Attribute Attention for Semantic Disambiguation in Zero-Shot Learning

Yang Liu, Jishun Guo†, Deng Cai*, Xiaofei He
State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou, China
Fabu Inc., Hangzhou, China
Alibaba-Zhejiang University Joint Institute of Frontier Technologies
†GAC R&D Center, Guangzhou, China
lyng_95@zju.edu.cn, guojishun@gacrnd.com, dengcai@cad.zju.edu.cn, xiaofeihe@fabu.ai

Abstract

Zero-shot learning (ZSL) aims to accurately recognize unseen objects by learning mapping matrices that bridge the gap between visual information and semantic attributes. Previous works implicitly treat attributes equally in compatibility score while ignoring that they have different importance for discrimination, which leads to severe semantic ambiguity. Considering both low-level visual information and global class-level features that relate to this ambiguity, we propose a practical Latent Feature Guided Attribute Attention (LFGAA) framework to perform object-based attribute attention for semantic disambiguation. By distracting semantic activation in dimensions that cause ambiguity, our method outperforms existing state-of-the-art methods on AwA2, CUB and SUN datasets in both inductive and transductive settings. The source code is released at https://github.com/ZJULearning/AttentionZSL.

1. Introduction

Zero-shot learning (ZSL), whose goal is to construct a classification model for classes that have no labeled samples before, is an active research topic recently [1, 18, 46, 33, 27, 39, 24, 43, 4]. Unlike supervised classification that directly assigns an unlabeled object to one of training accessible (*seen*) categories, ZSL aims to recognize objects that are *unseen* in training. To achieve this goal, auxiliary semantic attributes are provided for both seen and unseen classes [46, 44, 30]. ZSL then learns to predict in semantic space for the unseen object and infer its label by searching the class that attains the most similar semantic attribute.

Existing ZSL methods can be divided into *inductive* ZSL and *transductive* ZSL depending on whether images from unseen classes are available during training. Neither visual information nor side-information of unseen classes is available for inductive ZSL [49, 35, 17, 24, 16] while transductive ZSL [20, 13, 42, 47, 10, 39] has access to part of the unlabeled images. During the test phase, both *conventional* setting and *generalized* setting are considered in most recent works [42, 16, 39]. The search space for classifying new images is restricted to unseen classes in the conventional setting. While in generalized ZSL setting, we assume the test images come from both seen and unseen classes.

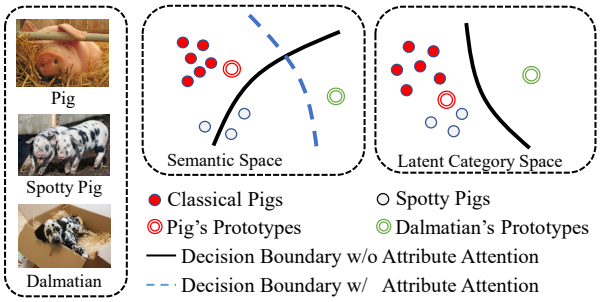


Figure 1: An illustrative diagram of semantic ambiguity where *spotty pig* is likely to be misclassified as *dalmatian* in semantic space. Based on objects' visual features and latent category features, we propose an attention based model to distract high activation in objects' sambiguous dimensions.

Compatibility score based measurement, which exploit linear or nonlinear function $F(x, y; W) = \theta(x)^T W \phi(y)$ to associate visual representation $\theta(x)$ and provided auxiliary side-information $\phi(y)$, have dominated the ZSL literature in the past few years [9, 1, 2, 35, 45]. Compatibility score could be adopted not only in user-defined semantic space ($\phi(y) = \boldsymbol{a}_y$) but also in latent feature space ($\phi(y) = \boldsymbol{\sigma}_y$) that recently introduced in several state-of-the-art ZSL methods [17, 24, 48, 31].

While previous works mostly focus on introducing various regularization targets [9, 2, 35, 1] to learn better mapping matrices W, semantic attribute itself is less concerned in the literature. Attributes are *implicitly* treated equally in

*Corresponding author

compatibility score for almost all existing works, while in this work, we argue that this kind of equal treatment leads to severe ambiguity in semantic attribute space. We use word *semantic-ambiguous* to describe those atypical objects that carry other classes' common attributes.

The misclassification of semantic-ambiguous objects can be described as follows: attribute P is typical in class A but is rarely found in class B; a semantic-ambiguous instance from B that carries attribute P will be categorized to class A since P is relatively common in class A. For example, *spotty pig*, as shown in Figure 1, is more likely to be classified to *dalmatian* since attribute *spots* is more typical in dalmatian than in pig (100.0 for dalmatian and 21.2 for pig). Since attribute with great value is generally thought to be common for instances within the class, the decision boundary of that attribute dimension is therefore closer to the class that owns greater attribute value. This kind of property makes classifications work well in most general cases, while at the same time, makes semantic-ambiguous objects that own atypical attributes be easily misclassified. Other examples include *ocean* for *polar bear* (35.0) against *humpback whale* (89.4) and *stripes* for *squirrel* (12.50) against *zebra* (98.9).

To alleviate the problem above, we propose third object-based attribute attention p apart from semantic attributes and latent category features to distract atypical semantic activations. Specifically, the attribute attention p_j at dimension j should be small if its corresponding attribute is not general in its potential class. By analyzing the factors that relate to the proposed attribute attention, we take both global class-level features and low-level visual information into considerations. In summary, the contributions are:

(1) We point out that semantic ambiguity exists in current ZSL methods and design an end-to-end attribute attention framework for semantic disambiguation.

(2) We propose an offline prototype learning strategy independent from visual-semantic training that shows effectiveness in both conventional and generalized ZSL settings.

(3) Combined with different prototype learning strategies, our method achieves the state-of-the-art performance in both inductive and transductive settings.

