

Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention

Dat Huynh
Northeastern University
huynh.dat@husky.neu.edu

Ehsan Elhamifar
Northeastern University
eelhami@ccs.neu.edu

Abstract

We address the problem of fine-grained generalized zero-shot recognition of visually similar classes without training images for some classes. We propose a dense attribute-based attention mechanism that for each attribute focuses on the most relevant image regions, obtaining attribute-based features. Instead of aligning a global feature vector of an image with its associated class semantic vector, we propose an attribute embedding technique that aligns each attribute-based feature with its attribute semantic vector. Hence, we compute a vector of attribute scores, for the presence of each attribute in an image, whose similarity with the true class semantic vector is maximized. Moreover, we adjust each attribute score using an attention mechanism over attributes to better capture the discriminative power of different attributes. To tackle the challenge of bias towards seen classes during testing, we propose a new self-calibration loss that adjusts the probability of unseen classes to account for the training bias. We conduct experiments on three popular datasets of CUB, SUN and AWA2 as well as the large-scale DeepFashion dataset, showing that our model significantly improves the state of the art.

1. Introduction

Fine-grained recognition, which is to classify categories that are visually very similar, is an important yet challenging task with a wide range of applications from fashion industry, e.g., recognition of different types of shoe or cloth [1, 2, 3], to face recognition [4, 5, 6] and environmental conservation, e.g., recognizing endangered species of birds or plants [7, 8, 9, 10, 11, 12]. However, training fine-grained classification systems is challenging, as collecting training samples from every class requires costly annotations by domain experts to distinguish between similar classes, e.g., ‘Parakeet Auklet’ and ‘Least Auklet’ bird species, see Figure 3. As a result, training samples often follow a long-tail distribution [13, 14], where many classes have few or no training samples. In this work, we aim to generalize fine-grained recognition to new classes without training samples

Figure 1: Traditional zero-shot classification (top) compresses visual features to perform global embedding with class semantic descriptions, hence, not efficiently capturing fine-grained discriminative visual information. Our method (bottom) finds local discriminative regions through dense attribute-based attention and individually embeds each attribute-based feature with the attribute semantic description, allowing for knowledge transfer to unseen classes while preserving all fine-grained details.

via capturing and transferring fine-grained knowledge from seen to unseen classes without overfitting on seen classes.

Although fine-grained classification has achieved remarkable performance by using feature pooling [15, 16, 9] and discriminative region localization [11, 7, 3, 1, 12, 2] techniques, it cannot generalize to unseen classes, as it requires training samples from every class, and cannot leverage auxiliary information such as class semantic vectors, which is fundamental for transferring knowledge to unseen classes. With the need for costly training data, conventional fine-grained classification methods cannot scale to a large number of classes. However, fine-grained classes can often be described in terms of attributes that are common among classes. Thus, effectively using these semantic descriptions to transfer knowledge among classes can significantly reduce the amount of annotations for training.

Zero-shot classification, on the other hand, leverages auxiliary information in the form of class semantic descriptions to generalize to unseen classes [17, 18, 19, 20]. A large group of existing work learns an embedding function

Figure 2: The overview of our proposed fine-grained zero-shot learning based on dense attribute-based attention with attribute embedding and self-calibration. Image features of R regions are extracted and fed into our dense attention mechanism to compute attention features for all attributes. The attention features are then aligned with attribute semantic vectors to measure the scores of attributes in the image, which are combined to form the final prediction.

to align the visual feature of an image with its class semantic vector, which allows to classify test images from both seen and unseen classes [18, 19, 20, 21, 12]. However, these works rely on holistic image features that are insufficient for distinguishing fine-grained classes, where the discriminative information is contained in a few regions corresponding to a few attributes. While feature synthesis techniques [22, 23, 24, 25], which learn to generate images features using class semantic vectors and convert the problem to the standard classification, have achieved strong results for zero-shot learning, they only synthesize high-level image features, which cannot capture fine-grained differences in details of seen/unseen classes.

Few works [26, 27, 28, 29] have explored localizing informative image regions for fine-grained zero-shot learning. [28] assumes access to ground-truth discriminative parts during both training and testing, which is restrictive. On the other hand, [26, 27, 29] can only scale up to a dozen attention modules without exploiting visual guidance from attribute semantics, with [26] requiring access to the costly part annotations during training. While emphasizing on discriminative regions of images, these work build a global image feature vector which, similar to the prior work, is aligned with the class semantic vector, incorporating all attributes, see Figure 1. Orthogonal to this direction, recent work [30, 31] show that adjusting the influence of different attributes can significantly improve the performance, yet, they rely on holistic image features, which cannot capture fine-grained discriminative regions.

Paper Contributions. In this paper, we develop a new framework for fine-grained zero-shot learning that addresses limitations of the existing work, discussed above. We propose a *dense attribute-based attention mechanism* that for each attribute focuses on the most relevant image regions, obtaining attribute-based features. Our attribute-based attention model is guided by each *attribute semantic vector*, hence, building the same number of feature vectors as the number of attributes. Instead of aligning a combination of all attribute-based features with the true class seman-

tic vector, we propose an *attribute embedding technique* that aligns each attribute-based feature with its attribute semantic vector, see Figure 2. Hence, we compute a vector of attribute scores, for the presence of each attribute in an image, whose similarity with the true *class semantic vector* is maximized. Moreover, we adjust each attribute score using an attention model over attributes to better capture the discriminative power of different attributes. Thus, our model handles classes that are different in only a few attributes.

