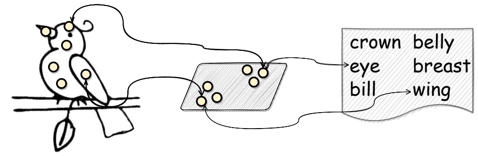
Dual Expert Distillation Network for Generalized Zero-Shot Learning

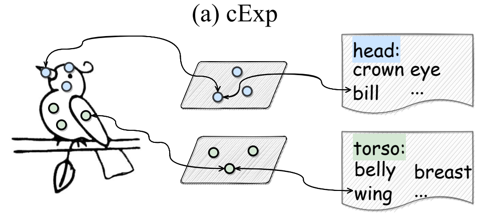
Zhijie Rao[1] , **Jingcai Guo**[1,2*] , **Xiaocheng Lu**[3] , **Jingming Liang**[1,4] , **Jie Zhang**[1] , **Haozhao Wang**[5] , **Kang Wei**[1] and **Xiaofeng Cao**[6]
[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China
[2]Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China
[3]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
[4]Department of Computer Science, The University of Iowa, Iowa City, United States
[5]School of Computer Science and Technology, HUST, Wuhan, China
[6]School of Artificial Intelligence, Jilin University, Changchun, China
jc-jingcai.guo@polyu.edu.hk

Abstract

Zero-shot learning has consistently yielded remarkable progress via modeling nuanced one-to-one visual-attribute correlation. Existing studies resort to refining a uniform mapping function to align and correlate the sample regions and subattributes, ignoring two crucial issues: 1) the inherent asymmetry of attributes; and 2) the unutilized channel information. This paper addresses these issues by introducing a simple yet effective approach, dubbed **D**ual **E**xpert **D**istillation **N**etwork (DEDN), where two experts are dedicated to coarse- and fine-grained visual-attribute modeling, respectively. Concretely, one coarse expert, namely **cExp**, has a complete perceptual scope to coordinate visual-attribute similarity metrics across dimensions, and moreover, another fine expert, namely **fExp**, consists of multiple specialized subnetworks, each corresponds to an exclusive set of attributes. Two experts cooperatively distill from each other to reach a mutual agreement during training. Meanwhile, we further equip DEDN with a newly designed backbone network, i.e., **D**ual **A**ttention **N**etwork (DAN), which incorporates both region and channel attention information to fully exploit and leverage visual semantic knowledge. Extensive experiments on various benchmark datasets indicate a new state-of-the-art. The code is available at github.com/zjrao/DEDN.



(a) cExp



(b) fExp

Figure 1: (a) *cExp*, i.e., the common practice in existing works, possesses complete attribute-awareness capability yet lacks the ability to process fine-grained semantic information. (b) *fExp*, i.e., consists of multiple specialized sub-networks, lacks a global perception field.

1 Introduction

Recognizing unknown categories in the open environment is a critical challenge for automatic recognition systems. Zero-Shot Learning (ZSL) [Lampert *et al.*, 2009] that serves as a promising solution has received increasing attention, which is inspired by human text-to-image reasoning capabilities. The objective of ZSL is to transfer the visual knowledge of seen classes to the unseen domain by virtue of shared semantic information, thus empowering the model to recognize the unseen classes. More trickily, Generalized Zero-Shot Learning (GZSL) [Chao *et al.*, 2016] requires recognizing samples from both seen and unseen classes in the inference phase.

Mainstream studies broadly follow two routes, generative [Xian *et al.*, 2018][Xie *et al.*, 2022][Li *et al.*, 2023] and embedding techniques [Zhang *et al.*, 2017][Liu *et al.*, 2020][Chen *et al.*, 2021b], where most of the schemes are devoted to mining and constructing class-wise visual-attribute relations. To strengthen the fine-grained perceptual capabilities of the model, recent research has invested considerable effort into modeling local-subattribute correlations [Xie *et al.*, 2019][Huynh and Elhamifar, 2020][Xu *et al.*, 2020]. The motivation is to build a refined pairwise relation map via searching and binding subattributes and the corresponding region visual features (Figure 1 (a)). Despite their contribution to boosting performance, the inherent asymmetry of attributes remains undiscussed, and the channel information is not fully exploited.

