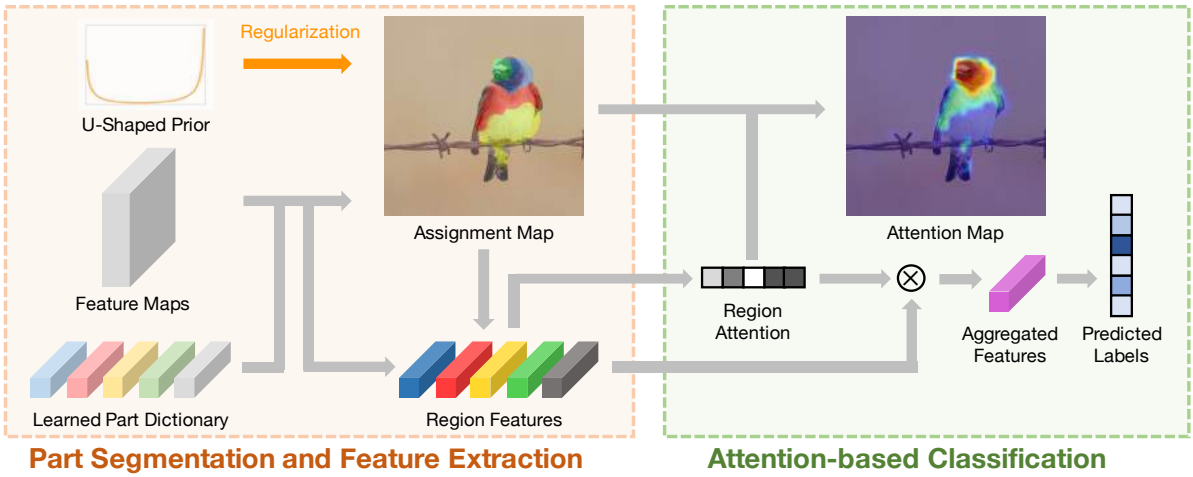

Figure 2. Overview of our method. With image-level labels, our model learns to group pixels into meaningful object part regions and to attend to these part regions for fine-grained classification. Our key innovation is a novel regularization of part occurrence that facilitates part discovery during learning. Once learned, our model can output (1) a part assignment map; (2) an attention map and (3) the predicted label of the image. We demonstrate that our model provides an accurate and interpretable deep model for fine-grained recognition.

3.1. Part Segmentation and Regularization

We now present the details for part segmentation and how to regularize the occurrence of parts.

Part Assignment. We adopt a similar projection unit proposed in previous work [33, 11]. More precisely, let q_{ij}^k be an element in the assignment matrix $\boldsymbol{Q}$, where i, j index the 2D position and k indexes the parts. q_{ij}^k indicates the probability for a feature vector $\boldsymbol{x}_{ij} \in \mathbb{R}^D$ at position (i, j) on $\boldsymbol{X}$ to be assigned to the k-th part $\boldsymbol{d}_k$ in $\boldsymbol{D}$. q_{ij}^k is computed as

$$q_{ij}^k = \frac{\exp(-\|(\boldsymbol{x}_{ij} - \boldsymbol{d}_k)/\sigma_k\|_2^2/2)}{\sum_k \exp(-\|(\boldsymbol{x}_{ij} - \boldsymbol{d}_k)/\sigma_k\|_2^2/2)}, \tag{1}$$

where $\sigma_k \in (0, 1)$ is a learnable smoothing factor for each part $\boldsymbol{d}_k$. Due to the softmax normalization, we have $q_{ij}^k > 0$ and $\sum_k q_{ij}^k = 1$. Moreover, We assemble all assignment vectors $\boldsymbol{q}_{ij} \in \mathbb{R}^K$ into our part assignment map $\boldsymbol{Q} \in \mathbb{R}^{K \times H \times W}$.

Part Occurrence. Given the part assignment map, our next step is to detect the occurrence of each part $\boldsymbol{d}_k$. A simple part detector can be implemented using a max pooling operation over the k-th assignment map $\boldsymbol{Q}^k = [q_{ij}^k]$. However, we found it beneficial to "smooth" the assignment map before pooling, e.g., using a Gaussian kernel with a small bandwidth. This smoothing operation helps to eliminate outliers on the feature map. Our part detector is thus defined as $t_k = \max_{ij} \mathcal{G} * \boldsymbol{Q}^k$, where $\mathcal{G}$ is a 2D Gaussian kernel and $*$ is the convolution operation. t_k lies in the range of $(0, 1)$. Furthermore, the outputs of k part detectors are concatenated into an occurrence vector $\boldsymbol{\tau} = [t_1, t_2, ..., t_K]^T \in (0, 1)^K$ for all parts.