To tackle the challenge of bias towards seen classes during testing, we propose a new self-calibration loss that adjusts the probability of unseen classes to account for the training bias. We conduct experiments on three popular datasets of CUB, SUN and AWA2. Moreover, we perform experiments on the much larger DeepFashion dataset for fine-grained generalized zero-shot cloth recognition. By extensive experiments, we show that our model significantly improves the state of the art.

2. Related Work

The goal of fine-grained recognition is to capture the small but discriminative features across different classes. [15, 16, 9, 32] captures the interaction between discriminative feature maps through pooling technique while [33, 34, 35] propose better ways to learn global image features that capture fine-grained details. On the other hand, [12, 36] localize discriminative parts of images through part-based supervision. To avoid the localization annotation of discriminative parts, [37, 10, 7, 38] localize them in a weakly supervised setting. Despite tremendous success in the fully supervised setting, these works cannot generalize to zero-shot learning where only high-level attribute descriptions are given for unseen classes.

To deal with unseen classes without training samples, [18, 19, 20, 21, 12] propose to learn an embedding function that aligns visual and semantic modalities where unseen classes are recognized based on the distance between visual features and unseen attribute descriptions. Recently, generative methods [22, 23, 24, 25] have shown great po-

Figure 3: Visualization of three fine-grained classes and their attribute descriptions from the CUB dataset. Notice that these classes are different only in a few attributes.

tential in zero-shot learning by synthesizing features of unseen classes based on attribute descriptions, transforming the problem into the traditional supervised learning with full training samples. However, they can only generate high-level image features and ignore discriminative local regions of images. [26, 27] introduce attention mechanisms for zero-shot learning to capture finer details. Attention can also be recursively learned in a hierarchical fashion [39, 40]. However, these works are designed for sequential input, thus are not suitable for image recognition.

Unseen class bias is the direct consequence of the domain shift [41] between training and testing time, where the model overfits on classes seen at training time. Thus, [41, 42] propose transductive zero-shot learning methods, where the model has access to unlabeled samples from unseen classes during training to learn the testing distribution. However, it is yet costly to collect unseen samples even without labels. Other approaches have been explored, such as prediction smoothing based on similarity between seen and unseen attribute descriptions [43], prediction calibration [44, 45] and novelty detection [46, 47]. These works either trade off discriminative power for unseen class accuracy or are non-differentiable and do not allow effective end-to-end training.

3. Visual Attention Review

Visual attention generates a feature from the most relevant region of an image and has been shown to be effective for image classification, saliency detection and captioning, among others [48, 49, 50, 51, 52, 53]. More specifically, one first divides an image I into R regions denoted by $\{I^r\}_{r=1}^R$, which can be arbitrary [54] or equal-size grid cells [52]. For simplicity and reproducibility, we use the latter approach. Let $\mathbf{f}^r = f(I^r)$ denote the feature vector of the region r , extracted using a ResNet-101 pretrained on ImageNet. Given region features $\{\mathbf{f}^r\}_{r=1}^R$, the goal of the attention module, $g(\cdot)$, is to find the most relevant regions for the task. This is done by finding an attention feature, \mathbf{f} , which is defined as

$$\mathbf{f} = g(\mathbf{f}^1, \dots, \mathbf{f}^R) = \sum_{r=1}^R \alpha_r(\mathbf{f}^r) \mathbf{f}^r, \quad (1)$$

where $\alpha_r(\mathbf{f}^r)$ denotes the weight or preference of selecting the region r . These weights are unknown and the task

of the attention module is to find them for an input image. In the soft-attention mechanism [52], which we use in the paper, one assumes that $\alpha_r \in [0, 1]$ and $\sum_{r=1}^R \alpha_r = 1$ to select different regions with different weights. The attention weights are often normalized using the softmax function.

4. Dense Attribute-Based Attention for Fine-Grained Generalized Zero-Shot Learning

In this section, we discuss our proposed attribute-based attention with attention over attributes for recognizing seen and unseen fine-grained classes in addition to our self-calibration loss to prevent bias towards seen classes. We first define the problem and then present our approach.

4.1. Problem Setting

Assume we have two sets of classes C_s and C_u , where C_s denotes seen classes that have training samples, C_u denotes unseen classes without training samples and $C = C_s \cup C_u$ denotes the set of all classes. Let $(I_1, y_1), \dots, (I_N, y_N)$ be N training samples, where I_i denotes the i -th training image and $y_i \in C_s$ corresponds to its class.

The goal of generalized zero-shot learning is to classify a test image that could belong to a seen or an unseen class.¹ Given that there are no training images for unseen classes, C_u , similar to existing work on (generalized) zero-shot learning [55, 56, 57, 20], we assume access to *class semantic* vectors $\{\mathbf{z}^c\}_{c \in C}$ that provide descriptions of classes. More specifically, $\mathbf{z}^c = [z_1^c, \dots, z_A^c]$ is the semantic vector of the class c with A attributes, where z_a^c denotes the score of having the a -th attribute in the class c [55, 56, 57, 20, 58, 59]. We normalize each \mathbf{z}^c to have unit Euclidean norm. Similar to [57], we assume both seen and unseen class semantic vectors are available at the training time. In addition, we assume access to *attribute semantic* vectors $\{\mathbf{v}_a\}_{a=1}^A$, where \mathbf{v}_a denotes the average GloVe representations of words in the a -th attribute, e.g., ‘yellow beak’. We allow the attribute semantic vectors to be refined during training (see below for details).

