

MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning

Shiming Chen¹, Ziming Hong¹, Guo-Sen Xie², Wenhan Yang³, Qinmu Peng¹, Kai Wang⁴,
 Jian Zhao⁵, Xinge You^{1*}

¹Huazhong University of Science and Technology (HUST), China

²Nanjing University of Science and Technology, China ³Nanyang Technological University, Singapore

⁴National University of Singapore, Singapore ⁵Institute of North Electronic Equipment, China

{shimingchen, pengqinmu, youxg}@hust.edu.cn hoongzm@gmail.com

Abstract

The key challenge of zero-shot learning (ZSL) is how to infer the latent semantic knowledge between visual and attribute features on seen classes, and thus achieving a desirable knowledge transfer to unseen classes. Prior works either simply align the global features of an image with its associated class semantic vector or utilize unidirectional attention to learn the limited latent semantic representations, which could not effectively discover the intrinsic semantic knowledge (e.g., attribute semantics) between visual and attribute features. To solve the above dilemma, we propose a Mutually Semantic Distillation Network (MSDN), which progressively distills the intrinsic semantic representations between visual and attribute features for ZSL. MSDN incorporates an attribute→visual attention sub-net that learns attribute-based visual features, and a visual→attribute attention sub-net that learns visual-based attribute features. By further introducing a semantic distillation loss, the two mutual attention sub-nets are capable of learning collaboratively and teaching each other throughout the training process. The proposed MSDN yields significant improvements over the strong baselines, leading to new state-of-the-art performances on three popular challenging benchmarks. Our codes have been available at: <https://github.com/shiming-chen/MSDN>.

1. Introduction

Recently, deep learning performs achievements on object recognition [12, 39, 40]. Based on the prior knowledge of seen classes, humans possess a remarkable ability to recognize new concepts (classes) using shared and distinct attributes of both seen and unseen classes [17]. Inspired by this cognitive competence, zero-shot learning (ZSL) is proposed under a challenging image classification setting to

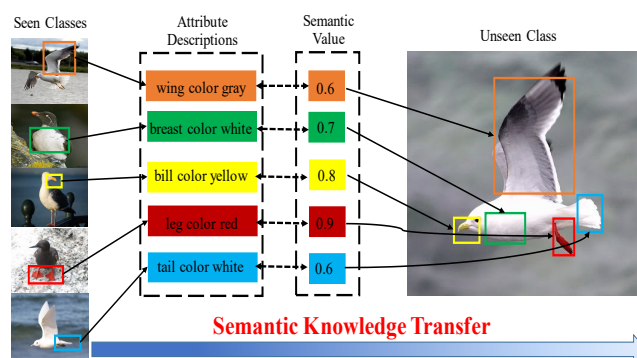


Figure 1. Motivation illustration. An unseen sample shares different partial information with a set of seen samples, and this partial information is represented as the abundant knowledge of semantic attributes (e.g., “bill color yellow”, “leg color red”). The key challenge of ZSL is how to infer the latent semantic knowledge between visual and attribute features on seen classes, allowing effective knowledge transfer to unseen classes. As such, properly distilling the intrinsic semantic knowledge/representations (e.g., attribute semantics) between visual and attribute features from seen to unseen classes is very necessary for ZSL.

mimic the human cognitive process [19, 28]. ZSL aims to tackle the unseen class recognition problem by transferring semantic knowledge from seen classes to unseen ones. It is usually based on the assumption that both seen and unseen classes can be described through the shared semantic descriptions (e.g., attributes) [18]. Based on the classes that a model sees in the testing phase, ZSL methods can be categorized into conventional ZSL (CZSL) and generalized ZSL (GZSL) [44], where CZSL aims to predict unseen classes, while GZSL can predict both seen and unseen ones.

ZSL has achieved significant progress, with many efforts focus on embedding-based methods, generative methods, and common space learning-based methods. As shown in Fig. 2 (a), embedding-based methods aim to learn a visual→semantic mapping to map the visual fea-

*Corresponding author

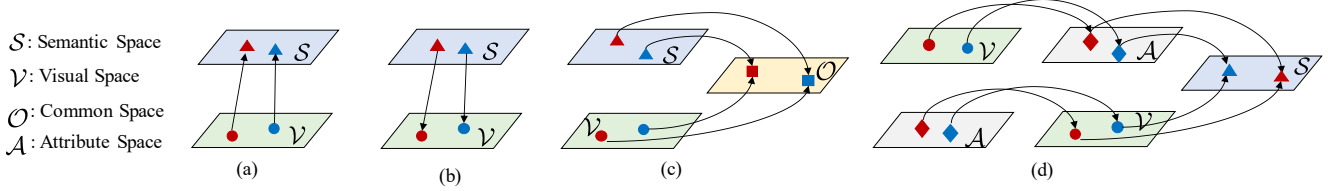


Figure 2. Four investigated ZSL paradigms. (a) Embedding-based method. (b) Generative method. (c) Common space learning-based method. (d) Ours proposed mutually semantic distillation network (MSDN). The semantic space \mathcal{S} is represented by the class semantic vector annotated by humans based on the attribute descriptions. The visual space \mathcal{V} is learned by a CNN backbone (e.g., ResNet101). The common space \mathcal{O} is a shared latent space between visual mapping and semantic mapping. The attribute space \mathcal{A} is learned by a language model (e.g., Glove [31]). Filled triangles, circles, squares and diamonds denote the sample features in \mathcal{S} , \mathcal{V} , \mathcal{O} and \mathcal{A} , respectively.

tures into the semantic space for visual-semantic interaction [2, 4, 5, 32, 46, 48]. The embedding-based methods usually have a large bias towards seen classes under the GZSL setting, since the embedding function is solely learned by seen class samples. To solve this issue, the generative ZSL methods (see Fig. 2(b)) are proposed to learn semantic \rightarrow visual mapping to generate visual features of unseen classes [3, 6, 8, 34, 35, 38, 43, 50], and thus converting ZSL into a conventional classification problem. As shown in Fig. 2(c), common space learning learns a common representation space where both visual features and semantic representations are projected for knowledge transfer [7, 10, 23, 34, 37, 41]. However, they simply utilize the global features representations and have neglected the fine-grained details in the training images.

