Recognizing Part Attributes with Insufficient Data

Xiangyun Zhao
Northwestern University

Yi Yang
Baidu Research

Feng Zhou
Baidu Research

Xiao Tan
Baidu Inc.

Yuchen Yuan
Baidu Inc.

Yingze Bao
Baidu Research

Ying Wu
Northwestern University

Abstract

Recognizing attributes of objects and their parts is important to many computer vision applications. Although great progress has been made to apply object-level recognition, recognizing the attributes of parts remains less applicable since the training data for part attributes recognition is usually scarce especially for internet-scale applications. Furthermore, most existing part attribute recognition methods rely on the part annotation which is more expensive to obtain. To solve the data insufficiency problem and get rid of dependence on the part annotation, we introduce a novel Concept Sharing Network (CSN) for part attribute recognition. A great advantage of CSN is its capability of recognizing the part attribute (a combination of part location and appearance pattern) that has insufficient or zero training data, by learning the part location and appearance pattern respectively from the training data that usually mix them in a single label. Extensive experiments on CUB-200-2011 [51], CelebA [35] and a newly proposed human attribute dataset demonstrate the effectiveness of CSN and its advantages over other methods, especially for the attributes with few training samples. Further experiments show that CSN can also perform zero-shot part attribute recognition. The code will be made available at https://github.com/Zhaoxiangyun/Concept-Sharing-Network.

1. Introduction

The computer vision community has seen tremendous progress in recognizing global features of objects, such as performing category detection [44, 15, 43, 68] and classification [24] (e.g. detect the bounding box and classify the category of a bird from an image). Meanwhile, recognizing attributes of object parts (e.g. localize the wing of a bird and classify its biologic feature) is still a very challenging problem due to multiple issues. First, attributes (e.g. the color of the wing of a bird) normally attach to a very limited area of an object, which are usually more difficult to be accurately localized from an image compared to the overall object. Most existing part attribute recognition methods [64, 30, 62] train a part detector with large extra annotations to detect the relevant part. However, such part annotations are very expensive to obtain. Therefore, these methods generally fail when the part annotation is not available. How to recognize the part attribute with only image-level annotation is still under-explored.

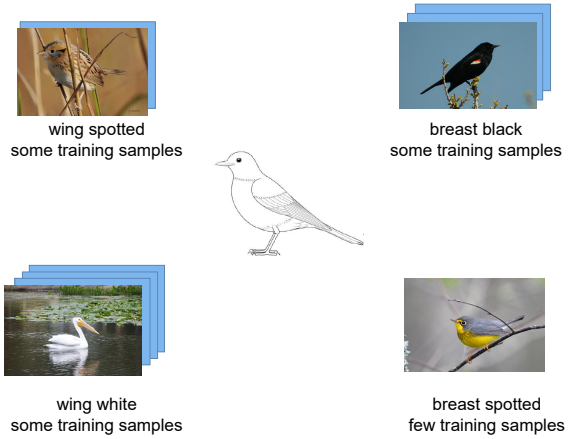

Figure 1. In many datasets and real applications, the labeling of part attributes is usually very limited. For example, as this figure shows, in CUB-200-2011 [51] dataset the labels of *breast spotted* is very few, whereas the number of the labels of *wing spotted*, *wing white*, and *breast black* are more but still limited. We propose to identify the relationships between different labels based on their locations and patterns, so as to re-use the labels of other attributes to facilitate the recognition of the attribute that lacks labels (e.g. *breast spotted* in this figure). Further, we find that the recognition of all these attributes can be jointly improved if individual concepts of different attributes can be shared.

Another important problem is that the training data is expensive to obtain and usually insufficient in the existing dataset. For example, in a commonly used bird parts attribute recognition dataset CUB-200-2011 [51], the number of training images for most attributes varies from merely

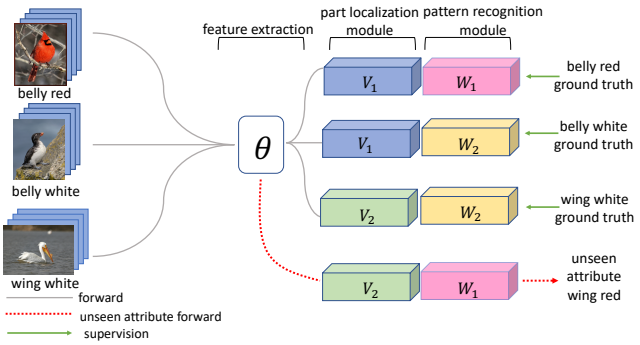

Figure 2. **Overview of the training.** Training images of different attributes are forwarded through the CNN to obtain the image representation, then attribute samples with the same part are forwarded through the same localization module and attribute samples with the same appearance pattern are forwarded through the same pattern recognition module. New attribute with no training data could be recognized as the combination of the learned modules.

a few dozens to at most a few hundred. Most existing attribute recognition methods (if not all) process each part attribute independently, and ignore the spatial correlation of different part attributes. As a result, their performance is simply limited by the volume of training data of each isolated attribute. How to solve the data insufficiency problem is rarely discussed.