4.2. Proposed Framework

In this section, we present our method for fine-grained generalized zero-shot classification. For each attribute, our method extracts a spatial attention feature from the most relevant regions of the input image, which will be subsequently used to find a compatibility score between the attribute semantic vector and the attribute-based image feature. We use the attribute-image compatibility scores and class semantic vectors to define the score of the image belonging to each class. To incorporate the utility of each attribute for computing the class score, we further scale

¹This is more challenging than the traditional zero-shot learning, which assumes a test image can only belong to an unseen class.

the attribute-image compatibility scores by attention over attributes. To learn the spatial and attribute attention networks and the parameters of the attention-image compatibility function, we propose a loss function that augments the standard cross-entropy loss with a new self-calibration loss that prevents the prediction bias towards only seen classes.

4.2.1 Attribute Localization via Dense Attention

The ability to learn visual models of attributes is crucial for transferring knowledge from seen to unseen classes. Recent work either embed image features into the class semantic space [18, 19, 20, 21, 12] or generate image features from class semantic vectors [22, 23, 24, 25]. However, without localizing each attribute, they ignore discriminative visual features of fine-grained classes, obtaining holistic features that contain information from non-discriminative or irrelevant image regions.

As the first component of our method, we propose an attribute-based spatial attention model, where for each attribute, we localize the most relevant image regions to the attribute to extract an *attribute-based attention feature* from a given image. Recall that $\{\mathbf{v}_a\}_{a=1}^A$ is the set of attribute semantic vectors and $\{\mathbf{f}_i^r\}_{r=1}^R$ denotes the region features of the image i . For the a -th attribute, we define its attention weights of focusing on different regions of image i as

$$(\mathbf{f}_i^r, \mathbf{v}_a) = \frac{\exp(\mathbf{v}_a^T \mathbf{W} \mathbf{f}_i^r)}{\sum_r \exp(\mathbf{v}_a^T \mathbf{W} \mathbf{f}_i^r)}, \quad (2)$$

where \mathbf{W} denotes a learnable matrix to measure the compatibility between each attribute semantic vector and the visual feature of each region. Using the set of attention weights $\{(\mathbf{f}_i^r, \mathbf{v}_a)\}_{r=1}^R$, we compute the *attribute-based attention feature* for the a -th attribute as

$$\mathbf{h}_i^a = \sum_{r=1}^R (\mathbf{f}_i^r, \mathbf{v}_a) \mathbf{f}_i^r. \quad (3)$$

Thus, \mathbf{h}_i^a represents the visual feature of the image i that is relevant to the a -th attribute according to the semantic vector \mathbf{v}_a . Notice that when an attribute is absent in the image, \mathbf{h}_i^a captures the visual evidence used to reject the attribute in the image. For instance, the model could focus on ‘back belly’, and later assigns a negative score to it, to indicate the absence of ‘white belly’, as shown in Figure 5.

4.2.2 Attribute Embedding with Attribute Attention

Given the set of attribute-based attention features $\{\mathbf{h}_i^a\}_{a=1}^A$ for each training image i , our goal is to compute the *class score* s_i^c of the image i belonging to a class c . During training, the class score would be optimized to be large for the ground-truth class $c = y_i$ and small for other classes $c \neq y_i$. To do so, we define A *attribute scores*, where each score

measures the strength of having each attribute in the image (recall A is the number of attributes). We fuse these scores using each class semantic vector to find the class score.

More specifically, we define the attribute score e_i^a , as the confidence of having the a -th attribute in the image i , by matching the attribute attention feature \mathbf{h}_i^a with the attribute semantic vector \mathbf{v}_a ,

$$e_i^a = \mathbf{v}_a^T \mathbf{W}_e \mathbf{h}_i^a, \quad (4)$$

where \mathbf{W}_e is an embedding matrix that embeds the attribute feature \mathbf{h}_i^a to the a -th attribute semantic space. In fact, when the attribute is visually present in an image, the associated image feature would be projected near its attribute semantic vector. One way to compute the class score s_i^c is to use the sum of products between each attribute score e_i^a and the strength of having the attribute a in class c , i.e. z_a^c , as

$$s_i^c = \sum_{a=1}^A e_i^a \times z_a^c. \quad (5)$$

As a result, when a class c has attribute a , i.e., $z_a^c > 0$, we would maximize the attribute score e_i^a .

However, one possible limitation of (5) is that all attributes contribute to the class score. In the fine-grained recognition, different classes often have many similar attributes and only a few of the attributes are different between them, see Figure 3. Hence, to focus on important attributes, we propose an attention mechanism over attributes to give the model the ability of selecting attributes that would be most informative for classification and distinguishing between similar classes. More specifically, we compute the utility of each attribute a based on its attention feature \mathbf{h}_i^a as

$$(\mathbf{h}_i^a, \mathbf{v}_a) = \frac{\exp(\mathbf{v}_a^T \mathbf{W} \mathbf{h}_i^a)}{\exp(\mathbf{v}_a^T \mathbf{W} \mathbf{h}_i^a) + 1}, \quad (6)$$

where \mathbf{W} is a learnable matrix. Having the attribute attention weights $\{(\mathbf{h}_i^a, \mathbf{v}_a)\}_{a=1}^A$, we propose to compute the class score by

$$s_i^c = \sum_{a=1}^A e_i^a \times z_a^c \times (\mathbf{h}_i^a, \mathbf{v}_a), \quad (7)$$

where the model adjusts the influence of each attribute on the final prediction by setting the attribute attention score. More specifically, it sets $(\mathbf{h}_i^a, \mathbf{v}_a) = 0$ when the attribute a encoded by \mathbf{h}_i^a cannot be aligned with the semantic vector \mathbf{v}_a and should not be used for prediction.