As shown in Fig. 1, an unseen sample shares different partial information with a set of seen samples, and this partial information is represented as the abundant knowledge of semantic attributes (e.g., “bill color yellow”, “leg color red”). Thus, the key challenge of ZSL is to infer the latent semantic knowledge between visual and attribute features on seen classes, and thus allowing desirable knowledge transfer to unseen classes. Recently, some attention-based ZSL methods [5, 25, 46–48, 54] leverage attribute descriptions as guidance to discover discriminative part/fine-grained features, enabling to match the semantic representations more accurately. Unfortunately, they simply utilize unidirectional attention, which only focuses on limited semantic alignments between visual and attribute features without any further sequential learning. As such, properly discovering the intrinsic and more sufficient semantic representations (e.g., attribute semantics) between visual and attribute features for knowledge transfer of ZSL is of great importance.

In light of the above observation, we propose a Mutually Semantic Distillation Network (MSDN) for ZSL, as shown in Fig. 2(d), to explore the intrinsic semantic knowledge between visual and attribute features. MSDN consists of an attribute \rightarrow visual attention sub-net, which learns attribute-based visual features, and a visual \rightarrow attribute attention sub-net, which learns visual-based attribute features. These two

mutual attention sub-nets act as a teacher-student network for guiding each other to learn collaboratively and teaching each other throughout the training process. As such, MSDN can explore the most matched attribute-based visual features and visual-based attribute features, enabling to effectively distill the intrinsic semantic representations for desirable knowledge transfer from seen to unseen classes (Fig. 1). Specifically, each attention sub-net is trained with an attribute-based cross-entropy loss with self-calibration [5, 14, 25, 48, 54]. To encourage mutual learning between the attribute \rightarrow visual attention sub-net and visual \rightarrow attribute attention sub-net, we further introduce a semantic distillation loss that aligns each other’s class posterior probabilities. The quantitative and qualitative results well demonstrate the superiority and great potential of MSDN.

Our contributions are summarized as: **i)** We propose a Mutually Semantic Distillation Network (MSDN), orthogonal to existing ZSL methods, which distills the intrinsic semantic representations for effective knowledge transfer from seen to unseen classes for ZSL. **ii)** We introduce a semantic distillation loss to enable mutual learning between the attribute \rightarrow visual attention sub-net and visual \rightarrow attribute attention sub-net in MSDN, encouraging them to learn attribute-based visual features and visual-based attribute features by distilling the intrinsic semantic knowledge for semantic embedding representations. **iii)** We conduct extensive experiments to show that our MSDN achieves significant performance gains over the counterparts on three benchmarks, i.e., CUB [42], SUN [30] and AWA2 [44].

2. Related Work

2.1. Zero-Shot Learning

To transfer semantic knowledge from seen to unseen classes, ZSL [6, 9, 11, 22, 27, 36, 43, 45, 50] learns a mapping between the visual and attribute/semantic domains. Targeting on this goal, embedding-based ZSL aims to learn a visual \rightarrow semantic mapping for visual-semantic interaction by mapping the visual features into the semantic space [2, 4, 32, 46, 48]. As the embedding is learned only on seen classes, these embedding-based methods inevitably overfit

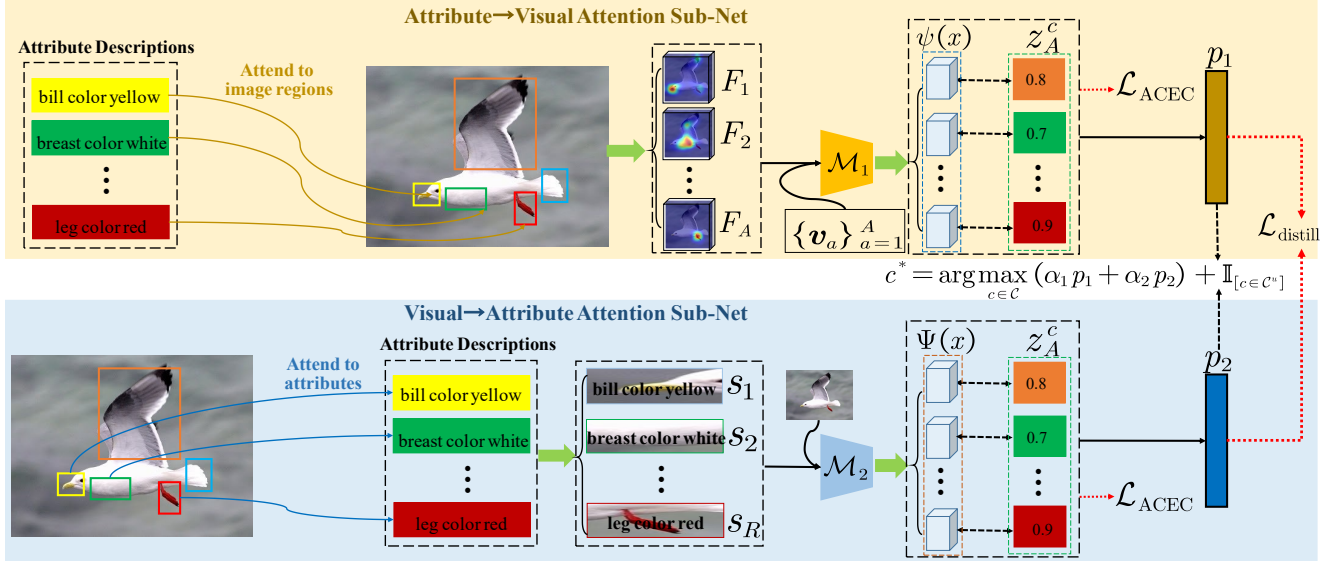


Figure 3. Illustrates of MSDN. MSDN consists of an attribute→visual attention sub-net and visual→attribute attention sub-net. Each sub-net is optimized with an attribute-based cross-entropy loss with self-calibration, and a semantic distillation loss to match the probability estimates of its peers for semantic distillation.

to seen classes under the GZSL setting. To mitigate this challenge, the generative ZSL models have been introduced to learn a semantic→visual mapping to generate visual features of of unseen classes [3, 6, 20, 34, 35, 38, 43, 50] for data augmentation. Currently, the generative ZSL usually based on variational autoencoders (VAEs) [3, 34], generative adversarial nets (GANs) [6, 16, 21, 38, 43, 50], and generative flows [35]. Furthermore, common space learning is also employed to learns a common representation space for interaction between visual and semantic domains [7, 10, 34, 37]. However, these methods still usually yield relatively undesirable results, since they cannot capture the subtle differences between seen and unseen classes. As such, attention-based ZSL methods [5, 25, 46–48, 54] utilize attribute descriptions as guidance to discover the more discriminative fine-grained features. They simply utilize unidirectional attention, which only focuses on limited semantic alignments between visual and attribute features without any further sequential learning. As such, properly exploring the intrinsic semantic representations between visual and attribute features for knowledge transfer of ZSL is necessary.