To address these challenges, we propose a novel Concept Sharing Network (CSN) for part attribute recognition. In CSN, the part attribute is defined as the combination of two concepts: part location and appearance pattern, as illustrated in Fig. 1. Our neural network models the two concepts as two modules, in contrast to individually modeling each attribute in different branches. Since the two modules in CSN could be shared among different attributes, the labeling of attributes (e.g. color and shape) belong to certain parts (e.g. wing) of an object can be used to facilitate the training of another attribute of the same parts, and vice versa. In such a manner, we maximize the usage of the precious training data to boost the attribute recognition performance independently and aggregately. Note that CSN only needs image-level attributes labels to train, so it would be more general than existing part attribute recognition methods [64, 30, 62] which depend on the part location annotation.

Furthermore, CSN can be used to discover new attributes, *i.e.* zero-shot part attribute recognition. Given a training set with certain attributes, the training result of part localization and pattern recognition in CSN could be used to recognize a new attribute that does not belong to the training set. As illustrated in Fig. 2, after the *wing* location and color (*red*) pattern are learned, a new attribute *wing red* could be recognized even though the training data of such attribute does not exist.

In this work, we also contribute a large-scale human attribute dataset, named as SurveilA, which contains 75,000 images with 10 carefully annotated attributes focusing on the fine-grained human activities for video surveillance. The human images are collected in the wild under different scenes, scales, poses and viewpoint variations. The dataset is challenging as shown in the experiments that simply fine-tuning standard networks cannot provide accurate enough estimation, and recognition would require a model to focus on local discriminative parts.

Overall, our work has the following contributions: 1) We aim at addressing the data insufficiency problem in part attribute recognition, which is rarely discussed in previous work. 2) We present an effective part attribute recognition framework which does not depend on the part annotation. 3) Our network is proven to be effective in zero-shot part attribute recognition. 4) We will release a new dataset for part attribute recognition, which consists of 75000 images of human in a real-world scenario with 10 attributes annotation.

2. Related Work

Attribute Recognition Attribute Recognition was first introduced as a computer vision problem in [12]. From then, attribute recognition have been studied extensively with numerous datasets and methods [11, 10, 26, 25, 27, 28, 55, 67]. Part Attribute recognition is a harder problem because it is only attached to a very limited area of an object. State of the art methods [5, 64, 30] usually rely on the part location annotations to train part detectors such as Poselets [5], Deformable Part Models [63] or R-CNN [16] to first localize parts then extract visual features to recognize attributes [22]. But the part annotation is very expensive to obtain. Although, recently some methods [57, 21, 69] are proposed to localize the important regions for recognition, they are not carefully designed for part attribute recognition and do not attempt to address the data insufficiency problem. [14, 32] used attribute recognition results to facilitate the fine-grained recognition, but both of them will fail when training data is insufficient.

Few-shot / Zero-shot Learning Besides collecting more data, few-shot learning [50] and zero-shot learning [39] attempt to directly address the data insufficiency problem - predicting novel concepts that were either very few or completely unseen from the training set. These problems are classical because almost all in-the-wild data follow a heavy-tail distribution [19] with new classes appearing frequently after the training and no finite set of samples can cover the diversity of the real world. Recently, few-shot learning is modeled as a meta learning problem [42, 13] through explicitly building training loss to enforce adaptation to new categories with few examples. On the other hand, due to the complete lack of training data, zero-shot models at-

tempt to learn to transfer knowledge from other external sources [1, 45, 7, 52, 65, 31, 58]. In contrast to these works, we make use of the visual attention mechanism to disentangle part location with appearance features and share the disentangled representations between attributes, which enable us to conduct zero-shot or few-shot generalization on novel attributes.

Visual Attention and Visual Question Answering Visual attention models [38, 4] have been widely used in object recognition [69, 53], fine-grained recognition [47, 33], image captioning [60] and visual question answering [8]. These models also decompose the location and appearance during representation, but do not focus on addressing the data insufficiency problem. CSN improves the visual attention models on data efficiency by sharing the attention mechanism across multiple attribute labels.

The attention region and feature sharing depend on attribute labels, which is similar to visual question answering (VQA) [3]. In VQA, existing methods [2, 37] also try to localize the relevant region according to the given question. But in VQA, all Q/A pairs use the same classifier (i.e. sharing image and language feature extraction, sharing answer classifier). In contrast, recognition of multiple attributes are usually considered as a multi-task problem with different classifiers to be trained. It is very expensive to collect sufficient data so as to well train each classifier, especially in large-scale recognition. This poses a special challenge of data insufficiency, which is the focus of our method.