Notice that unlike spatial attention in (2), where we use a softmax function, hence, ideally focus on one image region, in (7), we use a sigmoid function for each attribute individually, which allows to select multiple attributes with weights

close to one, and set the weight for the remaining attributes to be close to zero. It is worth noting that e_i^a in (4) and $a(h_i^a, v_a)$ in (6) have *complementary roles*: e_i^a captures the presence or absence of the attribute in an image, while $a(h_i^a, v_a)$ acts as a gating mechanism, which determines how much e_i^a should influence the final prediction.

Remark 1 Notice that instead of computing the class compatibility score between a class semantic vector and a global image feature, we first compute A compatibility scores between each attribute-based attention feature and each attribute semantic vector and automatically select a subset of these scores to form the class compatibility score. This gives our model the ability to use a rich set of features, based on localization of each attribute in an image, and incorporate the most discriminative ones for classification.

4.2.3 Loss Function with Self-Calibration Component

In order to find the parameters of our model, we need to optimize the cross-entropy loss between the model prediction and the ground-truth label over training images, i.e.,

$$L_{ce} \{s_i^c\}_{c \in C} = - \sum_i \log p(s_i^{y_i}). \quad (8)$$

Here, $p(s_i^c)$ is the probability that image i belongs to class c and is computed by applying softmax to class scores $\{s_i^c\}$,

$$p(s_i^c) = \frac{\exp(s_i^c)}{\sum_{c \in C} \exp(s_i^c)}. \quad (9)$$

However, optimizing the cross-entropy loss on training images that consist of only seen classes is prone to bias towards seen classes, as also observed in [57, 45]. In other words, given the fact that the model learns to suppress probabilities of unseen classes, during testing on images from unseen classes, the model will still predict high probabilities for seen classes, impeding the method to work well on unseen classes.

To overcome this challenge, we start by considering a calibration loss that allows to shift some of the prediction probabilities from seen to unseen classes during training. More specifically, we define

$$L_{cal} \{s_i^c\} = - \sum_i \log \sum_{u \in C_u} p(s_i^u), \quad (10)$$

where for brevity of notation, we have dropped the subscripts in $\{s_i^c\}_{c \in C}$, which can be inferred from the context. Thus, minimization of L_{cal} in conjunction with the cross-entropy loss, promotes to put nonzero probability on the unseen classes during training. Hence, at testing time, for an image from an unseen class, the model can produce a (large) non-zero probability for the true unseen class. However, the drawback of using (10), as it is, is that it reduces

the scores of seen classes and increases the scores of unseen classes during training on images from seen classes, which is not desired. Thus, to allow for nonzero prediction probability mass in unseen classes during training while keeping the scores of unseen classes low, we propose to augment the unseen scores and decrease seen scores using a margin (here, set to one), and use

$$L_{cal} \{s_i^c + 1_{C_u}(c)\}. \quad (11)$$

where $1_{C_u}(\cdot)$ is an indicator function taking the value of 1 when $c \in C_u$ and -1 otherwise. Notice that we pose the calibration process as an optimization problem. Thus, the whole model, including attention components and attribute embedding, get trained to avoid bias towards seen classes without introducing additional parameters.

Final Loss Function. Combining the cross-entropy and the self-calibration loss functions, we propose to minimize

$$\min_{W, W_e, W, \{v_a\}_{a=1}^A} L_{ce} \{s_i^c\} + L_{cal} \{s_i^c + 1_{C_u}(c)\}, \quad (12)$$

over the parameters of the two attention models (attributed-based spatial attention and attribute attention), attribute-image embedding and the attribute semantic vectors.

Remark 2 Notice that in our method, we are optimizing over attribute semantic vectors $\{v_a\}_{a=1}^A$, which results in visual grounding of each attribute meaning to the visual feature of training images. Also, by sharing $\{v_a\}_{a=1}^A$ among all classes, we effectively allow transferring fine-grained knowledge from seen to unseen classes. In the experiments, through ablation studies, we show that fine-tuning the attribute semantic vectors results in significant improvement of the performance.

Finally, at inference time, we predict the class of a test image as the class that has the maximum augmented score,

$$c = \arg\max_{c \in C} s_i^c + 1_{C_u}(c), \quad (13)$$

Thus, we make prediction based on the augmented seen and unseen scores, which we have explicitly calibrated to be sensitive toward unseen classes.

5. Experiments

We evaluate our proposed method, referred to as Dense-Attention Zero-shot LEarning (DAZLE), on three popular datasets of CUB [60], AWA2 [61] and SUN [62]. Moreover, to demonstrate the effectiveness of different components of our method, we perform experiments on DeepFashion [2], which is a dataset for fine-grained clothes recognition. Having almost 8 times more number of samples than the largest dataset among CUB, AWA2 and SUN, while having a thousand attributes, DeepFashion is a suitable dataset for studying the effectiveness of our method for fine-grained generalized zero-shot learning.