Regularization of Part Occurrence. Our key idea is to regularize the occurrence of each part. This is done by enforcing the empirical distribution of part occurrence to align with a U-shaped prior distribution. More precisely, given a set of N samples, e.g., those used in a mini-batch sampled from the full dataset, we first estimate the empirical distribution $p(\boldsymbol{d}_k|\boldsymbol{X}_{1:N})$ by concatenating all occurrence vectors $\tau_n, n = 1, 2, ..., N$ into an matrix $\boldsymbol{T} = [\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, ..., \boldsymbol{\tau}_N] \in (0, 1)^{K \times N}$. Moreover, we assume a prior distribution $\hat{p}(\boldsymbol{d}_k)$ is known, e.g., a Beta distribution. We propose to use 1D Wasserstein distance, also known as Earth-Mover distance to align $p(\boldsymbol{d}_k|\boldsymbol{X}_{1:N})$ and the prior $\hat{p}(\boldsymbol{d}_k)$, given by

$$W(p(\boldsymbol{d}_k|\boldsymbol{X}_{1:N}), \hat{p}(\boldsymbol{d}_k)) = \int_0^1 |F^{-1}(z) - \hat{F}^{-1}(z)| dz,$$

where $F(\cdot)$ and $\hat{F}(\cdot)$ are the Cumulative Distribution Functions (CDFs) for the empirical and prior distribution. And z spans the interval of $[0, 1]$.

During mini-batch training, the Wasserstein distance can be approximated by replacing the integration with a summation over the samples within the mini-batch, leading to the L1 distance between F^{-1} and $\hat{F}^{-1}$. In practice, we find that it is helpful to rescale the inverse of the CDFs using a logarithm function, which improves the stability of training.

$$\mathcal{L}_W = \frac{1}{N} \sum_{n=1}^{N} |\log([\boldsymbol{\tau}_k^*]_i + \epsilon) - \log(\hat{F}^{-1}\left(\frac{2n-1}{2N}\right) + \epsilon)|, \tag{2}$$

where $\boldsymbol{\tau}_k^*$ is the sorted version (in ascending order) of the k-th row vector of $\boldsymbol{T}$ (size N) and $[\boldsymbol{\tau}_k^*]_i$ is the i-th element of $\boldsymbol{\tau}_k^*$. ϵ is a small value added for numerical stability. Using the logarithm rescaling overcomes a gradient vanishing problem introduced by the softmax function in Eq. 1. Even if a part $\boldsymbol{d}_k$ is far away from all feature vectors in the current mini-batch, i.e., with a small value of q_{ij}^k in Eq. 1, $\boldsymbol{d}_k$ can still receive non-zero gradients due to the rescaling.

We note that there are different approaches to align two 1D distributions. We have previously experimented with Cramèr-von Mises criterion by shaping the CDFs as proposed in [5]. However, we found that our choice of 1D Wasserstein produces more robust results across datasets.

3.2. Region Feature Extraction and Attribution

Given the part assignment, our next step is to pool features from each region. This is done by using a nonlinear feature encoding scheme [33, 29, 42, 2], given by

$$\boldsymbol{z}'_k = \frac{1}{\sum_{ij} q^k_{ij}} \sum_{ij} q^k_{ij}(\boldsymbol{x}_{ij} - \boldsymbol{d}_k)/\sigma_k, \quad \boldsymbol{z}_k = \frac{\boldsymbol{z}'_k}{\|\boldsymbol{z}'_k\|_2}. \quad (3)$$

$\boldsymbol{z}_k$ is thus the region feature from pixels assigned to part $\boldsymbol{d}_k$. By combining $\boldsymbol{z}_k$ together, we obtain our region feature set $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_K] \in \mathbb{R}^{D\times K}$ from the input feature maps. We further transform $\boldsymbol{Z}$ using a sub-network f_z that has several residual blocks, where each one is a bottleneck block consisting of three 1x1 convolutions with Batch Normalization [26] and ReLU. The transformed features are thus $f_z(\boldsymbol{Z})$.