Attribute Dataset There are some attribute recognition datasets existing, from general objects and scenes [11, 27, 10, 46, 40, 28, 41, 66] to specific fine-grained classes such as faces [20, 26, 25, 35], birds [51], cars [61], clothes [34] or even butterflies [54]. Due to the importance of human attribute recognition [17, 5, 48, 70, 9, 64, 49, 29, 30] in real world applications and challenges, we propose a large scale challenging dataset for human part attribute recognition.

3. Concept Sharing Network

Part attribute recognition aims to predict whether or not the statement of the attribute of a part is true, e.g. whether or not 'the wing of the bird is black' or 'the bill of the bird is red'. In this work, we introduce a novel Concept Sharing Network(CSN) for part attribute recognition, as illustrated in Fig. 3. In CSN, each attribute is defined as a combination of two concepts: part localization and pattern recognition.

3.1. Network

In CSN, attributes are recognized based on two modules: the part localization module and the appearance pattern recognition module. In our training process, as illustrated in Fig. 2, training images of different attributes are forwarded through the CNN to obtain the image representation, then attribute samples with the same part are forwarded through the same localization module and attribute samples with the same appearance pattern are forwarded through the same pattern recognition module. The modules are learned by the training samples which are forwarded through them. In the inference process, as illustrated in Fig. 3, given an image and recognizing attribute $P_{i,j}$ corresponding to part i and appearance pattern j, the image is forwarded through the corresponding location module and pattern module to obtain the final prediction.

3.1.1 Part Localization module

One novelty of our work is the employment of attention mechanism within CSN neural network to localize the part for attributes. We propose an attention based method inspired by CAM [69]. Note that many other alternatives [53, 56] can also be incorporated with our framework. Given an image x, we use the stack of CNN to extract features at different location over the image. The output of the CNNs at a particular layer is $Q(x;\Theta) \in \mathbb{R}^{hw\times d}$ where h and w are the spatial height and width respectively, d is the number of channels and Θ is the parameters of the CNN adopted. For a specific part, we maintain a learnable vector $V_i \in \mathbb{R}^{d\times 1}$ which is called part representation to encode the associated part. We expect the inner-product value between V_i and features in the feature map $Q(X;\Theta)$ being high in associated regions, while the value being low at other places. As shown in Fig. 3, the localization module takes image representation Q and location representation V_i as input and outputs the inner-product map A_i:

$$A_i(x;V_i,\Theta) = A_i(Q;V_i) = Q(x;\Theta){V_i}^T. \quad (1)$$

We then normalize A_i over spatial domain by the soft-max operation to derive the attention map as

$$A'_i = s(A_i). \quad (2)$$

where s is spatial Softmax function. The attention map is broadcasted over channels and the results are then multiplied back to the feature map resulting an attention weighted feature map Q'_i:

$$Q'_i = Q(x;\Theta)^T \otimes A'_i. \quad (3)$$

where $\otimes$ is the operation of broadcasted multiplication. Note that, the attention weighted feature maps Q'_i are different for different i, though the input feature map Q is the same, which render output features to focus on different spatial location.

3.1.2 Pattern Recognition Module

Conventional attribute recognition methods usually adopt an average pool followed by a fully-connected layer and a

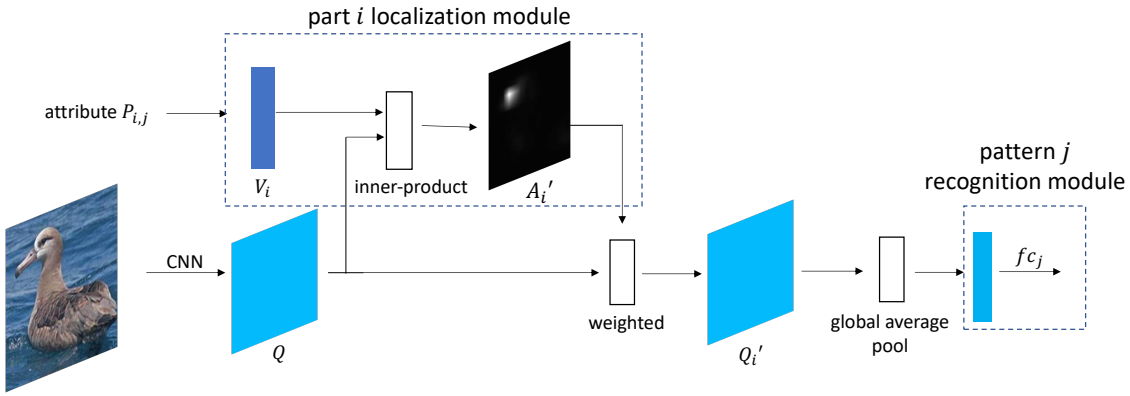

Figure 3. **Overview of the inference.** In the inference, given an image and recognizing attribute $P_{i,j}$, the image is forwarded through the corresponding location module i and pattern module j to obtain the final prediction.

soft-max layer to produce final probability of an attribute. In contrast, our method learns an attention map for each attribute to weight the feature map so that the following appearance pattern could be learned by aggregating all training data of different part locations. Specifically, for each appearance pattern j which is annotated a binary classifier label in part attribute recognition, the probability of predicted label is calculated as,

$$f(x;\Theta,V_i,W_j^T) = softmax(W_j^T \bar{Q}_i'). \tag{4}$$

where $\bar{Q}_i' \in \mathbb{R}^{d\times 1}$ is the global average pooling result of Q_i' over spatial domain, and $W_j \in \mathbb{R}^{d\times 2}$ are appearance specific weights for the binary classification task j.