2. Related Work

Zero-Shot Learning Traditional ZSL work [22, 3, 15, 29] follows a two-stage inference. The visual representations of unlabeled objects are firstly projected to semantic space, and classification is then performed via searching the class that attains the most similar attributes [11, 10, 34, 48]. Various semantic spaces have been investigated including user-defined attribute annotations [8] and unsupervised semantic representations (word2vec [26], GloVe [32]).

Although most works pay attention to learning within single semantic space, other feature spaces are also investigated recently. CDL [16] jointly aligns the class structures in both visual and semantic space. LAD [17] exploits dictionary learning to obtain a discriminative but semantic-preserving latent feature space. JSLA [31] learns latent representations by minimizing the intra-class distance. LDF [24] considers both intra-class and inter-class distances in latent feature space. In this work, we further connect those two spaces by (1) latent category features give guidance in semantic attribute attention; (2) semantic features provide hints in latent prototype construction.

Prototype Learning Prototype [10] is the most representative class-level embeddings used in classification. Existing ZSL methods learn prototypes for different purposes. Among them, DMaP [23] uses an iterative method to revise more semantically consistent prototypes within single semantic space. LDF [24] and JSLA [31] exploit ridge regression in learning class relatedness to obtain unknown prototypes in latent feature space. CDL [16] learns unseen class prototypes by sharing the structures between the visual and semantic space. Since our method also involves learning in multiple feature spaces, we also need latent prototypes in classification. We first adopt the same prototype learning framework as in [31, 24] to show our effectiveness in inductive ZSL setting and then propose another offline learning strategy to mitigate domain shift [10] in transductive ZSL setting. The proposed offline learning strategy differs from [23] in that (1) We separate prototype learning from visual-semantic training and make LFGAA train in an end-to-end manner. (2) We align prototypes across feature spaces rather than optimizing within single semantic space.

Attribute Selection As mentioned in NAS [12], attributes own different properties (*e.g*. class distribution, variance, and entropy) and have different importance in discrimination. NAS proposes to use a refined subset of attributes to build specific ZSL models. However, their refined attribute subset varies from models and datasets, which is a kind of dataset-based and model-based attribute selection. Our proposed attribute attention can also be regarded as a soft attribute selection. Different from NAS, our proposed attribute attention varies from objects even within the same class. To the best our knowledge, this is the first work to consider object-based attribute attention in ZSL.

3. Pre-analysis

3.1. Problem Formulation and Notations

We formulate ZSL problem as follows: a seen dataset $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N^s}$ that consists of N^s images is used as the training set, where x_i^s is the i-th image and $y_i^s \in \mathcal{Y}^{\mathcal{S}}$ is its corresponding label; a similar unseen dataset $\mathcal{U} =$

$\{(x_i^u, y_i^u)\}_{i=1}^{N^u}$ is used as the testing set. The seen and unseen classes are disjoint, *i.e.*, $\mathcal{Y}^{\mathcal{S}} \cap \mathcal{Y}^{\mathcal{U}} = \emptyset$, $\mathcal{Y}^{\mathcal{S}} \cup \mathcal{Y}^{\mathcal{U}} = \mathcal{Y}$. For each $y \in \mathcal{Y}$, there is an attribute vector $\boldsymbol{a}_y \in \mathbb{R}^k$ associated to it. We denote $\varphi(x) = \theta(x)^T W$ and $\sigma(x)$ to represent semantic prediction and latent feature prediction respectively in this work.

3.2. Importance of Object-based Attention

Attribute attention selectively concentrates on a discrete set of attributes and ignores less important ones. We demonstrate the importance of object-based attribute attention for dealing with the aforementioned semantic-ambiguous objects in this section.

We first view the nearest class search as multiple binary classifications among different combinations of categories, *e.g.*, deciding if object x closer to class y_1 or y_2 by:

$$\mathcal{D}(x, y_1, y_2; W) = F(x, y_1; W) - F(x, y_2; W) \tag{1}$$

where the sign of $\mathcal{D}$ indicates the binary classification result and $\mathcal{D} = 0$ is its decision hyperplane.

Consider the most straightforward form of compatibility score between semantic prediction and class-level attribute:

$$\begin{aligned} \mathcal{D}_{\text{ip}}(x, y_1, y_2) &= \varphi(x)^T \Delta(\boldsymbol{a}, y_1, y_2) \\ \Delta(\boldsymbol{a}, y_1, y_2) &= \boldsymbol{a}_{y_1} - \boldsymbol{a}_{y_2} \end{aligned} \tag{2}$$

where inner product is directly applied on them. It can be seen that $\mathcal{D}_{\text{ip}}$ is decided by both semantic prediction $\varphi(x)$ and attribute difference $\Delta(\boldsymbol{a}, y_1, y_2)$. We use l^1-normalized $\Delta(\boldsymbol{a}, y_1, y_2)$ as *information amount* to denote the discriminative information that each attribute dimension carries in binary classification, *i.e.*, attribute i is more discriminative than attribute j in classifying between class y_1 and y_2 if $|\Delta(\boldsymbol{a}, y_1, y_2)_i|$ is greater than $|\Delta(\boldsymbol{a}, y_1, y_2)_j|$. As shown in Figure 2, classification greatly depends on a small subset of attributes that carry greater discriminative information.

Another widely used similarity measurement [4, 16, 10, 17, 31] is cosine distance:

$$\begin{aligned} \mathcal{D}_{\cos}(x, y_1, y_2) &= \frac{1}{\|\varphi(x)\| \|\boldsymbol{a}_{y_1}\|} \varphi(x)^T \Delta'(\boldsymbol{a}, y_1, y_2) \\ \Delta'(\boldsymbol{a}, y_1, y_2) &= \boldsymbol{a}_{y_1} - \frac{\|\boldsymbol{a}_{y_1}\|}{\|\boldsymbol{a}_{y_2}\|} \boldsymbol{a}_{y_2} \end{aligned} \tag{3}$$

Compared with the simple inner product, cosine distance takes additional l^2-norm into consideration where classes with great attribute norm are unfavorable in discrimination.