2.2. Knowledge Distillation

To compress knowledge from a large teacher network to a small student network, knowledge distillation was proposed [13]. Recently, knowledge distillation is extended to optimize small deep networks starting with a powerful teacher network [29, 33]. By mimicking the teacher’s class probabilities and/or feature representation, distilling models convey additional information beyond the conventional supervised learning target [52, 53]. Motivated by these, we

design a mutually semantic distillation network to learn the intrinsic semantic by semantically distilling intrinsic knowledge. The mutually semantic distillation network consists of attribute→visual attention and visual→attribute attention sub-nets, which act as a teacher-student network to learn collaboratively and teach each other.

3. Mutually Semantic Distillation Network

Motivation. Prior works simply i) align the global features of an image with its associated class semantic vector, neglecting the fine-grained information for knowledge transfer, or ii) utilize unidirectional attention to learn the latent semantic representations, which only focuses on limited semantic alignments between visual and attribute features without any further sequential learning. However, an unseen sample can share different partial information with a set of seen samples, and this partial information is represented as the abundant knowledge of semantic attributes, as shown in Fig. 1. These observations prompt us to speculate that the current inferior performance of ZSL is closely related to the intrinsic semantic representations (*e.g.*, attribute semantics) between visual and attribute features, which offers effective knowledge transfer to unseen classes.

To properly learn the intrinsic semantic knowledge, we propose a Mutually Semantic Distillation Network (MSDN). Our strategy behind MSDN is to distill the intrinsic semantic knowledge from the attribute-based visual features and visual-based attribute features, which are leaned by two attention sub-nets optimized by a semantic distillation loss.

Overview. As illustrated in Fig. 3, our MSDN includes

an attribute→visual attention sub-net and visual→attribute attention sub-net. Under the constraint of attribute-based cross-entropy loss with self-calibration, the attribute→visual attention sub-net attempts to learn attribute-based visual features, and visual→attribute attention sub-net aims to discover visual-based attribute representations. A semantic distillation loss encourages the two mutual attention sub-nets to learn collaboratively and teach each other throughout the training process.

Notation. Assume that we have training data $\mathcal{D}^s = \{(x_i^s, y_i^s)\}$ with C^s seen classes, where $x_i^s \in \mathcal{X}$ denotes the visual sample i , and $y_i^s \in \mathcal{Y}^s$ is the corresponding class label. Another set of unseen classes C^u has unlabeled samples $\mathcal{D}^u = \{(x_i^u, y_i^u)\}$, where $x_i^u \in \mathcal{X}$ are the unseen class samples, and $y_i^u \in \mathcal{Y}^u$ are the corresponding labels. A set of class semantic vectors (semantic value annotated by humans according to attributes) of the class $c \in C^s \cup C^u = \mathcal{C}$ with $|A|$ attributes $z^c = [z_1^c, \dots, z_A^c]^\top = \phi(y)$ which helps knowledge transfer from seen to unseen classes. Note that we also use the semantic attribute vectors of each attribute $A = \{a_1, \dots, a_K\}$ learned by GloVe [31].

3.1. Attribute→Visual Attention Sub-net

Learning the fine-grained features for attribute localization is important in ZSL [46–48, 54]. As the first component of our MSDN, we proposed an attribute→visual attention sub-net to localize the most relevant image regions to the attribute to extract attribute-based visual features from a given image for each attribute. It expects two inputs: a set of visual features of the image $V = \{v_1, \dots, v_R\}$, such that each visual feature encodes a region in an image; a set of semantic attribute vectors $A = \{a_1, \dots, a_K\}$. We can attend to image regions with respect to each attribute, and compare each attribute to the corresponding attended visual region features to determine the importance of each attribute. For the k -th attribute, we first define its attention weights of focusing on the r -th region of one image as:

$$\beta_k^r = \frac{\exp(a_k^\top W_1 v_r)}{\sum_{k=1}^K \exp(a_k^\top W_1 v_r)}, \quad (1)$$

where W_1 is a learnable matrix to calculate the visual feature of each region and measure the similarity between each semantic attribute vector. As such, we can get a set of attention weights $\{\beta_k^r\}_{r=1}^R$.

We then extract the attribute-based visual features for each attribute based on the attention weights. For example, we get the k -th attribute-based visual feature F_k , which is relevant to the k -th attribute according to the semantic vector a_k . It is formulated as:

$$F_k = \sum_{r=1}^R \beta_k^r v_r. \quad (2)$$

Intuitively, F_k captures the visual evidence used to localize the corresponding semantic attribute in the image. If an image has an obvious attribute a_k , the model will assign a high positive score to the k -th attribute. Otherwise, the model will assign a negative score to the k -th attribute. Thus, we get a set of attribute-based visual features $F = \{F_1, F_2, \dots, F_K\}$.

After extracting the attribute-based visual features, we further introduce a mapping function \mathcal{M}_1 to map them into the semantic embedding space. To encourage the mapping to be more accurate, we take the semantic attribute vectors $A = \{a_1, a_2, \dots, a_K\}$ as support. Specifically, \mathcal{M}_1 matches the attribute-based visual feature F_k with its corresponding semantic attribute vector a_k , formulated as:

$$\psi_k = \mathcal{M}_1(F_k) = a_k^\top W_2 F_k, \quad (3)$$

where W_2 is an embedding matrix that embeds F into the semantic space. Essentially, ψ_k is an attribute score that represents the confidence of having the k -th attribute in an given image. Finally, MSDN obtains a mapped semantic embedding $\psi(x) = \{\psi_1, \psi_2, \dots, \psi_K\}$ for each image.

3.2. Visual→Attribute Attention Sub-net

Analogously, we design a visual→attribute attention sub-net to learn visual-based semantic attribute representations. They are complementary to the attribute-based visual features, enabling them to calibrate each other to discover the intrinsic semantic representations between visual and attribute features. We can first attend to semantic attributes with respect to each image region. Formally, we define its attention weights to focus on the k -th attribute as:

$$\tau_r^k = \frac{\exp(v_r^\top W_3 a_k)}{\sum_{r=1}^R \exp(v_r^\top W_3 a_k)}, \quad (4)$$

where W_3 is a learnable matrix to measure the similarity between the semantic attribute vector and each visual region feature. Thus, we can get a set of attention weights $\{\tau_r^k\}_{k=1}^K$, which is used to extract visual-based attribute features. It is formulated as:

$$S_r = \sum_{k=1}^K \tau_r^k a_k. \quad (5)$$

Intrinsically, S_r is the visual semantic representations, which is aligned to the F_k . We further introduce another mapping function \mathcal{M}_2 to map these visual-based attribute features $S = \{S_1, S_2, \dots, S_R\}$ into semantic space:

$$\bar{\Psi}_r = \mathcal{M}_2(S_r) = v_r^\top W_4 S_r, \quad (6)$$

where W_4 is an embedding matrix. Given a set of $V = \{v_1, \dots, v_R\}$, MSDN gets the mapped semantic embedding

$\bar{\Psi}(x) = \{\bar{\Psi}_1, \bar{\Psi}_2, \dots, \bar{\Psi}_R\}$ for the attributes of one image. To enable the learned semantic embedding $\bar{\Psi}(x_i)$ is R -dim to match with the dimension of class semantic vector (K -dim), it is further mapped into semantic attribute space with K -dim, formulated as $\Psi(x_i) = \bar{\Psi}(x_i) \times Att = \bar{\Psi}(x_i) \times (V^\top W_{att} A)$, where W_{att} is a learnable matrix.