Dataset	# attributes	# seen (val) / unseen classes	# training / testing samples
CUB	312	100 (50) / 50	7,057 / 4,731
SUN	102	580 (65) / 72	10,320 / 4,020
AWA2	85	27 (13) / 10	23,527 / 13,795
DeepFashion	1,000	30 (6) / 10	204,885 / 84,337

Table 1: Statistics of the datasets used in our experiments.

Below, we discuss the datasets, evaluation metrics and baseline methods. We then present and analyze the results on all datasets. We first report the traditional zero-shot performance on the well studied fine-grained CUB dataset and then show the effectiveness of our method for generalized zero-shot learning on CUB, AWA2 and SUN. Finally, we perform ablation studies on the DeepFashion dataset.

5.1. Experimental Setup

Datasets. Following [57], we conduct experiments on the three popular datasets of CUB, SUN and AWA2 and perform ablation studies on the larger-scale DeepFashion dataset, which allows us to study the effect of different components of our method. Table 1 shows the statistics of the four datasets.

CUB [60] contains images from fine-grained bird-species with 150 seen and 50 unseen classes. Since small discriminative regions are the key for distinguishing between fine-grained classes, the dataset also contains attribute location annotation to enable learning models for part detection. Notice that our method works in the weakly supervised setting, i.e., it uses the class label of each training image and does not require annotations of attribute locations. SUN [62] is a dataset of visual scenes having 645 seen and 72 unseen classes and has the largest number of classes among the datasets. However, it only contains 16 training images per class due to its small overall training set. AWA2 [61] has been proposed for animal classification with 40 seen and 10 unseen classes and has a medium size of 37,322 samples in total. For CUB, SUN, AWA2, we follow the proposed training, validation and testing splits in [57]. Finally, DeepFashion [2] contains 289,222 samples from 46 cloth categories. We partition the categories into 36 seen and 10 unseen classes, in order to have a sufficient number of unseen classes. We use the original training/testing split of the dataset to further divide seen classes into training and testing sets.

Evaluation Metrics. Following [57], we measure the top-1 accuracy on two settings: i) traditional zero-shot learning, where test images are only from unseen classes, thus all predictions are constrained to be from unseen classes; ii) generalized zero-shot learning, where test images are from seen and unseen classes. In the latter case, we report the accuracy on testing images from seen classes, acc_s , and from unseen classes, acc_u . Also, to capture the trade-off between seen and unseen performance, we compute the harmonic mean,

H , between seen and unseen accuracy, which is

$$H = 2 \times \frac{\text{acc}_s \times \text{acc}_u}{\text{acc}_s + \text{acc}_u}. \quad (14)$$

Baselines. We group all the baselines into 3 main categories, based on the type of features used for training. EZSL [59], SYNC [18], DeViSE [19], RNet [63], DCN [44] and TCN [43] work with holistic image features without localization and relate seen to unseen classes through attribute description during inference time. On the other hand, f-CLSWGAN [25], cycle-(U)WGAN [22], f-VAEGAN-D2 [23] and CADA-VAE [24] learn generative models to approximate the distribution of class images as a function of class semantic descriptions. Thus, given semantic descriptions of unseen classes, these models augment features of seen classes with generated features from the unseen ones and learn a discriminative classifier in the fully supervised setting. Finally, we report the results of S^2GA [26], which is designed for fine-grained classification, however, requires annotations of part locations to detect discriminative parts.

Implementation Details. Following the canonical setting in [57], We use a pretrained ResNet-101 with the input size of 224×224 for feature extraction in all methods without fine-tuning. We extract a feature map at the last convolutional layer whose size is $7 \times 7 \times 2048$ and treat it as a set of features from 7×7 regions.² We extract the semantic vectors $\{v_a\}_a$ using the GloVe model [64] trained on Wikipedia articles. We implement all methods in PyTorch and optimize with the default setting of RMSprop [65] with the learning rate 0.0001 and batch size of 50. We train all models on an NVIDIA V100 GPU for at most 20 epochs on CUB, AWA2, SUN and 2 epochs on DeepFashion. *In our method, we fix $\alpha = 0.1$ on all datasets*, which also shows that our self-calibration loss works on different datasets without the need for heavy hyper-parameter tuning. We consider two variants of our method: L_{ce} (seen classes) that optimizes only cross entropy on seen class and $L_{ce} + L_{cal}$ (all classes) that optimizes cross-entropy and self-calibration losses over both seen and unseen classes.

5.2. Experimental Results

Fine-Grained Zero-Shot Learning. We measure the fine-grained zero-shot performance on the CUB dataset that contains different bird species with small visual differences, hence, demands the ability to focus on discriminative regions for classification. Following [12, 26], we report the traditional zero-shot performance given unseen attribute description.³ Due to differences in the experiment settings among previous works, we conduct experiments on both the

²This is different from the setting of [27], which uses 448×448 images.

³This is different than [12, 8] where zero-shot learning is performed on noisy text features.