Both inner product and cosine distance work well in general cases; however, it fails to deal with those semantic-ambiguous cases. Those objects usually have great activations at their less typical attribute dimensions that leads to ambiguity. Worse still, there also exists correlation across attribute dimensions (*e.g.*, *spots* is highly related to *black* and *white*) that further aggregates such ambiguity. Based on this observation, we argue that the gap of semantic predictions among different dimensions should be reduced so that one or a few prominent attribute predictions would not dominate in classification. We then propose attribute attention to distract high activations on atypical attribute dimensions.

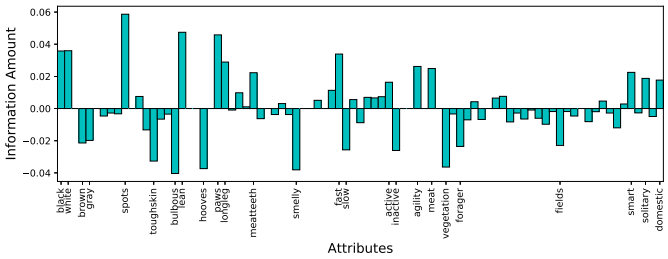


Figure 2: Attribute information amount in classifying between *pig* and *dalmatian*. Attributes with positive information amount are favourable to *dalmatian*.

Compatibility score with proposed object-based attribute attention can be written as:

$$F'(x, y; W) = \theta(x)^T W \text{diag}(p(x)) \boldsymbol{a}_y \tag{4}$$

where $p(x)$ is the proposed attention and W is the parameter for visual-semantic projection. The proposed compatibility score differs from the traditional ones [22, 3, 15, 29] in that: (1) Our attention $p(x)$ is object-based (a function of x) where low-level visual information is also exploited, while visual-semantic mapping matrices in traditional ZSL are learned directly from deep visual embedding $\theta(x)$. (2) Learning in semantic space alone makes projection matrices highly related to semantic attributes while aforementioned challenging cases are ambiguous in that space; in contrast, our proposed attention is learned independently from semantic space that won't enhance such ambiguity.

4. Latent Feature Guided Attribute Attention

Based on the description of semantic ambiguity problem, we adopt the idea that attribute attention should be highly related to both global category features and low-level visual information. The proposed Latent Feature Guided Attribute Attention (**LFGAA**) network is shown in Figure 3. At the core of our network, an Embedding Subnet learns projections from visual space to semantic space and latent feature space at the same time. The Embedding Subnet is decomposed into several branches according to their receptive fields. The Latent Guided Attention (**LGA**) module is attached within each branch to fuse visual information and global category features. Attribute attention from different visual levels aggregates at the end of the network.

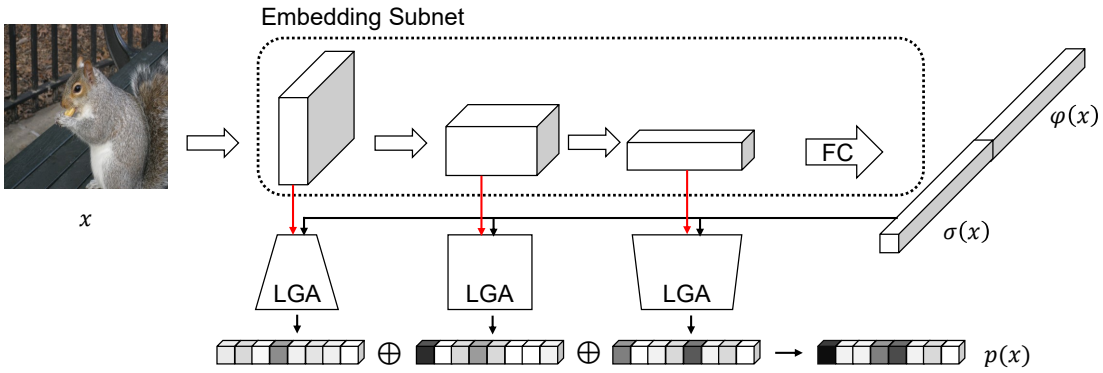


Figure 3: Overview of the proposed Latent Feature Guided Attribute Attention (LFGAA) network. Given an image, we first use Embedding Subnet to extract visual information. We build a fully connected layer on top to project visual embeddings into both user-defined semantic space and latent feature space. At the middle layers of Embedding Subnet, several Latent Guided Attention (LGA) modules are branched to perform object-based attribute attention. For each image input, LFGAA simultaneously produces semantic prediction $\varphi(x)$, latent feature prediction $\sigma(x)$ and semantic attribute attention $p(x)$. Notations: FC fully connected layer, $\oplus$ element-wise summation.

4.1. Attribute Embedding Subnet

Different from existing ZSL methods [17, 7, 28] that directly use pre-trained deep CNN features as their visual representations, we jointly optimize backbone CNN as well as other parts of Embedding Subnet in our work. Image features extracted from backbone CNN are fed to several fully connected layers with ReLU activation to be non-linearly projected to semantic and latent feature space respectively. Latent features are then used in both giving global class-related guidances in LGA module and making predictions in latent space, while semantic features are compared with attribute annotations via attention for semantic predictions.