3.3. Model Optimization

To optimize MSDN, each attention sub-net is trained with a supervised learning loss, *i.e.*, attribute-based cross-entropy loss with self-calibration. To encourage mutual learning between the two attention sub-nets, we introduce a semantic distillation loss that aligns each other's class posterior probabilities.

Attribute-Based Cross-Entropy Loss. Since the associated image and attribute embeddings are projected near their class semantic vector z^c when an attribute is visually present in an image, we take the attribute-based cross-entropy loss with self-calibration [14, 48, 54] (denoted as \mathcal{L}_{ACEC}) to optimize the parameters of the MSDN. This enables the image to have the highest compatibility score with its corresponding class semantic vector. Given a batch of n_b training images $\{x_i^s\}_{i=1}^{n_b}$ with their corresponding class semantic vectors z^c , \mathcal{L}_{ACEC} is defined as:

$$\mathcal{L}_{ACEC} = -\frac{1}{n_b} \sum_{i=1}^{n_b} \left[\log \frac{\exp(f(x_i) \times z^c)}{\sum_{\hat{c} \in \mathcal{C}^s} \exp(f(x_i) \times z^{\hat{c}})} - \lambda_{cal} \sum_{c' \in \mathcal{C}^u} \log \frac{\exp(f(x_i) \times z^{c'} + \mathbb{I}_{[c' \in \mathcal{C}^u]})}{\sum_{\hat{c} \in \mathcal{C}} \exp(f(x_i) \times z^{\hat{c}} + \mathbb{I}_{[\hat{c} \in \mathcal{C}^u]})} \right], \quad (7)$$

where $f(x_i) = \psi(x_i)$ for attribute \rightarrow visual attention sub-net and $f(x_i) = \Psi(x_i)$ for visual \rightarrow semantic attention sub-net, $\mathbb{I}_{[c \in \mathcal{C}^u]}$ is an indicator function (*i.e.*, it is 1 when $c \in \mathcal{C}^u$, otherwise -1), and λ_{cal} is a weight to control the self-calibration term. Intuitively, \mathcal{L}_{ACEC} encourages non-zero probabilities to be assigned to the unseen classes during training, thus MSDN produces a large probability for the true unseen class when given test unseen samples.

Semantic Distillation Loss. To enable the two mutual attention sub-nets to learn collaboratively and teach each other throughout the training process, we further introduce a semantic distillation loss $\mathcal{L}_{distill}$ for optimization. $\mathcal{L}_{distill}$ consists of a Jensen-Shannon Divergence (JSD) and an ℓ_2 distance between the predictions of the two attention sub-nets (*i.e.*, $p_1 = \{\psi(x_i) \times z^1, \dots, \psi(x_i) \times z^C\}$ and $p_2 = \{\Psi(x_i) \times z^1, \dots, \Psi(x_i) \times z^C\}$), formulated as:

$$\mathcal{L}_{distill} = \frac{1}{n_b} \sum_{i=1}^{n_b} \left[\underbrace{\frac{1}{2} (D_{KL}(p_1(x_i) || p_2(x_i)) + D_{KL}(p_2(x_i) || p_1(x_i)))}_{JSD} + \underbrace{\|p_1(x_i) - p_2(x_i)\|_2^2}_{\ell_2} \right], \quad (8)$$

where

$$D_{KL}(p || q) = \sum_{c=1}^{C^s} p^c \log\left(\frac{p^c}{q^c}\right). \quad (9)$$

Overall Loss. Finally, we define the overall loss function of MSDN as:

$$\mathcal{L}_{total} = \mathcal{L}_{ACEC} + \lambda_{distill} \mathcal{L}_{distill}, \quad (10)$$

where $\lambda_{distill}$ is a weight to control the semantic distillation loss.

3.4. Zero-Shot Prediction

After training MSDN, We first obtain the embedding features of a test instance x_i in the semantic space w.r.t. the semantic \rightarrow visual and visual \rightarrow semantic attention sub-nets, *i.e.*, $\psi(x)$ and $\Psi(x)$. Then, We fuse their predictions using two combination coefficients (α_1, α_2) to predict the test label of x_i with an explicit calibration, formulated as:

$$c^* = \arg \max_{c \in \mathcal{C}^u / \mathcal{C}} (\alpha_1 \psi(x_i) + \alpha_2 \Psi(x_i))^\top \times z^c + \mathbb{I}_{[c \in \mathcal{C}^u]}. \quad (11)$$

Here, $\mathcal{C}^u / \mathcal{C}$ corresponds to the CZSL/GZSL setting.

4. Experiments

Datasets. We evaluate our method on three challenging benchmark datasets, *i.e.*, CUB (Caltech UCSD Birds 200) [42], SUN (SUN Attribute) [30] and AWA2 (Animals with Attributes 2) [44]. Among these, CUB and SUN are fine-grained datasets, whereas AWA2 is a coarse-grained dataset. Following [44], we use the same seen/unseen splits and class embeddings. Specifically, CUB includes 11,788 images of 200 bird classes (seen/unseen classes = 150/50) with 312 attributes. SUN has 14,340 images from 717 scene classes (seen/unseen classes = 645/72) with 102 attributes. AWA2 consists of 37,322 images from 50 animal classes (seen/unseen classes = 40/10) with 85 attributes.

Evaluation Protocols. We evaluate the top-1 accuracy on unseen classes in the CZSL setting, denoted as *acc*. For GZSL setting, we evaluate the top-1 accuracies both on seen and unseen classes (*i.e.*, S and U), respectively. Furthermore, their harmonic mean (defined as $H = (2 \times S \times U) / (S + U)$) is also employed for evaluating the performance in the GZSL setting.