3.2. Concept Sharing Training

In this part, we describe the concept sharing training process. We denote an attribute by $P_{i,j}$ where the part location index is denoted by i and the pattern indexed is denoted by j. The attribute recognition model has parameters Θ, V_i, and W_j to optimize. We denote the training images for attribute $P_{i,j}$ as $X_{ij} = x_0, x_1, ..., x_{N_{ij}-1}$ whose labels are $Y_{ij} = y_0, y_1, ..., y_{N_{ij}-1}$ where N_{ij} denotes the total number of training samples for the attribute $P_{i,j}$. The Θ, V_i, W_j are trained by an end-to-end fashion using cross-entropy loss at final binary outputs of recognition modules. The overall loss is:

$$L = \sum_{i=0}^{N-1} L(f(x_i;\Theta,V_i,W_j^T), y_i). \tag{5}$$

Where L is the cross-entropy loss and N is the total number of training samples satisfying: $N = \sum_{i,j} N_{ij}$. All attributes in Eq. 5 share the same Θ which are used to extract the CNN features. In our concept sharing training, the localization module is learned from all training samples sharing the same part, and the pattern recognition module is learned from all training samples which sharing the same pattern. Consequently, the weights of sharing models V_i and W_j are updated by:

$$W_j^+ = W_j - \gamma \sum_k \sum_{n=0}^{N_{kj}-1} \frac{\partial L(f(x_n;\Theta,V_k,W_j^T), y_n)}{\partial W_j} \tag{6}$$

, and

$$V_i^+ = V_i - \gamma \sum_k \sum_{n=0}^{N_{ik}-1} \frac{\partial L(f(x_n;\Theta,V_i,W_k^T), y_n)}{\partial Q_i'} \times \frac{\partial Q_i'}{\partial V_i} \tag{7}$$

It is worthwhile to mention that, if the attributes are treated to be independent as in conventional recognition frameworks, training V_i and W_j will only involve training samples (X_{ij}, Y_{ij}). Since the number of training samples of a single attribute is usually small in practice, and therefore the performance of conventional recognition frameworks is limited by the insufficiency of training samples for each isolated attribute. In contrast, the number of training samples for a particular module would be enlarged in our conceptual sharing framework *i.e.* $\sum_k N_{kj}$ for the pattern recognition, and $\sum_k N_{ik}$ for part localization. The training data expansion improves the learning sufficiency with the limited data.

3.3. New Attribute Recognition

In large-scale applications where the number of attributes is large, it is almost infeasible to obtain reasonable amount of qualified training data for every attribute. In this part, we explain how CSN can be used to recognize a new attribute without requiring any training samples. As we discussed above, the localization and recognition module are shared among different attributes, one specific combination of localization and recognition modules determines the recognition for one specific attribute, such as the attribute *breast spotted* in Fig. 1. Notice that we do not have any training data for *breast spotted*. However, we can still train *breast* localization module and *spotted* pattern module from other attributes. Therefore, by combining the learned location *forehead* and pattern *spotted*, *forehead spotted* could be

understandably recognized without any such training data. Specifically, attribute $P_{i,j}$ recognition model is determined by its parameters Θ, W_j and V_i. Since W_j could be learned from attributes $P_{\alpha,j}(\alpha \neq i)$, and V_i could be learned from attributes $P_{i,\beta}(\beta \neq j)$, attribute $P_{i,j}$ recognition model could be determined even without any $P_{i,j}$ data.

4. Experiments

We conduct experiments on three attribute recognition datasets including CUB-200-2011 [51], CelebA [35] and our new SurveilA dataset. Since the positive and negative samples are highly imbalanced in CUB-200-2011 and SurveilA, we use average precision as our main evaluation metric.

4.1. Datasets

The CUB-200-2011 dataset [51] contains 11,788 images of 200 bird categories, where 5994 images are selected for training and the rest 5794 images for testing. Each image is annotated with 312 attributes among which 278 are part related attributes. These part related attributes are all binary attributes indicating whether a specific appearance pattern exists in one particular part such as attribute *wing-black* tells whether the wing is black. During experiments,we observe that the back and tail attributes are nosily labelled. Therefore, we exclude these labels and conduct experiments on the remaining 204 attributes. This is the main dataset for our experimental comparison and ablation study.

CelebA [35] consists of 202,599 face images and 40 binary face attributes. 16,000 images are selected for training and 20,000 images for testing and validation.