4.2. Latent Guided Attention Module

Despite learning in both semantic space and latent feature space achieves promising performance for most general cases in the literature [17, 31, 24], object-based attribute attention should be incorporated to deal with semantic-ambiguous objects. Latent Guided Attention module, as detailed in Figure 4, stems from the intuition that attribute attention is related to global class-level features as well as information from different visual levels. Given an visual feature map $\mathcal{M}_{i,l} \in \mathbb{R}^{C\times H\times W}$ of i-th image at specific layer l of Embedding Subnet and its corresponding latent feature embedding $\sigma(x_i) \in \mathbb{R}^k$, the proposed attribute attention $p_{i,l} \in \mathbb{R}^k$ from layer l is obtained as follows:

Visual feature map $\mathcal{M}_{i,l}$ is firstly mapped through a set of standard convolutional layers $\mathcal{F}$ to obtain $\mathcal{M}'_{i,l} \in \mathbb{R}^{k\times H'\times W'}$ that shares the same channel dimensions as latent features:

$$\mathcal{M}'_{i,l} = \mathcal{F}(\mathcal{M}_{i,l}) \tag{5}$$

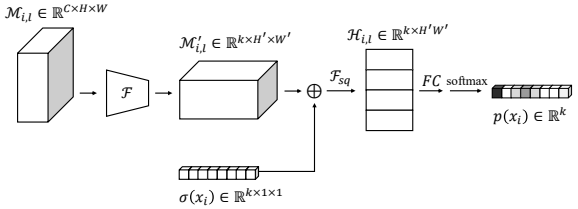


Figure 4: An illustration of Latent Guided Attention (LGA).

We use the same H' and W' for all LGA modules at different branch layers $l \in l_{\mathcal{B}}$ in this work.

We then combine the projected feature map $\mathcal{M}'_{i,l}$ and the latent feature $\sigma(x_i)$ to obtain attribute attention as follows:

$$\begin{aligned} \mathcal{H}_{i,l} &= \mathcal{F}_{sq}(\mathcal{M}'_{i,l} \oplus \sigma(x_i)) \\ p_l(x_i) &= \text{softmax}(W_l\mathcal{H}_{i,l} + b_l) \end{aligned} \tag{6}$$

where $\mathcal{V} = \mathcal{F}_{sq}(\mathcal{M})$ is a squeeze function that turns feature map $\mathcal{M} \in \mathbb{R}^{C\times H\times W}$ into feature vector $\mathcal{V} \in \mathbb{R}^{C\times HW}$ and $\oplus$ is channel-wise addition. W_l and b_l are parameters of single fully connected layer at specific branch layer l that assembles regional visual information.

Attribute attention obtained from specific layer l corresponds to a certain depth of visual perception. As information from different visual levels contributes to different semantic meanings, we fuse multiple attribute attention from different layers to obtain integrated attention.

4.3. Optimization

We consider both visual-semantic and visual-latent projection in our work and optimize them at the same time.

For visual-latent projection, we follow the same way as in LDF [24] that uses triplet loss [36] to learn discriminative latent category features by simultaneously enlarging inter-class distance and reducing intra-class distance:

$$\mathcal{L}_F = \frac{1}{N}\sum_{i}^{N}[\|\sigma(x_i)-\sigma(x_j)\|^2 - \|\sigma(x_i)-\sigma(x_k)\|^2 + \alpha]_+ \tag{7}$$

where x_i, x_j and x_k serve as anchor, positive and negative sample within a triplet respectively. Symbol $[\circ]_+$ is equivalent to $\max(\circ, 0)$. α is a preset parameter to control the desired margin between anchor-positive pair and anchor-negative pair.

For visual-semantic projection, we jointly learn parameters in Embedding Subnet as well as in LGA modules via widely used softmax loss:

$$\mathcal{L}_A = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(\varphi(x_i)^T \text{diag}(p(x_i))\boldsymbol{a}_{y_i})}{\sum\limits_{y\in\mathcal{Y}^S}\exp(\varphi(x_i)^T \text{diag}(p(x_i))\boldsymbol{a}_y)} \tag{8}$$

We combine these two optimization targets with a balancing factor β as our final optimization target, which can be written as:

$$\mathcal{L} = \mathcal{L}_F + \beta\mathcal{L}_A \tag{9}$$

4.4. Prototype Construction and ZSL Prediction

We investigate both inductive and transductive prototype construction in our work to prove the effectiveness of proposed attention. For inductive setting, we adopt the same construction process as in JSLA [31] and LDF [24], where mean latent features are directly used as prototypes for seen classes $\boldsymbol{\sigma}_{y^s} = \frac{1}{N}\sum_{x\in x_i^s}\sigma(x)$ and unseen prototypes are obtained in a *hybrid* way with ridge regression:

$$\begin{aligned}\beta_y^u &= \underset{y\in\mathcal{Y}^S}{\arg\min}\ \|\boldsymbol{a}_{y^u} - \sum\beta_y^u\boldsymbol{a}_y\|_2^2 + \lambda\|\beta_y^u\|_2^2 \\ \boldsymbol{\sigma}_{y^u} &= \sum_{y\in\mathcal{Y}^S}\beta_y^u\boldsymbol{\sigma}_y\end{aligned} \tag{10}$$

Then inductive hybrid ZSL prediction is performed by:

$$y_i^c = \underset{y^c\in\mathcal{Y}}{\arg\max}\ \boldsymbol{s}(\varphi_i', \boldsymbol{a}_{y^c}) + \boldsymbol{s}(\sigma(x_i), \boldsymbol{\sigma}_{y^c}) \tag{11}$$

It can be observed from Eq.(10) that *domain shift* [10] problem exists in hybrid prototypes where β_y^u is computed in semantic space and it cannot exactly reflects the true class relationship in latent space. Directly adopting this class-relatedness learned in semantic space without any adaptation to latent space causes an unknown shift.