Implementation Details. We take a ResNet101 [12] pre-trained on ImageNet as the CNN backbone to extract the feature map for each image without fine-tuning. We use the RMSProp optimizer with hyperparameters (momentum = 0.9, weight decay = 0.0001) to optimize our model. We set the learning rate and batch size to 0.0001 and 50, respectively. We empirically set the loss weights $\{\lambda_{cal}, \lambda_{distill}\}$ to $\{0.1, 0.001\}$ for CUB and AWA2, and $\{0.0, 0.01\}$ for SUN.

Table 1. Results (%) of the state-of-the-art CZSL and GZSL modes on CUB, SUN and AWA2, including generative methods, common space-based methods, and embedding-based methods. The best and second-best results are marked in **Red** and **Blue**, respectively. The symbol “–” indicates no results. The symbol “*” denotes attention-based methods.

Methods	CUB				SUN				AWA2			
	CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
	acc	U	S	H	acc	U	S	H	acc	U	S	H
Generative Methods												
f-CLSWGAN(CVPR’18) [43]	57.3	43.7	57.7	49.7	60.8	42.6	36.6	39.4	68.2	57.9	61.4	59.6
f-VAEGAN-D2(CVPR’19) [45]	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
E-PGN(CVPR’20) [50]	72.4	52.0	61.1	56.2	–	–	–	–	73.4	52.6	83.5	64.6
Composer*(NeurIPS’20) [15]	69.4	56.4	63.8	59.9	62.6	55.1	22.0	31.4	71.5	62.1	77.3	68.8
GCM-CF(CVPR’21) [51]	–	61.0	59.7	60.3	–	47.9	37.8	42.2	–	60.4	75.1	67.0
FREE(ICCV’21) [6]	–	55.7	59.9	57.7	–	47.4	37.2	41.7	–	60.4	75.4	67.1
Common Space Learning												
DeViSE(NeurIPS’13) [10]	52.0	23.8	53.0	32.8	56.5	16.9	27.4	20.9	54.2	17.1	74.7	27.8
DCN(NeurIPS’18) [23]	56.2	28.4	60.7	38.7	61.8	25.5	37.0	30.2	65.2	25.5	84.2	39.1
CADA-VAE(CVPR’19) [34]	59.8	51.6	53.5	52.4	61.7	47.2	35.7	40.6	63.0	55.8	75.0	63.9
SGAL(NeurIPS’19) [49]	–	40.9	55.3	47.0	–	35.5	34.4	34.9	–	52.5	86.3	65.3
HSVA(NeurIPS’21) [7]	62.8	52.7	58.3	55.3	63.8	48.6	39.0	43.3	–	59.3	76.6	66.8
Embedding-based Methods												
SP-AEN(CVPR’18) [4]	55.4	34.7	70.6	46.6	59.2	24.9	38.6	30.3	58.5	23.3	90.9	37.1
SGMA*(NeurIPS’19) [54]	71.0	36.7	71.3	48.5	–	–	–	–	68.8	37.6	87.1	52.5
AREN*(CVPR’19) [46]	71.8	38.9	78.7	52.1	60.6	19.0	38.8	25.5	67.9	15.6	92.9	26.7
LFGAA*(ICCV’19) [24]	67.6	36.2	80.9	50.0	61.5	18.5	40.0	25.3	68.1	27.0	93.4	41.9
DAZLE*(CVPR’20) [14]	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
APN*(NeurIPS’20) [48]	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
MSDN (Ours)	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3	70.1	62.0	74.5	67.7

Table 2. Ablation studies for different components of MSDN. The *baseline* is the visual feature extracted from CNN backbone with a global average pooling and then mapped into semantic embedding for ZSL.

Method	CUB		SUN	
	acc	H	acc	H
baseline	57.4	49.1	54.8	30.5
MSDN(V→A) w/o $\mathcal{L}_{\text{distill}}$	66.0	55.4	59.2	33.8
MSDN(A→V) w/o $\mathcal{L}_{\text{distill}}$	73.4	65.4	63.8	38.5
MSDN(V→A) w/ $\mathcal{L}_{\text{distill}}$	67.9	60.8	62.1	38.6
MSDN(A→V) w/ $\mathcal{L}_{\text{distill}}$	75.2	67.5	63.0	38.7
MSDN w/ $\mathcal{L}_{\text{distill}}$ (JSD)	74.3	67.0	64.7	39.4
MSDN w/ $\mathcal{L}_{\text{distill}}$ (ℓ_2)	74.4	67.6	64.9	40.8
MSDN (full)	76.1	68.1	65.8	41.3

4.1. Comparison with State-of-the-Arts

Conventional Zero-Shot Learning. We first compare our MSDN with the state-of-the-art methods in the CZSL setting. Table 1 presents the results of CZSL on various datasets. Our MSDN achieves the best accuracies of 76.1% and 65.8% on CUB and SUN, respectively. This shows that MSDN distills the intrinsic semantic representations for distinguishing fine-grained unseen classes. As shown in Fig. 1, MSDN can distill the semantic attributes of “bill color yellow”, “breast color white” and “leg color red” from *Aliforina Gull*, *Parakeet Auklet*, *Pigeon Guillemot*, transferring to unseen classes (e.g., *Red Legged Kittiwake*). As

for the coarse-grained dataset (i.e., AWA2), MSDN still obtains competitive performance, with a top-1 accuracy of 70.1%. Compared to other embedding-based methods, MSDN achieves new state-of-the-art on all datasets.

Generalized Zero-Shot Learning. Table 1 also shows the results of different methods in the GZSL setting, i.e., embedding-based methods, generative methods, and common space learning methods. Interestingly, most state-of-the-art methods achieve good results on seen classes but fail on unseen classes on CUB and AWA2, while our MSDN generalizes well to unseen classes with high seen and unseen accuracies. As such, MSDN achieves good results of Harmonic mean, e.g., 68.1% and 67.7% on CUB and AWA2, respectively. These benefits come from the semantic distillation of MSDN, enabling to discover the intrinsic semantic representations for effective knowledge transfer from seen to unseen classes. Compared to attention-based ZSL methods [24, 46–48, 54] that utilize attribute descriptions as guidance to discover the more discriminative fine-grained features, our MSDN also achieves significant improvement on Harmonic mean at least 3.7% on SUN. This demonstrates the superiority and potential of the proposed MSDN for ZSL.