Our new SurveilA dataset contains 75,000 images with 10 binary attributes, among which 70,000 are selected for training and the rest 5,000 for testing. The dataset focuses on human part attribute (*e.g.* whether carrying an item, whether wearing shorts or pants), collected under real surveillance scenarios with large pose and appearance variations. We will release the dataset to facilitate the research of attribute recognition.

4.2. Implementation Details

We first resize the image to be 512×512, then randomly cropped 446×446 for training. We use ResNext50 [59] as the visual representation module for feature extraction. The outputs from the layer 'conv5' in ResNext50 are used as the feature representation for part localization *i.e.* Q in eq. 1.The network is trained for 100 epochs with ADAM[23] where initial learning rate is set to 0.0001 with a learning rate decay of 0.1 after 50 epochs.

attribute	no share	part share	pattern+part share
belly-solid	83.6%	85.9%	85.5%
breast-white	81.3%	82.1%	81.9%
bill-grey	44.7%	46.0%	47.0%
bill-black	78.2%	79.4%	77.0%

Table 1. Average precision given 500 training images

attribute	no share	part share	pattern+part share
belly-solid	79.2%	84.0%	84.9%
breast-white	76.5%	79.7%	80.5%
bill-grey	38.5%	42.2%	46.8%
bill-black	74.9%	78.6%	76.0%

Table 2. Average precision given 200 training images

4.3. Experiments on CUB-200-2011 Dataset

4.3.1 Study the Number of Training Samples

In this part, we empirically study the the effectiveness of CSN as the number of training samples varies. In order to study this, we select attributes *belly-solid*, *breast-white*, *bill-grey* and *bill-black* which could share the same part and have relatively larger positive samples. All experiments are run on joint training of 91 attributes(*i.e.* all bill relevant attributes, all grey attributes, all black attributes, all belly, all breast, all white). We evaluate the performance of the CSN w/o(without) share, CSN w/part share, CSN w/part + pattern share on the 3 attributes selected. The results are shown in Tab. 2 and Tab. 1 We see the benefits of sharing attributes is more obvious when the number of training samples is small. The pattern features extracted at different location still varies, forcing them to be the same pattern when they have relatively large data will harm the performance. But when the training data size decreases, their own data is insufficient to learn the pattern module. Pattern sharing will improve the learning efficiency of the pattern module.

We also try to decrease the training samples to zero. CSN still obtains promising results for *belly-solid* (84.7% AP), *breast-white* (79.9% AP), *bill-grey* (46.3% AP), *bill-black* (64.2% AP) that are comparable to the supervised training. We will further investigate the effectiveness of CSN for zero-shot recognition in the following section.

4.3.2 Study the Number of Sharing Attributes

In this part, we empirically study the effects of the number of sharing attributes on the overall performance. In Tab. 5, we compare sharing different number of attributes in pattern module and localization module. We select attribute *bill-curved*,*bill-brown*,*bill-orange* and *bill-red*. We perform experiments on joint training of all bill attributes, all brown attributes, all orange attributes and all red attributes. We com-

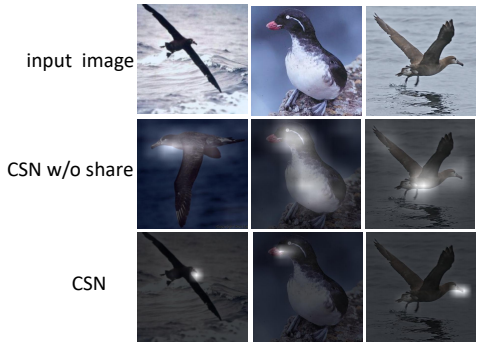

Figure 4. Comparison of heat map generated by CSN w/o sharing and full CSN for the part "bill".

Baseline	CAM [69]	STN [21]	PANDA [64]	R-CNN [62]	CSN
63.1%	63.2%	63.7%	64.9%	65.5%*	65.2%

Table 3. Average Precision(AP) comparison. The numbers are shown only for the 32 attributes that contain more than 1000 positive samples in this dataset. If we perform such comparison for attributes with smaller training set, the baseline always produces very poor result.

CSN w/o share	CSN	CSN-soft	CSN-soft-1
63.9%	65.2%	65.1%	63.8%

Table 4. Average Precision(AP) comparison. The numbers are shown only for the 32 attributes that contain more than 1000 positive samples in this dataset. If we perform such comparison for attributes with smaller training set, the baseline always produces very poor result.

pare the four attributes performance under different number of sharing part attributes and sharing pattern attributes.

From (a) - (c) in Tab. 5, as the number of sharing part attributes increases, the overall performance steadily increases. This illustrates that it is more effective when number of sharing part attributes become larger. Note that as the number of sharing pattern attributes increases to be 9, the improvement is observed, this also indicates the sharing pattern is effective.