Inspired by NCM classifier [25] that assigns images to the class with the closest mean, we propose an offline **Self-Adaptation** (**SA**) strategy to build prototypes directly in latent feature space in transductive ZSL setting. The basic ideas for self-adaptation are that (1) samples should locate near their corresponding prototypes; (2) samples close in semantic space tend to be close in latent space. We denote attended semantic predictions $\varphi'(x_i) = \varphi(x_i)^T\text{diag}(p(x_i))$ and use cosine-similarity as the basic measurement of such closeness for simplicity in this section:

$$\boldsymbol{s} = \frac{\varphi^T\boldsymbol{a}_y}{\|\varphi\|\cdot\|\boldsymbol{a}_y\|} \tag{12}$$

We first introduce a single SA step as follows: pseudo labels y^c are first assigned to unlabeled objects based on their attended semantic predictions:

$$y_i^c = \underset{y^c\in\mathcal{Y}}{\arg\max}\ \boldsymbol{s}(\varphi'(x_i), \boldsymbol{a}_{y^c}) \tag{13}$$

The latent feature prototype of unseen class $u \in \mathcal{U}$ can then be obtained by averaging latent feature predictions of instances within pseudo-labeled class u:

$$\boldsymbol{\sigma}_{y^u} = \frac{\sum_i\sigma(x_i)\mathbb{1}(y^u, y_i^c)}{\sum_i\mathbb{1}(y^u, y_i^c)} \tag{14}$$

$\mathbb{1}(x, y)$ returns 1 if $x = y$ or 0 otherwise. Prototypes obtained in Eq.(14) alleviate domain shift problem by directly averaging in latent feature space, and semantic space only provides a hint in this process.

Self-Adaptation then, as described below, iteratively revises the semantic prototypes and aligns the latent prototypes with initialization $\boldsymbol{\sigma}_{y^u}^0 = \boldsymbol{\sigma}_{y^u}$. We use the latest pseudo labels as our transductive prediction labels $y_i^u = y_{i,T}^c$ in this work.

Self-Adaptation: simultaneously build latent feature prototypes, revise semantic prototypes and make predictions.

Initialize:

$\boldsymbol{\sigma}_{y^u}^0 \leftarrow \boldsymbol{\sigma}_{y^u}$, $\boldsymbol{a}_{y^u}^0 \leftarrow \boldsymbol{a}_{y^u}$, $y_{i,0}^c \leftarrow y_i^c$

for t = 1 to T **do**

$y_{i,t}^c \leftarrow \underset{y^u}{\text{argmax}}\ \boldsymbol{s}(\varphi'(x_i), \boldsymbol{a}_{y^u}^{t-1}) + \boldsymbol{s}(\sigma(x_i), \boldsymbol{\sigma}_{y^u}^{t-1})$

$\boldsymbol{\sigma}_{y^u}^t \leftarrow \dfrac{\sum_i\sigma(x_i)\mathbb{1}(y^u, y_{i,t}^c)}{\sum_i\mathbb{1}(y^u, y_{i,t}^c)}$

$\boldsymbol{a}_{y^u}^t \leftarrow \dfrac{\sum_i\varphi'(x_i)\mathbb{1}(y^u, y_{i,t}^c)}{\sum_i\mathbb{1}(y^u, y_{i,t}^c)}$

---

5. Experiments

5.1. Setting

Datasets Experiments are conducted on three representative ZSL datasets: Animals with Attribute 2 [46] (AwA2), Caltech-UCSD Birds 200-2011 [44] (CUB) and SUB Attribute Database [30] (SUN). AwA2 is a coarse-grained

Table 1: Comparisons in the conventional ZSL setting (%). For each dataset, the best performance is marked in **bold** font for both inductive and transductive methods. For LFGAA$^{\text{G}}$, LFGAA$^{\text{V}}$ and LFGAA$^{\text{R}}$, the visual embedding function is implemented with GoogleNet[40], VGG19[37], and ResNet101[14] respectively. Both standard split (SS) and proposed split (PS) are considered. Notations: $\mathcal{I}$ inductive ZSL methods, $\mathcal{T}$ transductive ZSL methods.

	Method	AwA2		CUB		SUN	
		SS	PS	SS	PS	SS	PS
$\mathcal{I}$	DAP[22]	58.7	46.1	37.5	40.0	38.9	39.9
	SSE[49]	67.5	61.0	43.7	43.9	25.4	54.5
	CONSE[29]	67.9	44.5	36.7	34.3	44.2	38.8
	DEVISE[9]	68.6	59.7	53.2	52.0	57.5	56.5
	ESZSL[35]	75.6	55.1	43.7	53.9	57.3	54.5
	ALE[1]	80.3	62.5	53.2	54.9	59.1	58.1
	SJE[2]	69.5	61.9	55.3	53.9	57.1	53.7
	SYNC[5]	71.2	46.6	54.1	55.6	59.1	56.3
	LAD[17]	78.4	67.8	56.6	57.9	51.7	62.6
	CDL[16]	79.5	67.9	54.5	54.5	61.3	**63.6**
	Y. Annadani *et al*. [4]	-	63.8	-	56.0	-	61.4
	LFGAA+Hybrid (Ours)	**84.3**	**68.1**	**67.6**	**67.6**	**62.0**	61.5
$\mathcal{T}$	TAAw[21]	82.0	-	51.0	-	57.0	-
	SE-ZSL[42]	80.8	69.2	60.3	59.6	64.5	63.4
	QFSL[39]	84.8	79.7	69.7	72.1	61.7	58.3
	LFGAA$^{\text{V}}$+SA (Ours)	94.0	75.5	**80.0**	76.9	**66.7**	61.4
	LFGAA$^{\text{G}}$+SA (Ours)	**95.1**	76.6	78.1	**81.1**	63.1	64.8
	LFGAA$^{\text{R}}$+SA (Ours)	94.4	**84.8**	79.7	78.9	64.0	**66.2**

dataset that is medium-scale regarding the number of images, *i.e*. 37,322 images from 50 animal classes with 85 user-defined attributes. CUB is a fine-grained dataset consisting of 11,788 images from 200 different bird species with 312 user-defined attributes. SUN is another fine-grained dataset including 14,340 images from 717 different scenes provided 102 user-defined attributes. Standard 40/10, 150/50, 645/72 zero-shot splits are adopted on AwA2, CUB and SUN respectively for both standard split and proposed split [46].