Model	Approach	CUB			SUN			AWA2		
		acc _s	acc _u	H	acc _s	acc _u	H	acc _s	acc _u	H
EZSL [59]	Holistic Feature	63.8	12.6	21.0	27.9	11.0	12.1	77.8	5.9	11.0
SYNC [18]		70.9	11.5	19.8	43.3	7.9	13.4	90.5	10.0	18.0
DeViSE [19]		53.0	23.8	32.8	27.4	16.9	20.9	74.7	17.1	27.8
RNet [63]		61.1	38.1	47.0	–	–	–	93.4	30.0	45.3
DCN [44]		60.7	28.4	38.7	37.0	25.5	30.2	–	–	–
TCN [43]	Holistic Feature Generation	52.0	52.6	52.3	37.3	31.2	34.0	65.8	61.2	63.4
f-CLSWGAN [25]		57.7	43.7	49.7	36.6	42.6	39.4	68.9	52.1	59.4
cycle-(U)WGAN [22]		59.3	47.9	53.0	33.8	47.2	39.4	–	–	–
f-VAEGAN-D2 [23]		60.1	48.4	53.6	38.0	45.1	41.3	70.6	57.6	63.5
CADA-VAE [24]		53.5	51.6	52.4	35.7	47.2	40.6	75.0	55.8	63.9
DAZLE L _{ce} (seen classes)	Dense Attention	65.3	42.0	51.1	31.9	21.7	25.8	82.5	25.7	39.2
DAZLE L _{ce} + L _{cal} (all classes)		59.6	56.7	58.1	24.3	52.3	33.2	75.7	60.3	67.1

Table 2: Generalized zero-shot classification performance on CUB, SUN and AWA2. We report accuracy per seen class, acc_s, and accuracy per unseen class, acc_u, as well as their harmonic mean, H.

Method	Bounding Box Annotations	Accuracy	
		SS	PS
RNet [63]	Not Required	62.0	55.6
DCN [44]		55.6	56.2
TCN [43]		–	59.5
f-CLSWGAN [25]		–	57.3
cycle-(U)WGAN [22]		–	58.6
f-VAEGAN-D2 [23]	Required	–	61.0
S ² GA (one-attention-layer) [26]		67.1	–
S ² GA (two-attention-layer) [26]	Not Required	68.9	–
DAZLE L _{ce} (seen classes)		64.1	62.3
DAZLE L _{ce} + L _{cal} (all classes)		67.8	65.9

Table 3: Zero-shot classification performance on CUB dataset.

standard split (SS) and on the proposed split (PS) in [57] for comparison with the state-of-the-art methods. Notice that some unseen classes in SS appear in the ImageNet training set for the feature extractor, thus performances on SS are often higher than on PS.

Table 3 shows the accuracies of different methods on the two splits of the CUB. Notice that on SS, we achieve at least 5.8% improvement over methods trained on holistic image features, while we have comparable performance (within 1% difference) to methods that use ground-truth bounding box annotations of the discriminative parts during training (we do not use this information). In fact, this shows the effectiveness of our dense attribute-based attention on capturing fine-grained details, achieving similar performance to S²GA without the need for the costly annotations of the discriminative parts locations. On the other hand, on PS, we outperform other methods with at least 4.9% improvement, in particular, with respect to the state-of-the-art generative methods, which lack the ability to synthesize local discriminative regions of images. Also, notice that having the self-calibration loss facilitates knowledge transfer from seen to unseen classes, boosting the accuracy on PS by 3.6% compared to not using it.

Fine-Grained Generalized Zero-Shot Learning Table 2 shows the performance of different methods for generalized zero-shot learning, where both seen and unseen classes ap-

pear at the test time. As the results show, and expected, the unseen accuracy, acc_u, of all methods is much lower than the seen accuracy, acc_s.

Notice that compared to SYNC [18], which achieves the best seen accuracy, our method (L_{ce}) generalizes better to unseen classes with high seen accuracy. This shows the effectiveness of our dense attention mechanism in generalization to unseen classes by only focusing on transferable attribute features instead of holistic visual appearance features, which often contain irrelevant background information. However, without the self-calibration loss, our method has lower unseen accuracy especially compared to feature generation techniques, which simulate the inference distribution by augmenting training samples with synthesized features from unseen classes.

On the other hand, using the calibration loss, L_{ce} + L_{cal}, our method significantly outperforms other algorithms on unseen accuracy, in particular, improves over the state-of-the-art generative model CADA-VAE [24] on unseen accuracy by 5.1%, 5.1% and 4.5% on CUB, SUN and AWA2, respectively. In addition, our method improves the harmonic mean score by 5.7% and 3.2%, respectively, on CUB and AWA2. However, it does not achieve the best harmonic mean on SUN. We believe this is due to having only 16 training samples for all seen classes, which does not allow to effectively train our dense attention model and results in even low seen performance compared to SYNC [18]. Please see the supplementary materials for more detailed analysis of different components of our method.

Ablation Study. We evaluate the effectiveness of different components of our method by performing fine-grained generalized zero-shot classification for clothes recognition on the DeepFashion dataset, with 1,000 attributes.