4.3.3 Comparison with the State-of-the-art

We first compare with state-of-the-art recognition method with only image level annotation CAM [69], Spatial Transform Network(STN) [21] and our baseline (a multi task learning framework with different branches). STN aims at explicitly localizing important regions for recognition. We used the public implementation and replace the recognition by our part attribute recognition. Our CAM implementation follows [69] (i.e. localize and crop) for recognition. Tab. 3 shows the performance of our method against other methods. Since the baseline always shows very low performance

	bill-curved	bill-brown	bill-orange	bill-red
(a) $N_1 = 0, N_2 = 0$	19.1%	25.6%	38.6%	35.8%
(b) $N_1 = 4, N_2 = 0$	20.1%	25.9%	39.4%	36.7%
(c) $N_1 = 24, N_2 = 0$	21.4%	26.7%	42.7%	38.1%
(d) $N_1 = 24, N_2 = 9$	-	26.9%	43.0%	38.3%

Table 5. Average Precision(AP) comparison with different sharing attribute number. N_1 is the number of sharing part attributes, and N_2 is the number of sharing pattern attributes. Note: *curved* attribute only exists in *bill-curved*, so (d) for *bill-curved* is not available.

with attributes with small number of training images, we select the attributes with more than 1000 positive images for a fair comparison. Such selection leads to 32 attributes (e.g. as white relevant attributes and black relevant attributes) in the experiments. CSN produces AP of 65.2% compared to the competitors: 63.2% from CAM, 63.7% from STN and 63.1% of CSN baseline. This illustrates the advantage of CSN over alternatives given a reasonably large training set. We investigate all 204 part attributes to further understand the overall performance of different methods. Fig. 5 shows the AP difference between of CSN and the baseline on all 204 attributes. Significant improvement over the baseline is observed on average.

We then compare with state-of-the-art attributes recognition methods [64, 62]. We used the public implementation [64, 6, 62] to train a parts detector(i.e. poselets [6] and R-CNN [62]) with part annotation and then recognize the attributes. Without part annotation, we still obtain comparable performance with PANDA [64] and R-CNN based method [62]. Since they both depend on the part annotations to train the part detector, they fail when the part annotation is not available. Furthermore, they both can not be used for zero shot recognition while ours can.

We also visualize the localization heat map(*i.e.* the attention map) obtained by CSN w/o share and CSN in Fig. 4. The localization heat map obtained by CSN is obviously better than that obtained by CSN w/o share. This further verifies the effectiveness of the concept sharing.

4.3.4 Soft Sharing among Attributes

In the above section, attributes with the same part are manually grouped to the same part localization module, and attributes with the same pattern are manually grouped to the same pattern module. Accordingly, attributes with different concepts (localization / pattern) can not share the same module. In this part, we study the soft sharing among parts. We add a learnable soft weight vector $T_i = t_1, t_2, .., t_m$ for attribute k, where m is number of parts modules. This weight vector is used to combine the attention map from different part. That is

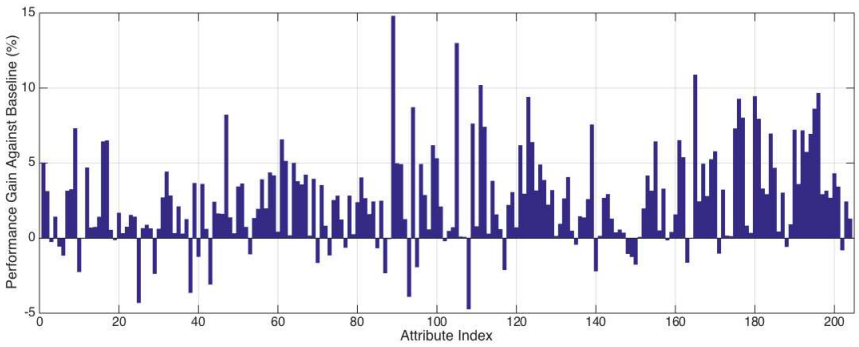

Figure 5. CSN performance gain against baseline on joint training of 204 attributes. The baseline method refers to a multi task learning framework with 204 branches. Please refer to the supplementary materials for the specific name of each attribute id.

Figure 6. Qualitative results of CSN on CUB-200-2011 test set. For each pair of images in the figure, the left shows the input image. The right shows the location heat map predicted by CSN. The bottom text shows the predicted attribute. More visualization results are in supplementary materials.

$$A'_k = \sum_{i=0}^{m-1} t_i A_i. \tag{8}$$

We first initialize the vector to be one hot vector(*i.e.* one value is 1, others are 0 in T_i). Let the attributes sharing the same part have the same initialization, and let them learnable during the training. This obtains 65.1%(*i.e.* CSN-soft in Tab. 3) which is comparable with CSN. We then initialize the vectors as 1(*i.e.* all values in T_i are initialized as 1). This indicates that we do not have the prior knowledge which attributes are of the same part. We obtain 63.8%(*i.e.* CSN-soft-1 in Tab. 3) which is lower than 65.1% of the previous one. This indicates that this prior knowledge provide important information.