4.2. Ablation Studies

To provide further insight into our MSDN, we conduct ablation studies to evaluate the effectiveness of our

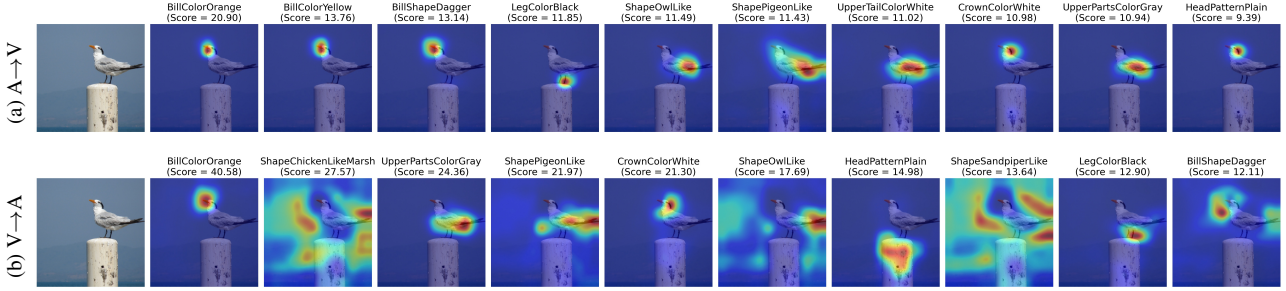


Figure 4. Visualization of attention maps for the two mutual attention sub-nets (i.e., MSDN($A \rightarrow V$) and MSDN($V \rightarrow A$)). The scores are the attribute scores.

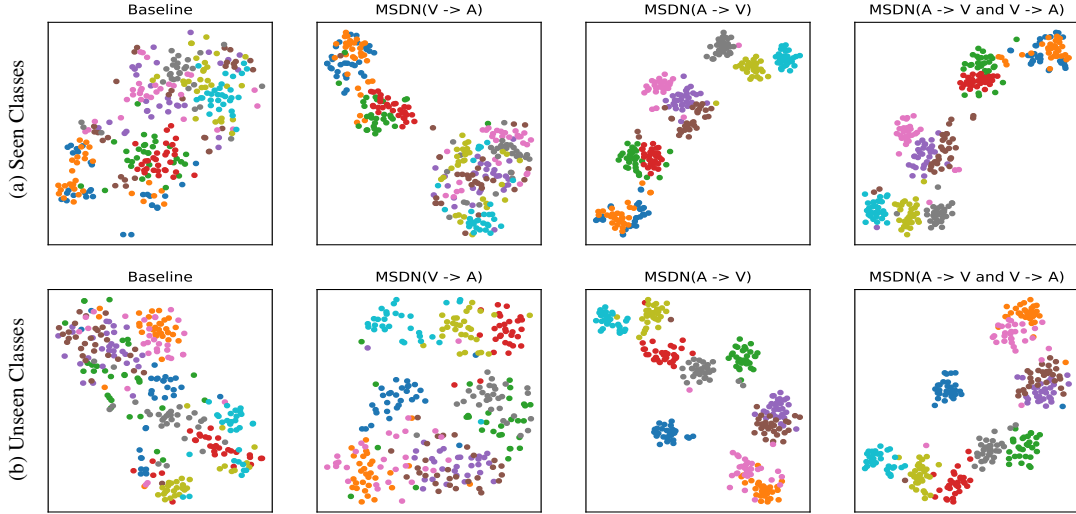


Figure 5. t-SNE visualizations of visual features for (a) seen classes and (b) unseen classes, learned by the baseline, MSDN($V \rightarrow A$), MSDN($A \rightarrow V$), and MSDN($V \rightarrow A$ and $A \rightarrow V$). The 10 colors denote 10 different seen/unseen classes randomly selected from CUB.

MSDN in terms of the $V \rightarrow A$ attention sub-net (denoted as MSDN($V \rightarrow A$) w/o $\mathcal{L}_{\text{distill}}$), $A \rightarrow V$ attention sub-net (denoted as MSDN($A \rightarrow V$) w/o $\mathcal{L}_{\text{distill}}$), semantic distillation loss (i.e., MSDN($V \rightarrow A$) w/ $\mathcal{L}_{\text{distill}}$, MSDN($A \rightarrow V$) w/ $\mathcal{L}_{\text{distill}}$), semantic distillation loss with JSD (i.e., MSDN w/ $\mathcal{L}_{\text{distill}}$ (JSD)) and ℓ_2 (i.e., MSDN w/ $\mathcal{L}_{\text{distill}}$ (ℓ_2)). Our results are shown in Table 2. Compared to the baseline, MSDN only employs the single attention sub-net without semantic distillation achieving significant improvements. For example, MSDN($V \rightarrow A$) w/o $\mathcal{L}_{\text{distill}}$ achieves the gains of acc/H by 8.6%/6.3% and 4.4%/3.3% on CUB and SUN respectively, MSDN($A \rightarrow V$) w/o $\mathcal{L}_{\text{distill}}$ gets the acc/H improvements of 16.0%/16.3% and 9.0%/8.0% on CUB and SUN respectively. This is beneficial from that MSDN refines the visual features, alleviating the cross-dataset bias problem [6]. If MSDN is optimized by the semantic distillation loss, its results can be further improved, e.g., MSDN($V \rightarrow A$) improves the Harmonic mean by 5.4% and 4.8% on CUB and SUN, respectively. These results show that semantic distillation encourages the two mutual attention sub-nets to learn collaboratively and teach each other, and thus the intrinsic semantic representations can be distilled for knowledge transfer. When the semantic distillation loss only uses

one distance, i.e., JSD or ℓ_2 , the distilling capacity of MSDN are limited. Moreover, our full model ensembles the complementary embeddings learned by the two mutual attention sub-nets to further improve the feature representations, achieving acc/Harmonic mean improvements of 18.7%/19.0% and 11.0%/10.8% on CUB and SUN over the baseline, respectively.

4.3. Qualitative Results

Visualization of Attention Maps. To intuitively show the effectiveness of our MSDN at distilling the intrinsic semantic, we visualize the attention maps learned by the two mutual attention sub-nets, e.g., MSDN($A \rightarrow V$) and MSDN($V \rightarrow A$). As shown in Fig. 4, MSDN($A \rightarrow V$) and MSDN($V \rightarrow A$) sub-nets effectively learn the attribute-based visual features and visual-based attribute features for representing the discriminative attribute localizations. MSDN($A \rightarrow V$) and MSDN($V \rightarrow A$) can similarly learn the most important semantic representation, which is beneficial from mutual learning for semantic distillation. Furthermore, the two attention sub-nets also learn the complementary attribute feature localization for each other. For example, MSDN($A \rightarrow V$) can well learn the key semantic of