Moreover, an attention module is attached on top of $\boldsymbol{Z}$ to predict the importance of each region. This is realized by a sub-network f_a, given by $\boldsymbol{a} = \text{softmax}(f_a(\boldsymbol{Z}^T))$, where f_a consists of multiple 1x1 convolutions with Batch Normalization and ReLU in-between. The result attention $\boldsymbol{a} \in \mathbb{R}^K$ is further used for classification.

3.3. Attention Based Classification

Finally, we re-weight the transformed region features $f_z(\boldsymbol{Z})$ using the attention vector $\boldsymbol{a}$, followed by a linear classifier. Therefore, the final prediction is given by

$$\hat{y} = \text{softmax}(\boldsymbol{W} f_z(\boldsymbol{Z})\boldsymbol{a}) \quad (4)$$

where $\boldsymbol{W} \in \mathbb{R}^{C\times D}$ is the weights of a linear classifier for C-way classification. Note that the attention $\boldsymbol{a}$ serves as a modulator of the region features $\boldsymbol{Z}$. Thus, large values in $\boldsymbol{a}$ suggest a more important region for classification.

Pixel Attribution. Given the attention $\boldsymbol{a}$, we can easily back-track the contribution of each pixel on the feature map. This can be done by using $\mathcal{Q}^T\boldsymbol{a}$, where $\mathcal{Q} \in \mathbb{R}^{K\times HW}$ is simply a reshaped version of the part assignment map $\boldsymbol{Q}$.

3.4. Implementation

We present our implementation details on loss functions, network architecture, as well as training and inference.

Loss Function. Our model was trained by minimizing cross entropy loss for classification plus the 1D Wasserstein distance based regularization loss in Eq. 2 for part regularization. We varied the weights balancing the loss terms and the prior Beta distribution used for Wasserstein distance during our experiments.

Network Architecture. We replaced the last convolutional block of a baseline CNN (ResNet101 [23]) with our proposed module. We roughly matched the number of parameters in our final model to the baseline.

Training and Inference. We used standard mini-batch SGD for all datasets. The hyper-parameters were chosen differently across datasets due to varied tasks and will be discussed in the experiments. We applied data augmentation including random crop, random horizontal flip and color jittering, and adopted learning rate decay as in [23]. The convolutional layers in our models were initialized from ImageNet pre-trained models. The new parameters, including the part dictionary, were randomly initialized following [22]. All parameters were jointly learned on the target dataset. For all experiments, we report results using a single center crop unless further notice.

4. Experiments and Results

We now describe our experiments and discuss the results. We first introduce datasets and metrics used in our experiments. We then present our experiments and results on individual dataset, followed by an ablation study. For all experiments, results are reported on both accuracy and interpretability and compared against latest methods.

Datasets. Three fine-grained recognition datasets are considered in our experiments, including CelebA [36], CUB-200-2011 [56] and iNaturalist 2017 [55]. These datasets span over a range of tasks and sizes. CelebA is a medium scale dataset for facial attribute recognition and facial landmark detection. CUB-200 is a small scale dataset for bird species recognition that also comes with bird keypoint annotations. Finally, iNaturalist 2017 is a large-scale dataset for fine-grained species recognition and detection, with over 5000 categories spanning from mammals to plants.

Evaluation Metric. We evaluate both accuracy and interpretability of our model for fine-grained visual recognition. For accuracy, we report the standard instance level or average class accuracy as previously considered for fine-grained classification. As a proxy of interpretability, we measure object part localization error using annotated object landmarks, since our model is designed to discover object parts. This localization error has been previously considered for part segmentation models, such as Hung et al. [25]. For the dataset, e.g., iNaturalist 2017 that does not come with part annotations, we follow the protocol of Pointing Game [63] and report object localization error using the annotated object bounding boxes. Pointing Game has been widely used to evaluate interpretable deep models [63, 48, 43].