4.3.5 Zero-Shot Attribute Recognition

In this part, we study the capability of CSN on zero shot attribute recognition. The experiments are carried out on all 204 attributes where we randomly select 20 attributes as unseen attributes and leave the other 184 attributes for training the network. In Tab. 8, zero-shot algorithm refers to CSN w/ sharing pattern and part with no training data on the 20 attributes, this is the algorithm for zero shot in Tab. 8, and supervised refers to CSN w/o sharing trained on all available data. Since the performance gap between zero-shot and supervised is highly relevant to the size of train data, we also list the number of positive samples in statistics. As shown in Tab. 8, zero-shot obtains promising results on the condition of zero shot learning. Note that attributes with very small data, such as *'forehead purple'*, *'wing olive'*, *'underparts green'*, zero shot even outperforms results trained on their own data. For most of attributes, zero shot algorithm is surprisingly effective as it shows accuracy comparable with trained on their own data.

4.4. Experiments on CelebA

Our method can also be applied on the general attributes(i.e. global and parts attributes), so we also perform experiments on a general attributes dataset CelebA [35]. We follow the protocol in [18]. In table 7, we evaluate the average accuracy which is usually reported for CelebA. Our CSN obtains better performance than our baseline and beats the state of the art methods. This is because all other methods fail to explicit localize the important regions for recognition. In CSN, we group the parts attributes to share the same localization module such as nose related and mouth related. We observe further improvement by sharing the localization module. This indicates that the concept sharing is still effective.

Attribute ID	1	2	3	4	5	6	7	8	9	10
# of training imgs	51968	12050	4982	3183	1334	542	320	276	209	147
Baseline	89.5%	47.1%	57.0%	38.0%	49.0%	7.1%	5.9%	6.3%	2.2%	0.5%
CSN	**96.3%**	**65.8%**	**75.6%**	**67.3%**	**84.1%**	**23.4%**	**32.2%**	**38.3%**	**7.9%**	**30.3%**

Table 6. Average Precision(AP) on our new human attribute test set. The attributes 1-10 are some of the most useful ones in security surveillance applications: 1 *the length of sleeve (short / long)*, 2 *the length of pants (long / short)*, 3 *using a cell phone or not*, 4 *carrying an item or not*, 5 *dragging a luggage or not*, 6 *smoking or not*, 7 *wearing a glove or not*, 8 *holding a baby in arm or not*, 9 *wearing a mask or not*, 10 *holding an umbrella or not*. The 2nd row shows the number of the training samples. It can be observed that CSN obtains higher performance gain for attributes with fewer training samples.

Baseline	Multi-task [18]	Fully [36]	CSN w/o share	CSN
91.1%	91.2%	91.3%	91.7%	**91.8%**

Table 7. Comparison of attribute classification accuracy on CelebA test set.

attribute	#images	zero-shot	supervised
belly_solid	2455	86.1%	88.1%
upperparts_black	1561	74.5%	78.8%
crown_black	1434	75.7%	82.9%
underparts_black	730	67.3%	74.5%
breast_multi-colored	626	30.8%	37.9%
crown_buff	413	38.1%	39.2%
underparts_brown	360	41.2%	47.8%
belly_striped	319	36.5%	41.5%
wing_yellow	316	52.3%	63.9%
belly_brown	313	36.7%	44.8%
wing_spotted	300	37.0%	51.6%
bill_yellow	215	9.2%	50.3%
throat_red	175	65.6%	70.0%
belly_red	140	68.3%	73.1%
wing_olive	127	33.2%	30.6%
upperparts_orange	81	18.7%	28.1%
wing_iridescent	74	12.0%	13.0%
belly_olive	67	13.9%	23.9%
underparts_green	38	25.2%	16.0%
forehead_purple	22	11.2%	4.2%

Table 8. Average Precision(AP) comparison between zero shot learning and supervised learning on 20 attributes. Zero shot is CSN with part and pattern sharing, supervised is CSN w/o sharing.

4.5. Experiments on SurveilA Dataset

Tab. 6 shows that in this dataset CSN achieves 51.2% mAP compared to baseline mAP 30.3%. Such a huge performance improvement is achieved since most of these human attributes only depend on a small part of the image (e.g. *wearing a mask or not* are only related to the face area). Hence, the localization function of CSN establishes higher importance than the CUB-200-2011 dataset. In Fig. 7, we visualize the parts localization obtained by our methods.

Figure 7. Exemplar images of our human dataset. The heat map placed on the right of each image visualizes the localization information predicted in the CSN inference process. We see that most attributes are only associated with a very small area in images.

The experiments on humans further verify that our proposed method is effective and reliable for recognizing parts of different types of objects.