As the results in Table 4 show, without self-calibration, while different variants of our method do well for classifying seen classes, they do not generalize to unseen classes. Using our dense attention instead of no attention improves the seen accuracy by 1.4% (without self-

Dense Attention	Self Calibration	Attention on Attribute	acc _s	acc _u	H
No	No	No	45.3	4.8	8.7
Yes	L _{ce} (seen classes)	No	46.7	6.1	10.8
Yes		Yes	38.7	8.2	13.5
No		No	36.2	18.9	24.8
Yes	Yes L _{ce} + L _{cal} (all classes)	No	37.1	20.3	26.2
Yes (fixed v _a)		Yes	31.6	21.5	25.6
Yes		Yes	38.1	21.5	27.5

Table 4: Ablation study for generalized zero-shot learning on the DeepFashion dataset.

calibration) and improves the harmonic mean by 1.4% (with self-calibration), which demonstrates the importance of attending to fine-grained attributes. When using self-calibration, attention on attributes further boosts the harmonic mean by 1.3%. Notice that without refining the attribute semantic vectors, i.e., when v_a 's are fixed, the harmonic mean drops by 1.9% (compared to when refining them), showing that semantic representations learned from GloVe is not initially compatible with visual features and learning of the attribute semantics is necessary.

Effect of Hyperparameters and Attribute Selection.

Given that our framework produces one attention per attribute, we investigate the effect of the number of used attributes, hence the number of attentions, by learning from different subsets of attributes on the DeepFashion dataset. To do so, we rank attributes by their discriminative power measured via the entropy on the probability of each attribute appearing in all classes. If an attribute appears in all classes, then it is non-discriminative and will have high entropy and if an attribute is only present in one class, then it will have zero-entropy, indicating its discriminative power.

As the results in Figure 4 (left) shows by learning an attention for each of the top 300 discriminative attributes, our method achieves high harmonic mean accuracy, which shows the importance of attribute selection. Notice that by dynamically weighting the importance of each attribute through attention over attributes, our method further improves the performance by 1.3%. As the results show, when we only use attention on attributes (without attribute-based features, by using mean of features from all image regions) the performance drops by more than 3% compared to using dense attention and attention over attributes. This shows the importance of our dense attention mechanism, which provides inputs for the attribute attention module.

Figure 4 (right) shows the performance of our method as a function of α on DeepFashion. Notice that for very small values of α , we obtain high seen accuracy and nearly zero unseen accuracy and as α increases, the seen accuracy decreases and unseen accuracy increases and finally saturates. As a result, the Harmonic mean, which captures a trade-off between seen and unseen accuracies, achieves the optimal score when seen and unseen accuracies are similar, which corresponds to when $\alpha \in [0.1, 0.3]$.

Qualitative Results. Figure 5 visualizes the dense-

Figure 4: Left: Effect of the number of used attributes for learning dense attention and effect of attention over attributes on Harmonic mean. Right: Effect of α on seen, unseen and the harmonic mean accuracy. Both experiments are performed on DeepFashion.

Figure 5: Visualization of attention maps of attributes with positive attribute scores (left) and with negative attribute scores (right).

attention maps on the CUB dataset. Notice that our model is able to localize fine-grained information given weak supervision, i.e., only image labels. Moreover, our model correctly assigns positive scores to present attributes and negative scores to absent attributes by focusing on regions that support or reject the existence of each attribute. This demonstrates the capability of learning different levels of abstraction/granularity through the hierarchical structure of W and W_e , where the input of (4) depends on the output of (3). We observe that W well localizes different parts of a bird for W_e to determine the presence (e.g., 'belly color black' in first image) or absence (e.g., 'belly color white' in first image) of different patterns and colors of these parts.

6. Conclusion

We proposed a dense attribute-based attention mechanism with attention over attributes that focuses on the most relevant image regions of each attribute and grounds visual attribute descriptions to discriminative regions in a weakly supervised setting. To transfer knowledge from seen to unseen classes, we proposed a self-calibration loss that adjusts the prediction distribution in advance to better adapt to unseen classes at the inference time. By extensive experiments on three well-studied datasets and the DeepFashion dataset, we showed the effectiveness of our proposed method.

Acknowledgements

This work is partially supported by DARPA Young Faculty Award (D18AP00050), NSF (IIS-1657197), ONR (N000141812132) and ARO (W911NF1810300).