Concretely, part localization errors are reported on CelebA and CUB-200. Following a similar protocol in [25], we convert our assignment maps to a set of landmark locations by learning a linear regression model. The regression model maps the 2D geometric centers of part assignment

Method	Acc (%) ↑
LNet+ANet [36]	87.0
MOON [46]	90.9
Lu et al. [37]	91.0
Hand et al. [20]	91.2
Kalayeh et al. [30]	91.8
He et al. [21]	91.8
PS-MCNN [7]	**93.0**
ResNet101	91.5
Ours	91.5

Table 1. Results of facial attribute recognition on CelebA dataset. Average class accuracy is reported.

	DFF	SCOPS	**Ours**
Error (%) ↓	31.3	15.0	**8.4**

Table 2. Results of facial landmark localization on CelebA dataset. Normalized L2 distance (%) is reported.

maps into 2D object landmarks. The predicted landmarks are compared against ground-truth on the test set. We report normalized mean L2 distance between the prediction and ground-truth. For iNaturalist 2017, Pointing Game results are reported. We calculate the error rate by counting the cases where the peak location of our output attention map lies outside the ground-truth object bounding boxes.

4.1. Results on CelebA

Dataset. CelebA [36] is a facial attribute and landmark detection dataset that contains 202,599 celebrity face images collected from Internet. Each face image is annotated with 40 facial attributes and 5 landmark locations (eyes, noise and mouth corners). We consider two different splits of the data from [36, 25]. The first split from [36] includes 162,770, 19,867, and 19,962 images for training, validating and testing respectively, and is used to evaluate facial attribute recognition. Faces in this split are aligned to image center. The second split from [25] has 45,609 images for training, 5,379 images for fitting the linear regressor and 283 images for testing. This split is used to report part localization error. Faces are not aligned in this split.

Implementation Details. We trained two models on the two splits using the same architecture. We attached a separate attention-based binary classification head for each facial attribute, as these attributes are not mutually exclusive. For attribute recognition, our model was trained on the training set and evaluated on test set. The validation set was used to select hyper-parameters. For landmark localization, we followed the training procedure from [25]. Our models were trained using a learning rate of 5e-3 with batch size of 32 and a weight decay of 1e-4 for 30 epochs. We set the weights between the two loss terms to 10:1 and used a prior Beta distribution with α=1 and β=1e-3 (close to Bernoulli with $p = 1$). All input images were resized to 256×256 and fed into the model without cropping. A dictionary of 9 parts was used. When reporting the part localization error, we normalize the L2 distance by the inter-ocular distance [25].

Figure 3. Visualization of part assignment maps on CelebA. From left to right: input images, DFF results [12], SCOPS results [25] and our results. Our results are better aligned with the facial parts.

Figure 4. Visualization of attention maps from our model on CelebA. From left to right: input images, attention maps corresponding to different attributes. Our model is able to identify regions that are discriminative for facial attributes.

Recognition Results (Accuracy). Our results on attribute recognition are summarized in Table. 1. We compare our results with state-of-the-art methods [36, 46, 37, 20, 30, 21, 7] as well as a baseline ResNet101 (pre-trained on ImageNet). Surprisingly, the baseline ResNet101 already achieves similar or even higher accuracy in comparison to many previous methods, including those require auxiliary face pars-

Method	Acc (%) ↑
STN [27]	84.1
Kernel Pooling [14]	86.2
MA-CNN [68]	86.5
PC-DenseNet161 [16]	86.9
KERL [9]	87.0
DFL-CNN [57]	**87.4**
NTS-Net [60]	**87.5**
DCL [10]	**87.8**
TASN [69]	**87.9**
ResNet101	**87.7**
Ours	**87.3**

Table 3. Results of bird species recognition on CUB-200-2011. Instance level accuracy is reported.

Method	CUB-001	CUB-002	CUB-003
DFF	22.4	21.6	22.0
SCOPS	18.5	18.8	21.1
Ours	**15.6**	**15.9**	**13.8**

Table 4. Results of landmark localization errors on CUB-200-2011. Normalized L2 distance (%) is reported.

ing [30, 21]. Our model performs on par with the strong ResNet101 baseline. The only method that is significantly better than our model and ResNet101 baseline is [7], which uses extra labels of face identities. To summarize, our model achieves state-of-the-art accuracy.