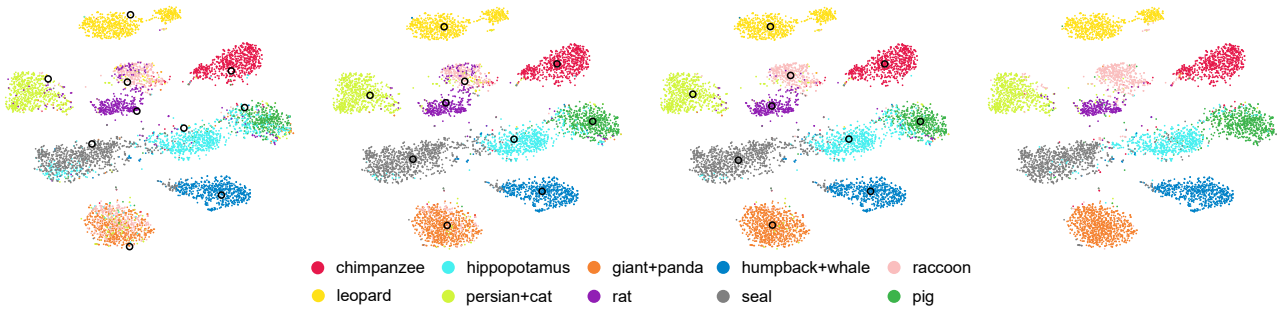
Evaluation Metrics We use average per-class top-1 accuracy ($Acc_{\mathcal{Y}}$) as the primary metric in our experiments and conduct the experiments in both conventional and generalized setting [46]. Unlabeled objects in conventional setting only come from unseen classes ($\mathcal{Y} = \mathcal{Y}^{\mathcal{S}}$), while in the generalized setting, they come from both seen and unseen classes ($\mathcal{Y} = \mathcal{Y}^{\mathcal{S}} \cup \mathcal{Y}^{\mathcal{U}}$). We report $Acc_{\mathcal{Y}^{\mathcal{U}}}$ in conventional setting and $Acc_{\mathcal{Y}^{\mathcal{S}}}$, $Acc_{\mathcal{Y}^{\mathcal{U}}}$, harmonic mean $\mathcal{H} = \frac{2 * Acc_{\mathcal{Y}^{\mathcal{U}}} * Acc_{\mathcal{Y}^{\mathcal{S}}}}{Acc_{\mathcal{Y}^{\mathcal{U}}} + Acc_{\mathcal{Y}^{\mathcal{S}}}}$ in generalized setting.

Implementations Different backbone networks including GoogleNet [40], VGG19 [37] and ResNet101 [14] are used to initialize our Embedding Subnet, and images are randomly cropped to the corresponding size before fed into the LFGAA network. The triplet margin α and ridge regression λ are both set to 1.0 for all the experiments. We select four feature maps of different sizes and learn attribute attention in different visual levels. We set iteration SA steps $T = 10$ but in practice it converges on the first few iterations. We perform online triplet mining with batch-hard strategy and the whole LFGAA is trained in an end-to-end manner with Adam optimizer [19] throughout all the experiments.

5.2. Conventional Comparison

We make the conventional ZSL (CZSL) comparison with several state-of-the-art transductive ZSL methods [39, 21, 42] and competitive inductive ZSL methods [22, 29, 5, 17, 9, 35, 2, 4, 16]. We conduct both inductive **LFGAA+Hybrid** and transductive **LFGAA+SA** that has no differences except for prototype construction described in Section 4.4 on CZSL and results are shown in Table 1.

Overall Performance It can be seen from Table 1 that proposed attribute attention achieves the state-of-the-art performances in both inductive and transductive settings on CZSL. Verified from experimental results on different backbone networks, our method is not only effective to a specific CNN model or specific data split. We also find the attribute (*shape*, *color* of different body regions) in CUB is much simpler than SUN that involves scene understanding and attribute attention benefits from low-level visual information.



(a) LFGAA+Hybrid (b) LFGAA+SA ($T = 1$) (c) LFGAA+SA ($T = 5$) (d) Ground Truth

Figure 5: Visualization of latent feature predictions and latent class prototypes on AwA2 unseen objects. Comparisons among different methods are shown in (a-c) where those methods share the same feature network (that leads to the same point distributions) but differ in prototype constructions only. Colors in (a-c) represent **prediction labels** while colors in (d) represent the **ground truth**. We use black circles to mark latent prototypes in each method.

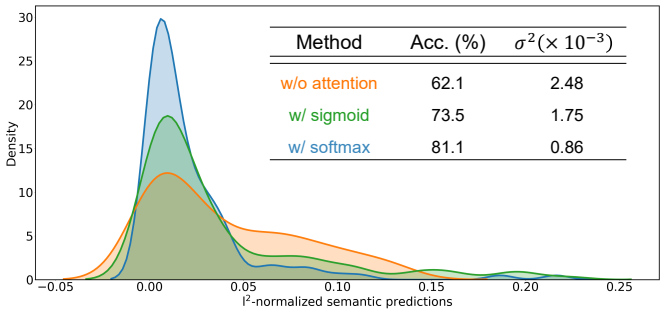


Figure 6: Comparisons on CUB dataset with proposed data split [46] and LFGAA^G+SA. σ^2 is the l^2-normalized prediction variance of 312 semantic dimensions.