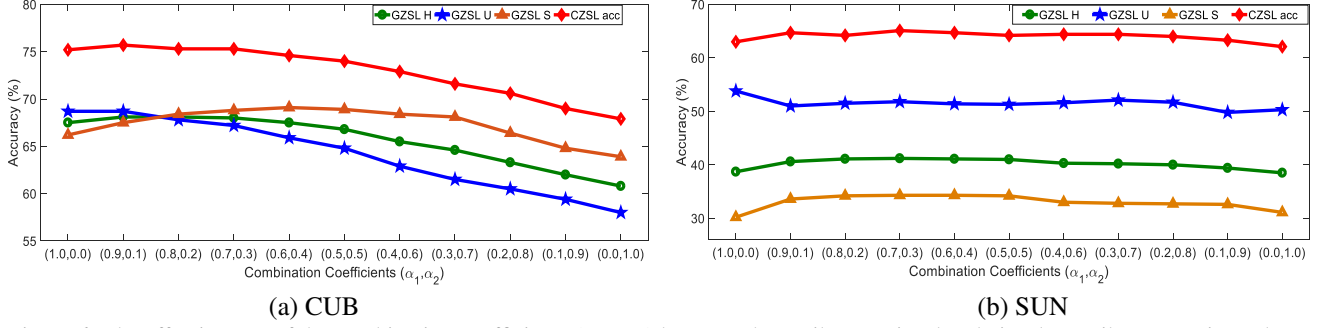


Figure 6. The effectiveness of the combination coefficients (α_1, α_2) between the attribute \rightarrow visual and visual \rightarrow attribute attention sub-nets.

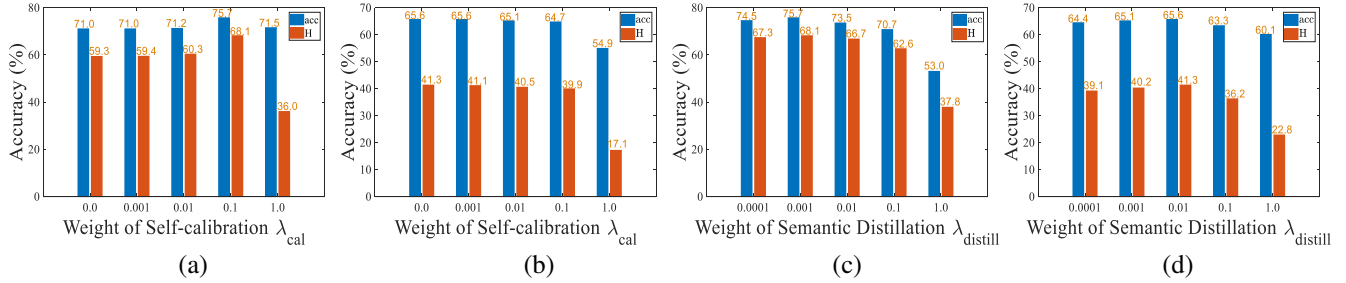


Figure 7. The effects of λ_{cal} on (a) CUB and (b) SUN. The effects of $\lambda_{distill}$ on (c) CUB and (d) SUN. (Best viewed in color)

“under tail white” but no “shape chicken-like marsh” for *Carolina Wren*, while MSDN(V \rightarrow A) confidently learn the important semantic “shape chicken-like marsh” but not “under tail white”. Thus, our full MSDN achieves significant performance both in seen and unseen classes.

t-SNE Visualizations. As shown in Fig. 5, we also present the t-SNE visualization [26] of visual features for seen and unseen classes on CUB, learned by the baseline, MSDN(A \rightarrow V), MSDN(V \rightarrow A), and combination of the two attention sub-nets (*i.e.*, MSDN(V \rightarrow A and A \rightarrow V)). Compared to the baseline, our models learn the intrinsic semantic representations both in seen and unseen classes. This shows that our MSDN can simultaneously learn the discriminative and transferable features for effective knowledge transfer in ZSL. As such, our MSDN achieves significant improvement over baseline.

4.4. Hyperparameter Analysis

Effects of Combination Coefficients. we conduct experiments to determine the effectiveness of the combination coefficients (α_1, α_2) between attribute \rightarrow visual and visual \rightarrow attribute attention sub-nets. As shown in Fig. 6, MSDN performs poorly when α_1/α_2 is set too small or large, because both the attribute-based visual features and visual-based attribute features are complementary for discriminative semantic embedding representations. When the combination coefficients α_1, α_2 are set to (0.9,0.1) and (0.7,0.3) on CUB and SUN, respectively, MSDN achieves the best results.

Effects of Loss Weights. Here we study how to set the related loss weights of MSDN: λ_{cal} and $\lambda_{distill}$, which control the self-calibration term and semantic distillation loss,

respectively. Based on the results in Fig. 7 (a) and (b), we choose λ_{cal} as 0.1 for CUB/AWA2. Since the number of seen classes is much larger than the number of unseen classes and per class only contains 16 training images on SUN, thus it usually overfits unseen classes. As such, we set λ_{cal} to 0.0 for SUN. According to the results in Fig. 7 (c) and (d), we set $\lambda_{distill}$ to 0.001 and 0.01 for CUB/AWA2 and SUN, respectively.

5. Conclusion and Discussion

In this paper, we propose a novel mutually semantic distillation network (MSDN) for ZSL. MSDN consists of two mutual attention sub-nets, *i.e.*, attribute \rightarrow visual and visual \rightarrow semantic attention sub-nets, which learns attribute-based visual features and visual-based attribute features for semantic embedding representations, respectively. To encourage mutual learning between the two attention sub-nets, we introduce a semantic distillation loss that aligns each other’s class posterior probabilities. Thus, MSDN distills the intrinsic semantic representations between visual and attribute features for effective knowledge transfer of ZSL. Extensive experiments on three popular benchmarks show the superiority of MSDN. we believe that our work will also facilitate the development of other visual-and-language learning systems, *e.g.*, visual question answering [1].

Acknowledgements

This work is partially supported by NSFC (61772220, 62172177, 62006244), Key R&D Plan of Hubei Province (2020BAB027), NSF of Hubei Province (2021CFB332), YESS Program of China Association for Science and Technology (YESS20200140).