Localization Results (Interpretability). We further evaluate facial landmark localization results, as shown in Table 2. Our results are compared to the latest methods of DFF [12] and SCOPS [25]. DFF performs non-negative matrix factorization on the feature maps of a pre-trained CNN (VGG[1]) to generate the part segmentation. SCOPS explores spatial coherence, rotation invariance, semantic consistency and visual saliency in self-supervised training for object part segmentation. Our model outperforms both methods by a significant margin in localization error, achieving a 6.6% and 21.9% error reduction when compared to SCOPS and DFF, respectively. These results suggest that our model can localize facial landmarks with high accuracy, thus supporting the interpretability of our model.

Visualization. Our model achieves state-of-the-art results on facial attribute recognition and provides new capacity for facial landmark localization. We further visualize the assignment maps from our model and compare them to those from DFF [12] and SCOPS [25] in Fig. 3. Moreover, we display the attention maps from our model in Fig. 4. Note that our attention maps are attribute specific, as we used a separate classification head for each attribute. These qualitative results show that our model is able to segment faces into meaningful part regions (e.g., hair, forehead, eyes, nose, mouth and neck), and to attend to those regions that are discriminative for attribute recognition (e.g., eye regions for "narrow eyes" or hair regions for "black hair").

[1]Using ResNet leads to worse results of DFF as reported in [12].

Figure 5. Visualization of sample assignment and attention maps from our model on CUB-200-2011 test set. From top to bottom: input images, assignment maps and attention maps. Our method can consistently identify body parts under different poses.

4.2. Results on CUB-200-2011

Dataset. Caltech-UCSD Birds-200-2011 [56] (CUB-200) is a small scale dataset for fine-grained bird species recognition. CUB-200 contains 5,994/5,794 images for training/test from 200 different bird species. Each image is annotated with a species label, 15 bird landmarks and a bounding box of the bird.

Implementation Details. We trained a single model for both classification and landmark localization using a learning rate of 1e-3, batch size of 32 and a weight decay of 5e-4 for 150 epochs. We set the weights between the two loss terms to 2:1, used a prior distribution the same as CelebA and a dictionary of 5 parts. We resized the input images by scaling the shortest side to 448 and randomly crop a region of 448x448 for training. When reporting the part localization error, we normalized the L2 distance using the size of the bird's bounding box, similar to [25].

Recognition Results (Accuracy). We present our results of recognition accuracy and compare them to state-of-the-art methods [27, 14, 68, 16, 9, 57, 60, 10, 69] in Table. 3. Again, the baseline ResNet101 already achieves state-of-the-art results on CUB-200. Our model is slightly worse than ResNet101 (-0.4%) and performs on par with previous part-based models like MA-CNN [68].

Localization Results (Interpretability). Moreover, we evaluate part localization error and compare our results to DFF [12] and SCOPS [25]. To make a fair comparison, we report the errors on the first three categories following [25], as shown in Table 4. Again, our model significant reduces the localization error (2.9%–6.2%). When fitting with all 200 categories, our model achieves an average localization error of 11.51%. These results provide further evidences towards the interpretability of our model.

Visualization. We also visualize the assignment maps and the attention maps from our model, as presented in Fig.

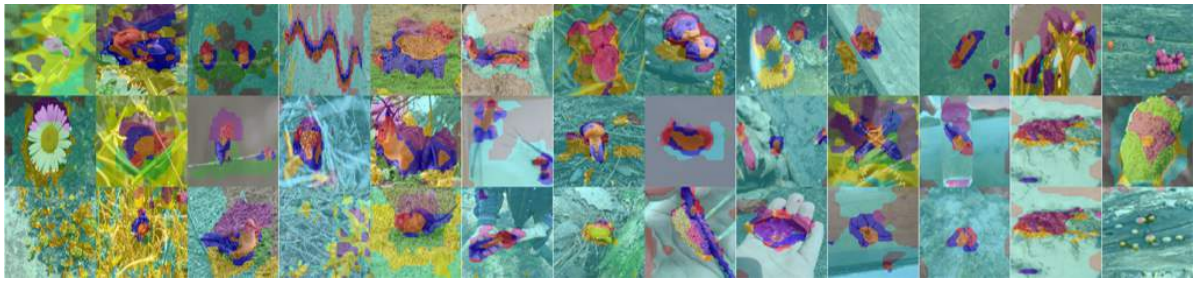

Figure 6. Visualization of sample assignment maps on iNaturalist 2017 test set. Each column comes from one super category.