Comparisons with state-of-the-art methods As shown in Table 1, our attention based method outperforms the state-of-the-art on three ZSL datasets. Compared in inductive settings, our LFGAA+Hybrid shows superiority on both AwA2 and CUB and achieves comparable performance on SUN. In transductive settings, our proposed LFGAA+SA has an impressive gain over the state-of-the-art QFSL [39] by a large margin of $5.1\% \sim 10.3\%$ on both AwA2 and challenging CUB datasets. Our performance improvement on SUN is slightly lower than on the other datasets because of the scarcity (about 20 unseen instances within the class) where our self-adapted prototypes are not representative enough for 72 unseen categories.

Effective Attribute Attention The architecture of our LFGAA network follows LDF [24] to learn in both semantic and latent feature space. Differently, we introduce a third attention branch to make predictions more discriminative for those challenging objects. Our baseline LFGAA+Hybrid (84.3% on AwA2, 67.6% on CUB) outperforms the results reported in LDF [24] (81.4% on AwA [22], 65.9% on CUB) and achieves the state-of-the-art.

Table 2: Ablation results (%) on all three datasets (proposed split) with the same backbones.

Method	AwA2	CUB	SUN
LFGAA+Hybrid+w/o attention	62.4	62.1	61.2
LFGAA+Hybrid	68.1	67.6	61.5
LFGAA+SA+w/o attention	70.9	67.4	62.8
LFGAA+SA	75.5	78.9	66.2

To gain an insight into the effectiveness of proposed attention, we conduct another experiment on the CUB dataset by comparing with methods **LFGAA w/o attention** and **LFGAA w/ sigmoid**, and the results are shown in Figure 6. It can be observed that, with proposed attention mechanism, the *variance* of predictions among different dimensions is reduced which shows the basic idea discussed in Section 3.2 that *prominent attributes shouldn't completely dominate in discrimination*. It can also be found that the distracting effect of **softmax-based** attention is better since it creates competitions among different attribute dimensions in training. We also visualize the normalized semantic predictions of each unseen class for those three methods and their approximate density functions with kernel density estimation, which also shows this distracting effect.

To further probe the efficacy of attribute attention, we perform an ablation study for the CZSL setting. The results in Table 2 clearly demonstrate that proposed attention benefits in both inductive and transductive settings.

Effective Self-Adaptation Domain shift problem exists in methods [17, 31, 24] that learn in both semantic and latent space. By directly building prototypes in latent space, our self-adaptation based method outperforms hybrid based one by an obvious margin. To prove that our performance boost

Table 3: Comparisons in the generalized ZSL setting (%). For each dataset, the best result is marked in **bold** font.

Method	AwA2			CUB			SUN		
	$Acc_{\mathcal{Y}^\mathcal{U}}$	$Acc_{\mathcal{Y}^\mathcal{S}}$	$\mathcal{H}$	$Acc_{\mathcal{Y}^\mathcal{U}}$	$Acc_{\mathcal{Y}^\mathcal{S}}$	$\mathcal{H}$	$Acc_{\mathcal{Y}^\mathcal{U}}$	$Acc_{\mathcal{Y}^\mathcal{S}}$	$\mathcal{H}$
SYNC[5]	10.0	90.5	18.0	11.5	70.9	19.8	7.9	**43.3**	13.4
DEVISE[9]	17.1	74.7	27.8	23.8	53.0	32.8	14.7	30.5	19.8
ESZSL[35]	5.9	77.8	11.0	12.6	63.8	21.0	11.0	27.9	15.8
CMT[38]	0.5	90.0	1.0	7.2	49.8	12.6	8.1	21.8	11.8
CMT*[38]	8.7	89.0	15.9	4.7	60.1	8.7	8.7	28.0	13.3
CDL[16]	29.3	73.9	41.9	23.5	55.2	32.9	21.5	34.7	26.5
Y. Annadani *et al*. [4]	20.7	73.8	32.3	24.6	54.3	33.9	20.8	37.2	26.7
f-CLSWGAN[47]	57.9	61.4	59.6	43.7	57.7	49.7	**42.6**	36.6	**39.4**
SE-ZSL[42]	**58.3**	68.1	62.8	41.5	53.3	46.7	40.9	30.5	34.9
LFGAA+Hybrid (Ours)	27.0	**93.4**	41.9	36.2	**80.9**	50.0	18.5	40.0	25.3
LFGAA+SA (Ours)	50.0	90.3	**64.4**	**43.4**	79.6	**56.2**	20.8	34.9	26.1

mainly comes from the correction of latent prototypes, we use t-SNE [41] to visualize this correction process in Figure 5 (a-c). It can be seen that most of the hybrid prototypes (*e.g. persian cat*, *giant panda* and *rat*) are at the edge of clusters while self-adapted prototypes gradually move to the center. With self-adaptation strategy, our final label predictions are also separable in latent space alone.

5.3. Generalized Comparisons

We also apply our method in generalized ZSL (GZSL) to further demonstrate its effectiveness. Latent feature prototypes of seen classes and unseen classes are jointly obtained by self-adaptation with pseudo labels extending from $\mathcal{Y}^\mathcal{U}$ to $\mathcal{Y} = \mathcal{Y}^\mathcal{S} \cup \mathcal{Y}^\mathcal{U}$ and the results are shown in Table 3.

Our performance boost mainly comes from the improvement both in $Acc_{\mathcal{Y}^\mathcal{S}}$ and $Acc_{\mathcal{Y}^\mathcal{U}}$ but the issue of the bias towards seen classes [6, 46] still exists since our LFGAA model has no access to the unseen images. Although we achieve outstanding CZSL performance on SUN, our GZSL performance is not as good as other transductive methods on SUN as it is an extremely biased dataset (seen classes are $\sim$ 9 times of unseen classes). Methods like SE-ZSL [42] and f-CLSWGAN [47] use synthetic unseen examples to remove this bias at the cost of $Acc_{\mathcal{Y}^\mathcal{S}}$ drop.