References

- [1] W. Wang, Y. Xu, J. Shen, and S. C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. **1**
- [2] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. **1, 5, 6**
- [3] K. E. Ak, A. A. Kassim, J.-H. Lim, and J. Y. Tham, "Learning attribute representations with localization for flexible fashion search," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. **1**
- [4] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *British Machine Vision Conference*, 2015. **1**
- [5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," *European Conference on Computer Vision*, 2016. **1**
- [6] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," *IEEE International Conference on Computer Vision*, 2015. **1**
- [7] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective sparse sampling for fine-grained image recognition," *IEEE International Conference on Computer Vision*, 2019. **1, 2**
- [8] M. Elhoseiny, Y. Zhu, H. Zhang, and A. M. Elgammal, "Link the head to the "beak": Zero shot learning from noisy text description at part precision," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6288–6297, 2017. **1, 6**
- [9] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," *IEEE International Conference on Computer Vision*, 2015. **1, 2**
- [10] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," *IEEE International Conference on Computer Vision*, 2017. **1, 2**
- [11] X. Zhao, Y. Yang, F. Zhou, X. Tan, Y. Yuan, Y. Bao, and Y. Wu, "Recognizing part attributes with insufficient data," *IEEE International Conference on Computer Vision*, 2019. **1**
- [12] Z. Zhang and V. Saligrama, "Zero-shot learning via joint latent similarity embedding," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. **1, 2, 4, 6**
- [13] D. Huynh and E. Elhamifar, "Interactive multi-label CNN learning with partial labels," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. **1**
- [14] D. Wertheimer and B. Hariharan, "Few-shot learning with localization in realistic settings," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. **1**
- [15] S. Kong and C. C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. **1, 2**
- [16] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. **1, 2**
- [17] D. Huynh and E. Elhamifar, "A shared multi-attention framework for multi-label zero-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. **1**
- [18] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. **1, 2, 4, 6, 7**
- [19] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Neural Information Processing Systems*, 2013. **1, 2, 4, 6, 7**
- [20] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," *International Conference on Learning Representations*, 2014. **1, 2, 3, 4**
- [21] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. **2, 4**
- [22] R. Felix, B. G. V. Kumar, I. D. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," *European Conference on Computer Vision*, 2018. **2, 4, 6, 7**
- [23] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "f-vaegand2: A feature generating framework for any-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. **2, 4, 6, 7**
- [24] E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero- and few-shot learning via aligned variational autoencoders," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. **2, 4, 6, 7**
- [25] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5542–5551, 2018. **2, 4, 6, 7**
- [26] Y. Yu, Z. Ji, Y. Fu, J. Guo, Y. Pang, and Z. Zhang, "Stacked semantics-guided attention model for fine-grained zero-shot learning," *Neural Information Processing Systems*, 2018. **2, 3, 6, 7**
- [27] Y. Zhu, J. Xie, Z. Tang, X. Peng, and A. Elgammal, "Learning where to look: Semantic-guided multi-attention localization for zero-shot learning," *Neural Information Processing Systems*, 2019. **2, 3, 6**
- [28] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele, "Multi-cue zero-shot learning with strong supervision," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. **2**
- [29] G. S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, and L. M. Shao, "Attentive region embedding network for zero-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. **2**
- [30] Y. Guo, G. Ding, J. Han, and S. Tang, "Zero-shot learning with attribute selection," *AAAI Conference on Artificial Intelligence*, 2018. **2**

- [31] Y. Liu, J. Guo, D. Cai, and X. He, "Attribute attention for semantic disambiguation in zero-shot learning," *IEEE International Conference on Computer Vision*, 2019. **2**
- [32] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," *European Conference on Computer Vision*, 2018. **2**
- [33] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a cnn for fine-grained recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. **2**
- [34] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," *European Conference on Computer Vision*, September 2018. **2**
- [35] L. Zhang, S. Huang, W. Liu, and D. Tao, "Learning a mixture of granularity-specific experts for fine-grained categorization," *IEEE International Conference on Computer Vision*, 2019. **2**
- [36] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep lac: Deep localization, alignment and classification for fine-grained recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. **2**
- [37] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," *European Conference on Computer Vision*, 2018. **2**
- [38] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. **2**
- [39] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," *IEEE International Conference on Computer Vision*, 2019. **3**
- [40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Neural Information Processing Systems*, 2017. **3**
- [41] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. **3**
- [42] E. Kodirov, T. Xiang, Z.-Y. Fu, and S. Gong, "Unsupervised domain adaptation for zero-shot learning," *IEEE International Conference on Computer Vision*, 2015. **3**
- [43] H. Jiang, R. Wang, S. Shan, and X. Chen, "Transferable contrastive network for generalized zero-shot learning," *IEEE International Conference on Computer Vision*, 2019. **3, 6, 7**
- [44] S. Liu, M. Long, J. Wang, and M. I. Jordan, "Generalized zero-shot learning with deep calibration network," *Neural Information Processing Systems*, 2018. **3, 6, 7**
- [45] W. L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," *European Conference on Computer Vision*. **3, 5**
- [46] Y. Atzmon and G. Chechik, "Adaptive confidence smoothing for generalized zero-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. **3**
- [47] R. Socher, M. Ganjoo, C. D. Manning, and A. Y. Ng, "Zero-shot learning through cross-modal transfer," *Neural Information Processing Systems*, 2013. **3**
- [48] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Neural Information Processing Systems*, 2014. **3**
- [49] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *International Conference on Learning Representations*, vol. abs/1412.7755, 2015. **3**
- [50] Z. Wang, T. Chen, G. Li, G. Li, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," *IEEE International Conference on Computer Vision*, 2017. **3**
- [51] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3668–3677, 2016. **3**
- [52] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2015. **3**
- [53] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018. **3**
- [54] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *Neural Information Processing Systems*, 2015. **3**
- [55] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. **3**
- [56] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. **3**
- [57] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning — the good, the bad and the ugly," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. **3, 5, 6, 7**
- [58] M. Bucher, S. Herbin, and F. Jurie, "Generating visual representations for zero-shot classification," *IEEE International Conference on Computer Vision Workshops*, 2017. **3**
- [59] B. Romera-Paredes and P. H. Torr, "An embarrassingly simple approach to zero-shot learning," *International Conference on Machine Learning*, 2015. **3, 6, 7**
- [60] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," California Institute of Technology, Tech. Rep. CNS-TR-2010-001, 2010. **5, 6**
- [61] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. **5, 6**

- [62] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 5, 6
- [63] F. Sung, Y. Yang, N. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6, 7
- [64] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," *Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 6
- [65] T. Tijmen and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSEERA: Neural networks for machine learning 4.2*, 2012. 6