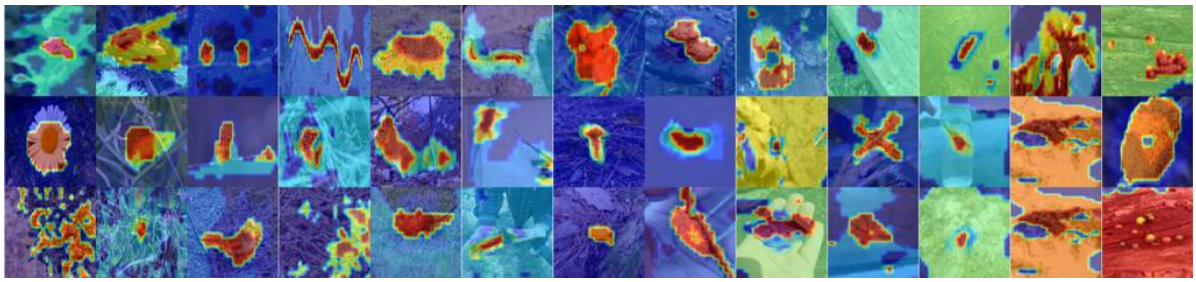

Figure 7. Visualization of sample attention maps on iNaturalist 2017 test set. Each column comes from one super category.

5. Our model demonstrates the ability to discover coherent parts of the birds (e.g, beak/legs, head, wings/tail, body) and select the important regions (beak/legs and wings/tails) to recognize the species.

4.3. Results on iNaturalist 2017

Dataset. iNaturalist 2017 [56] is a large-scale dataset for fine-grained species recognition. It contains 579,184 and 95,986 for training and testing from 5,089 species organized into 13 super categories. Some images also come with bounding box annotations of the object. Since the dataset does not include part annotations, we report Pointing Game results to evaluate interpretability of our model. This dataset is very challenging for mining meaningful object parts, as the objects in different super-categories have drastically different visual appearance (e.g., plants vs. mammals).

Implementation Details. We trained a single model for classification and localization. Our model was trained using a learning rate of 1e-3 with batch size of 128 and a weight decay of 5e-4 for 75 epochs. During training, we resize the input images by scaling the shortest side to 320 and randomly crop a region of 224x224. We set the weights between the two loss terms to 10:1. A dictionary of 8 parts was used and a prior Beta distribution with α=2e-3 and β=1e-3 was considered. We also explored fully convolutional test by feeding the full image (shortest side of 320) into the model.

Recognition Results (Accuracy). Table 5 summarizes our results and compares them to baseline ResNet101 models and the latest methods including SSN [44] and TASN [69]. Both SSN and TASN make use of attention-based upsampling to zoom in discriminative regions for classification.

Method	Acc (%) ↑
SSN	65.2
TASN	**68.2**
ResNet101 [69]	59.6
ResNet101 (ours)	61.1
Ours	64.8
Ours + FC test	66.8

Table 5. Results of species classification on iNaturalist 2017. Instance level accuracy is reported. FC test: fully convolutional test.

Method	Error (%) ↓
CAM [71] / Grad-CAM [48]	11.8
Guided Grad-CAM [48]	8.2
Ours	**7.6**

Table 6. Results of Pointing Game on iNaturalist 2017. CAM/Grad-CAM & Guided Grad-CAM use a ResNet101 model.

Unlike CelebA and CUB-200, the baseline ResNet101 result is much worse than state-of-the-art models (SSN and TASN). Our model improves the baseline ResNet101 by at least 3.7%. Using test time augmentation (full convolutional testing) further boosts our results by 2%. However, our model is still worse than TASN (-1.4%). We speculate that a similar upsampling mechanism as SSN and TASN can be used by our model to further improve the accuracy.

Localization Results (Interpretability). Furthermore, we report Pointing Game results in Table 6. Our results are further compared to widely used saliency methods using the baseline ResNet101 model, including CAM/Grad-CAM [71, 48] and Guided Grad-CAM [48]. Note that CAM and Grad-CAM are the same when visualizing features from the last convolutional layer of a ResNet. Our model achieves the lowest localization error (4.2% and 0.6% improvements for CAM/Grad-CAM and Guided Grad-CAM).