5. Conclusion

In this paper, we proposed a new neural network structure (CSN) for part attribute recognition. By identifying part locations and appearance patterns from the training data that do not explicitly label these two concepts, CSN can increase part attribute recognition accuracy especially if the number of labels is small. In the special case of data limitation where none data is available, CSN is still valid to recognize attributes (*i.e.* zero-shot part attribute recognition).

Acknowledgements

This work was supported in part by National Science Foundation grant IIS-1619078, IIS-1815561, and the Army Research Office ARO W911NF-16-1-0138.

References

[1] Ziad Al-Halah and Rainer Stiefelhagen. How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes. In *WACV*, pages 837–843. IEEE, 2015. 3

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, pages 39–48, 2016. 3

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 3

[4] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014. 3

[5] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, pages 1543–1550. IEEE, 2011. 2, 3

[6] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, pages 1365–1372. IEEE, 2009. 6

[7] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, pages 52–68. Springer, 2016. 3

[8] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015. 3

[9] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792. ACM, 2014. 3

[10] Ali Farhadi, Ian Endres, and Derek Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, pages 2352–2359. IEEE, 2010. 2, 3

[11] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785. IEEE, 2009. 2, 3

[12] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *NIPS*, pages 433–440, 2008. 2

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017. 2

[14] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1349–1358, 2017. 2

[15] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 1

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158, 2016. 2

[17] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, pages 1–7. Citeseer, 2007. 3

[18] 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*, 2017. 7, 8

[19] Bruce M Hill. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174, 1975. 2

[20] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 3

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015. 2, 6

[22] Jungseock Joo, Shuo Wang, and Song-Chun Zhu. Human attribute recognition by rich appearance dictionary. In *CVPR*, pages 721–728, 2013. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 1

[25] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011. 2, 3

[26] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *ICCV*, pages 365–372. IEEE, 2009. 2, 3

[27] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958. IEEE, 2009. 2, 3

[28] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 2, 3

[29] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016. 3

[30] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *ECCV*, pages 684–700. Springer, 2016. 1, 2, 3

[31] Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, and Yueting Zhuang. Zero-shot recognition using dual visual-semantic mapping paths. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5207–5215, 2017. 3

[32] Xiao Liu, Jiang Wang, Shilei Wen, Errui Ding, and Yuanqing Lin. Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[33] 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

[34] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016. 3

[35] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 1, 3, 5, 7

[36] 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. 8

[37] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *CVPR*, pages 4942–4950, 2018. 3

[38] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014. 3

[39] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, pages 1410–1418, 2009. 2

[40] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758. IEEE, 2012. 3

[41] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *ECCV*, pages 85–100. Springer, 2016. 3

[42] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016. 2

[43] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 1

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1

[45] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015. 3

[46] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. In *ECCV*, pages 1–14. Springer, 2010. 3

[47] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. *arXiv preprint arXiv:1412.7054*, 2014. 3

[48] Gaurav Sharma and Frederic Jurie. Learning discriminative spatial representation for image classification. In *BMVC*, pages 1–11. BMVA Press, 2011. 3

[49] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *ICCV Workshops*, pages 87–95, 2015. 3

[50] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? In *NIPS*, pages 640–646, 1996. 2

[51] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 3, 5

[52] Donghui Wang, Yanan Li, Yuetan Lin, and Yueting Zhuang. Relational knowledge transfer for zero-shot learning. In *AAAI*, volume 2, page 7, 2016. 3

[53] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017. 3

[54] Josiah Wang, Katja Markert, and Mark Everingham. Learning models for object recognition from natural language descriptions. In *BMVC*, volume 1, page 2, 2009. 3

[55] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *ICCV*, pages 531–540, 2017. 2

[56] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[57] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, September 2018. 2

[58] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. *arXiv preprint arXiv:1703.04394*, 2017. 3

[59] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5987–5995. IEEE, 2017. 5

[60] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015. 3

[61] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, pages 3973–3981, 2015. 3

[62] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, pages 834–849. Springer, 2014. 1, 2, 6

[63] Ning Zhang, Ryan Farrell, Forrest Iandola, and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *CVPR*, pages 729–736, 2013. 2

[64] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, pages 1637–1644, 2014. 1, 2, 3, 6

[65] Ziming Zhang and Venkatesh Saligrama. Zero-shot recognition via structured prediction. In *ECCV*, pages 533–548. Springer, 2016. 3

[66] Bo Zhao, Yanwei Fu, Rui Liang, Jiahong Wu, Yonggang Wang, and Yizhou Wang. A large-scale attribute dataset for zero-shot learning. *arXiv preprint arXiv:1804.04314*, 2018. 3

[67] Xiangyun Zhao, Haoxiang Li, Xiaohui Shen, Xiaodan Liang, and Ying Wu. A modulation module for multi-task learning with applications in image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–416, 2018. 2

[68] Xiangyun Zhao, Shuang Liang, and Yichen Wei. Pseudo mask augmented object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4061–4070, 2018. 1

[69] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 2, 3, 6

[70] Jianqing Zhu, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *CVPR Workshops*, pages 331–338, 2013. 3