References

- [1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *IJCV*, 123:4–31, 2015. [8](#)
- [2] Zeynep Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *TPAMI*, 38:1425–1438, 2016. [2](#)
- [3] Gundeep Arora, V. Verma, Ashish Mishra, and P. Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, pages 4281–4289, 2018. [2, 3](#)
- [4] Long Chen, Hanwang Zhang, Jun Xiao, W. Liu, and S. Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR*, pages 1043–1052, 2018. [2, 6](#)
- [5] Shiming Chen, Ziming Hong, Yang Liu, Guosen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. Trans-zero: Attribute-guided transformer for zero-shot learning. In *AAAI*, 2022. [2, 3](#)
- [6] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. In *ICCV*, 2021. [2, 3, 6, 7](#)
- [7] Shiming Chen, Guo-Sen Xie, Yang Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. In *NeurIPS*, 2021. [2, 3, 6](#)
- [8] Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, and Zheng Zhang. Semantics disentangling for generalized zero-shot learning. In *ICCV*, 2021. [2](#)
- [9] Yu-Ying Chou, Hsuan-Tien Lin, and Tyng-Luh Liu. Adaptive and generative zero-shot learning. In *ICLR*, 2021. [2](#)
- [10] Andrea Frome, G. S. Corrado, Jonathon Shlens, S. Bengio, J. Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013. [2, 3, 6](#)
- [11] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *CVPR*, 2021. [2](#)
- [12] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [1, 5](#)
- [13] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. [3](#)
- [14] D. Huynh and E. Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, pages 4482–4492, 2020. [2, 5, 6](#)
- [15] Dat T. Huynh and E. Elhamifar. Compositional zero-shot learning via fine-grained dense feature composition. In *NeurIPS*, 2020. [6](#)
- [16] Rohit Keshari, R. Singh, and Mayank Vatsa. Generalized zero-shot learning via over-complete distribution. In *CVPR*, pages 13297–13305, 2020. [3](#)
- [17] Christoph H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009. [1](#)
- [18] Christoph H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 36:453–465, 2014. [1](#)
- [19] H. Larochelle, D. Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, pages 646–651, 2008. [1](#)
- [20] J. Li, Mengmeng Jing, K. Lu, Z. Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *CVPR*, pages 7394–7403, 2019. [3](#)
- [21] K. Li, Martin Renqiang Min, and Yun Fu. Rethinking zero-shot learning: A conditional visual classification perspective. In *ICCV*, pages 3582–3591, 2019. [3](#)
- [22] Y. Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Discriminative learning of latent features for zero-shot recognition. In *CVPR*, pages 7463–7471, 2018. [2](#)
- [23] Shichen Liu, Mingsheng Long, J. Wang, and Michael I. Jordan. Generalized zero-shot learning with deep calibration network. In *NeurIPS*, 2018. [2, 6](#)
- [24] Yang Liu, Jishun Guo, Deng Cai, and X. He. Attribute attention for semantic disambiguation in zero-shot learning. In *ICCV*, pages 6697–6706, 2019. [6](#)
- [25] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and T. Harada. Goal-oriented gaze estimation for zero-shot learning. In *CVPR*, 2021. [2, 3](#)
- [26] L. V. D. Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *JMLR*, 9:2579–2605, 2008. [8](#)
- [27] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Z. Zha, and Yongdong Zhang. Domain-aware visual bias eliminating for generalized zero-shot learning. In *CVPR*, pages 12661–12670, 2020. [2](#)
- [28] Mark Palatucci, D. Pomerleau, Geoffrey E. Hinton, and Tom Michael Mitchell. Zero-shot learning with semantic output codes. In *NeurIPS*, pages 1410–1418, 2009. [1](#)
- [29] Emilio Parisotto, Jimmy Ba, and Ruslan Salakhutdinov. Actor-mimic: Deep multitask and transfer reinforcement learning. In *ICLR*, 2016. [3](#)
- [30] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758, 2012. [2, 5](#)
- [31] Jeffrey Pennington, R. Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. [2, 4](#)
- [32] B. Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. [2](#)
- [33] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. [3](#)
- [34] Edgar Schönfeld, S. Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. In *CVPR*, pages 8239–8247, 2019. [2, 3, 6](#)
- [35] Yuming Shen, J. Qin, and L. Huang. Invertible zero-shot recognition flows. In *ECCV*, 2020. [2, 3](#)
- [36] Jie Song, Chengchao Shen, Yezhou Yang, Y. Liu, and Mingli Song. Transductive unbiased embedding for zero-shot learning. *CVPR*, pages 1024–1033, 2018. [2](#)
- [37] Yao-Hung Hubert Tsai, Liang-Kang Huang, and R. Salakhutdinov. Learning robust visual-semantic embeddings. In *ICCV*, pages 3591–3600, 2017. [2, 3](#)

- [38] M. R. Vyas, Hemanth Venkateswara, and S. Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *ECCV*, 2020. 2, 3
- [39] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *CVPR*, pages 6897–6906, 2020. 1
- [40] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 1
- [41] Q. Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *IJCV*, 124:356–383, 2017. 2
- [42] P. Welinder, S. Branson, T. Mita, C. Wah, Florian Schroff, Serge J. Belongie, and P. Perona. Caltech-ucsd birds 200. *Technical Report CNS-TR-2010-001, Caltech.*, 2010. 2, 5
- [43] Yongqin Xian, T. Lorenz, B. Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018. 2, 3, 6
- [44] Yongqin Xian, B. Schiele, and Zeynep Akata. Zero-shot learning — the good, the bad and the ugly. In *CVPR*, pages 3077–3086, 2017. 1, 2, 5
- [45] Yongqin Xian, Saurabh Sharma, B. Schiele, and Zeynep Akata. F-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, pages 10267–10276, 2019. 2, 6
- [46] Guo-Sen Xie, L. Liu, Xiaobo Jin, F. Zhu, Zheng Zhang, J. Qin, Yazhou Yao, and L. Shao. Attentive region embedding network for zero-shot learning. In *CVPR*, pages 9376–9385, 2019. 2, 3, 4, 6
- [47] Guo-Sen Xie, L. Liu, Xiaobo Jin, F. Zhu, Zheng Zhang, Yazhou Yao, J. Qin, and L. Shao. Region graph embedding network for zero-shot learning. In *ECCV*, 2020. 2, 3, 4, 6
- [48] Wenjia Xu, Yongqin Xian, Jiuniu Wang, B. Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020. 2, 3, 4, 5, 6
- [49] H. Yu and B. Lee. Zero-shot learning via simultaneous generating and learning. In *NeurIPS*, 2019. 6
- [50] Y. Yu, Zhong Ji, J. Han, and Z. Zhang. Episode-based prototype generating network for zero-shot learning. In *CVPR*, pages 14032–14041, 2020. 2, 3, 6
- [51] Zhongqi Yue, Tan Wang, Hanwang Zhang, Qianru Sun, and Xiansheng Hua. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, 2021. 6
- [52] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. In *ECCV*, 2020. 3
- [53] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 3
- [54] Yizhe Zhu, Jianwen Xie, Z. Tang, Xi Peng, and A. Elgammal. Semantic-guided multi-attention localization for zero-shot learning. In *NeurIPS*, 2019. 2, 3, 4, 5, 6