Method	Acc (%) ↑	Error (%) ↓
w/o attention	**91.5**	**7.6**
w/o regularization	**91.3**	12.3
full model	**91.5**	8.4

Table 7. Ablation study on CelebA using the split from [25]. Both recognition accuracy and localization error are reported.

Finally, the visualization of the assignment and attention maps are shown in Fig. 6 and Fig. 7.

4.4. Ablation Study, Limitation and Discussion

Ablation. We conduct an ablation study on CelebA to evaluate our model components. Our study considers two variants, one without regularization and one without attention. Table 7 presents both recognition accuracy and localization error on the split from [25]. The accuracy of all our variants are quite similar, yet our regularization largely improves the localization accuracy (3.9%). Our model without attention has slightly better part localization performance, yet lacks the critical ability of region and pixel attribution compared to our full model. Our full model has small localization error for all landmarks–7.4%, 7.5%, 9.1%, 9.3% and 8.6% for left eye, right eye, nose, left mouth corner and right mouth corner, respectively.

Limitation and Discussion. Many failure cases of our model were found on iNaturalist dataset, as shown in Fig. 6 and 7. Our model may fail to group pixels into part regions and sometimes produce incorrect saliency maps. We speculate these failure cases are produced by our prior Beta distribution. With over 5K fine-grained categories on iNaturalist, a single U-shaped distribution for all parts might fail to describe the occurrence of parts. Moreover, our model does not model the interaction between parts and requires a moderate to large batch size to estimate the empirical distribution of part occurrence. A promising future direction is thus to explore better priors of object parts.

5. Conclusion

We presented an interpretable deep model for fine-grained classification. Our model leveraged a novel prior of object part occurrence and integrated region-based part discovery and attribution into a deep network. Trained with only image-level labels, our model can predict an assignment map of object parts, an attention map of the part regions and the object label, demonstrating strong results for object classification and object part localization. We believe our model provides a solid step towards interpretable deep learning and fine-grained visual recognition.

Acknowledgment: The authors acknowledge the support provided by the UW-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. The authors also thank Fangzhou Mu for helpful suggestions on the writing.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34(11):2274–2282, 2012. 3

[2] Relja Arandjelovic and Andrew Zisserman. All about VLAD. In *CVPR*, pages 1578–1585, 2013. 5

[3] Salim Arslan, Sofia Ira Ktena, Ben Glocker, and Daniel Rueckert. Graph saliency maps through spectral convolutional networks: Application to sex classification with brain connectivity. In *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities*, 2018. 3

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, pages 6541–6549, 2017. 2

[5] Babak Ehteshami Bejnordi, Tijmen Blankevoort, and Max Welling. Batch-shaped channel gated networks. In *ICLR*, 2020. 4

[6] Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In *ICLR*, 2019. 2

[7] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *CVPR*, pages 4290–4299, 2018. 6

[8] Alina Barnett Jonathan Su Cynthia Rudin Chaofan Chen, Oscar Li. This looks like that: Deep learning for interpretable image recognition. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2019. 2

[9] Tianshui Chen, Liang Lin, Riquan Chen, Yang Wu, and Xiaonan Luo. Knowledge-embedded representation learning for fine-grained image recognition. In *IJCAI*, pages 627–634, 2018. 7

[10] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *CVPR*, pages 5157–5166, 2019. 7

[11] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *CVPR*, pages 433–442, 2019. 1, 3, 4

[12] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In *ECCV*, pages 352–368, 2018. 3, 6, 7

[13] Mark Craven and Jude W Shavlik. Extracting tree-structured representations of trained networks. In *NeurIPS*, pages 24–30, 1996. 2

[14] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *CVPR*, pages 2921–2930, 2017. 7

[15] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *CVPR*, pages 4829–4837, 2016. 2

[16] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification. In *ECCV*, pages 70–86, 2018. 7

[17] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. In *ICML Workshop on Learning Feature Hierarchies*, 2009. 2

[18] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, pages 3429–3437, 2017. 2

[19] Chunhui Gu, Joseph J Lim, Pablo Arbeláez, and Jitendra Malik. Recognition using regions. In *CVPR*, pages 1030–1037, 2009. 3

[20] Emily M Hand and Rama Chellappa. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *AAAI*, pages 4068–4074, 2017. 6