6. Quality of Disambiguation

The *information amount* (IA) introduced in Section 3.2 reflects the importance of attribute j in binary classifying class y_1 from y_2. IA is originally image-irrelevant but becomes relevant in our method as we propose to re-weigh the importance of attribute according to image I.

$$\mathrm{IA}^*(I, j, y_1, y_2) = \frac{p(I)_j(\boldsymbol{a}_{y_1,j} - \boldsymbol{a}_{y_2,j})}{\sum_i \|\boldsymbol{a}_{y_1,i} - \boldsymbol{a}_{y_2,i}\|_1 p(I)_i} \tag{15}$$

We show the disambiguation of misleading attribute IA in Figure 7 where IA is drastically reduced for ambiguous examples but not impacted much for its counterpart.

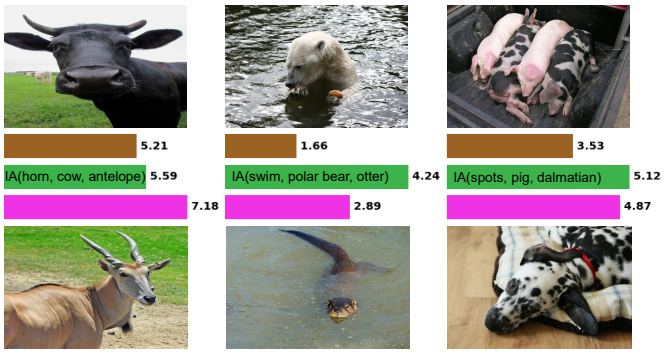


Figure 7: IA in binary classification. Green bar is w/o attention, brown bar is w/ attention for ambiguous example and purple bar is w/ attention for its counterpart.

7. Conclusion

In this paper, we present the drawback of equal treatment on different semantic dimensions, especially when dealing with semantic-ambiguous objects. It is reasonable to think that classifications should depend on multiple factors instead of one or a few prominent semantic predictions. Motivated by this, we propose an end-to-end attention framework to distract semantic predictions that may cause ambiguity. The proposed attention, which is learned independently from semantic space, integrates both low-level visual information and global category features in discrimination. Various experimental results conducted on different datasets demonstrate its efficiency in both inductive and transductive settings.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (Grant Nos: 61751307) and the National Youth Top-notch Talent Support Program.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2013.

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

[3] Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5975–5984, 2016.

[4] Yashas Annadani and Soma Biswas. Preserving semantic relations for zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5327–5336, 2016.

[6] 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 *European Conference on Computer Vision*, pages 52–68. Springer, 2016.

[7] Zhengming Ding, Ming Shao, and Yun Fu. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6005–6013, 2017.

[8] Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785, 2009.

[9] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26*, pages 2121–2129, 2013.

[10] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):23322345, Nov 2015.

[11] Yasuhiro Fujiwara and Go Irie. Efficient label propagation. In *Proceedings of The 31st International Conference on Machine Learning*, pages 784–792, 2014.

[12] Yuchen Guo, Guiguang Ding, Jungong Han, and Sheng Tang. Zero-shot learning with attribute selection. In *AAAI-18 AAAI Conference on Artificial Intelligence*, 2018.

[13] Yuchen Guo, Guiguang Ding, Xiaoming Jin, and Jianmin Wang. Transductive zero-shot recognition via shared model space learning. In *AAAI*, volume 3, page 8, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[15] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *Advances in neural information processing systems*, pages 3464–3472, 2014.

[16] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Learning class prototypes via structure alignment for zero-shot recognition. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[17] Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, and Xilin Chen. Learning discriminative latent attributes for zero-shot classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4223–4232, 2017.

[18] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. Gaze embeddings for zero-shot image classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[19] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *international conference on learning representations*, 2015.

[20] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015.

[21] Soheil Kolouri, Mohammad Rostami, Yuri Owechko, and Kyungnam Kim. Joint dictionaries for zero-shot learning. *national conference on artificial intelligence*, 2018.

[22] 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.

[23] Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, and Yueting Zhuang. Zero-shot recognition using dual visual-semantic mapping paths. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3279–3287, 2017.

[24] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Discriminative learning of latent features for zero-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7463–7471, 2018.

[25] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2624–2637, 2013.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *neural information processing systems*, pages 3111–3119, 2013.

[27] Pedro Morgado and Nuno Vasconcelos. Semantically consistent regularization for zero-shot recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[28] Pedro Morgado and Nuno Vasconcelos. Semantically consistent regularization for zero-shot recognition. In *CVPR*, volume 9, page 10, 2017.

[29] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S. Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *international conference on learning representations*, 2014.

[30] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758, 2012.

[31] Peixi Peng, Yonghong Tian, Tao Xiang, Yaowei Wang, Massimiliano Pontil, and Tiejun Huang. Joint semantic and latent attribute modelling for cross-class transfer learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1625–1638, 2018.

[32] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[33] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[34] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. In *Advances in Neural Information Processing Systems 26*, pages 46–54, 2013.

[35] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *international conference on learning representations*, 2015.

[38] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. *neural information processing systems*, pages 935–943, 2013.

[39] Jie Song, Chengchao Shen, Yezhou Yang, Yang Liu, and Mingli Song. Transductive unbiased embedding for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1024–1033, 2018.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[41] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[42] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR18)*, 2018.

[43] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. *computer vision and pattern recognition*, 2018.

[44] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[45] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh N. Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 69–77, 2016.

[46] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 11, 2018.

[47] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018.

[48] Meng Ye and Yuhong Guo. Zero-shot classification with discriminative semantic representation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[49] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015.