[21] Keke He, Yanwei Fu, Wuhao Zhang, Chengjie Wang, Yu-Gang Jiang, Feiyue Huang, and Xiangyang Xue. Harnessing synthesized abstraction images to improve facial attribute recognition. In *IJCAI*, pages 733–740, 2018. 6

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 5

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[24] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *CVPR*, pages 1173–1182, 2016. 2

[25] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. SCOPS: Self-supervised co-part segmentation. In *CVPR*, pages 869–878, 2019. 1, 3, 5, 6, 7, 9

[26] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 5

[27] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015. 7

[28] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *ECCV*, pages 352–368, 2018. 1, 3

[29] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311, 2010. 5

[30] Mahdi M Kalayeh, Boqing Gong, and Mubarak Shah. Improving facial attribute prediction using semantic segmentation. In *CVPR*, pages 6942–6950, 2017. 6

[31] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(3):302–324, 2009. 3

[32] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, pages 9167–9176, 2019. 1, 3

[33] Yin Li and Abhinav Gupta. Beyond grids: Learning graph representations for visual recognition. In *NeurIPS*, pages 9225–9235, 2018. 1, 3, 4, 5

[34] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*, pages 1666–1674, 2015. 2

[35] Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016. 3

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 1, 2, 5, 6

[37] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *CVPR*, pages 5334–5343, 2017. 6

[38] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, pages 4765–4774, 2017. 2

[39] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *CVPR*, pages 5188–5196, 2015. 2

[40] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *NeurIPS*, pages 7775–7784, 2018. 2

[41] Omkar M Parkhi, Andrea Vedaldi, CV Jawahar, and Andrew Zisserman. The truth about cats and dogs. In *ICCV*, pages 1427–1434, 2011. 2

[42] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale

image classification. In *ECCV*, pages 143–156, 2010. 5

[43] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018. 2, 5

[44] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *ECCV*, pages 51–66, 2018. 8

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *SIGKDD*, pages 1135–1144. ACM, 2016. 2

[46] Ethan M Rudd, Manuel Günther, and Terrance E Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *ECCV*, pages 19–35, 2016. 6

[47] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *NeurIPS*, pages 3856–3866, 2017. 2

[48] 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 *ICCV*, pages 618–626, 2017. 2, 5, 8

[49] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *ICCV*, pages 1143–1151, 2015. 2

[50] Marcel Simon, Erik Rodner, and Joachim Denzler. Part detector discovery in deep convolutional neural networks. In *ACCV*, pages 162–177. Springer, 2014. 2

[51] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[52] Austin Stone, Huayan Wang, Michael Stark, Yi Liu, D Scott Phoenix, and Dileep George. Teaching compositionality to cnns. In *CVPR*, pages 5058–5067, 2017. 3

[53] Yiyou Sun, Sathya N. Ravi, and Vikas Singh. Adaptive activation thresholding: Dynamic routing type behavior for interpretability in convolutional neural networks. In *ICCV*, 2019. 2

[54] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328, 2017. 2

[55] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018. 1, 2, 3, 5

[56] 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, 2, 5, 7, 8

[57] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *CVPR*, pages 4148–4157, 2018. 2, 7

[58] 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 *CVPR*, pages 842–850, 2015. 2

[59] Junjie Yan, Yinan Yu, Xiangyu Zhu, Zhen Lei, and Stan Z Li. Object detection by labeling superpixels. In *CVPR*, pages 5107–5116, 2015. 3

[60] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *ECCV*, pages 420–435, 2018. 7

[61] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014. 2

[62] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *CVPR*, pages 1143–1152, 2016. 2

[63] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *IJCV*, 126(10):1084–1102, 2018. 5

[64] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, pages 834–849, 2014. 2

[65] Quanshi Zhang, Ruiming Cao, Ying Nian Wu, and Song-Chun Zhu. Growing interpretable part graphs on convnets via multi-shot learning. In *AAAI*, pages 2898–2906, 2017. 3

[66] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *CVPR*, pages 8827–8836, 2018. 2

[67] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *CVPR*, pages 1134–1142, 2016. 2

[68] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, pages 5209–5217, 2017. 3, 7

[69] 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 *CVPR*, pages 5012–5021, 2019. 7, 8

[70] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene CNNs. In *ICLR*, 2015. 2

[71] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 2, 8