The asymmetry of attributes stems from the fact that 1) the semantic dimensions between attributes are heterogeneous or

*Corresponding author: *Jingcai Guo*.

even antagonistic. Take the SUN dataset [Patterson and Hays, 2012] as an example, where 38 attributes (*studying, playing, etc.*) describe the function of one scene, while 27 attributes (*trees, flowers, etc.*) describe the entities in the scene. It can be obviously observed that the former are abstract and global, while the latter are concrete and local; 2) the visual features corresponding to attributes are intertwined. For example, neighboring regions tend to be more semantically similar, a phenomenon that is exacerbated by the local information fusion mechanism of the convolutional kernel, which leads to difficulties in accurately locating fine-grained attributes such as *head, crown*, and so on.

In this paper, we revisit the task of modeling visual-attribute relations from the perspective of attribute annotations. Given the inherent complexity of attribute descriptions, existing learning paradigms are virtually forcing a single model to undertake a multi-objective hybrid task, which is ideally appealing yet empirically challenging. Naturally, we employ the idea of divide-and-conquer to release the pressure of a single model. We meticulously decompose the hybrid task into multiple subtasks, i.e., dividing the attributes into multiple disjoint clusters and assigning specialized learnable networks to them. Our approach is referred to as, Dual Expert Distillation Network, abbreviated DEDN. As shown in Figure 1, our approach sets up two experts. *cExp*, in line with common practices, is equipped with complete attribute perception capability to harmonize holistic visual-attribute measure results. *fExp*, consists of multiple subnetworks, where each subnetwork is only responsible for capturing the characteristics of a specific attribute cluster. During the training phase, we encourage the two to learn cooperatively to compensate for their respective deficiencies in a mutually distilling manner. The decision results of the two experts are combined for final inference.

For the issue of underutilized channel information, we design a novel attention network, Dual Attention Network (DAN), as the backbone. DAN employs a dual-attention mechanism that fully exploits the potential semantic knowledge of both regions and channels to facilitate more precise visual-attribute correlation metrics. To further boost performance, we present Margin-Aware Loss (MAL) as the training loss function to address the confidence imbalance between seen and unseen classes.

Our contributions are summarized below:

- We rethink the issue of modeling visual-attribute relations from the perspective of attribute annotations and point out that the inherent complexity of attributes is one of the major bottlenecks. We propose a simple yet effective strategy of establishing two experts working on distinct attribute perception scopes to learn and infer collaboratively in a complementary manner.
- We present a novel attention network, dubbed DAN, which incorporates both region and channel attention information to better capture correlations between visuals and attributes. Furthermore, a new learning function named MAL is designed to balance the confidence of seen and unseen classes.
- We conduct extensive experiments on mainstream evaluation datasets, and the results show that the proposed method effectively improves the performance.

2 Related Work

In ZSL/GZSL, attributes are the only ties that bridge seen and unseen classes, hence exploring and constructing the link between visuals and attributes is a core subject. Existing methods fall into class-wise visual-attribute modeling, which treats both visual features and attribute vectors as a whole, and regional visual-subattribute modeling, which seeks to explore the correlation between local visual information and subattributes.

2.1 Class-Wise Visual-Attribute Modeling

Mainstream researches broadly follow two technical routes, generative and embedding techniques. Generative techniques utilize the latent distribution fitting ability of generative models such as GAN and VAE to implicitly learn the relationship between attributes and categories to construct hallucinatory samples of unseen classes [Xian *et al.*, 2018][Verma *et al.*, 2018][Felix *et al.*, 2018][Li *et al.*, 2019][Vyas *et al.*, 2020][Keshari *et al.*, 2020][Xie *et al.*, 2022][Li *et al.*, 2023]. The technical bottleneck of this route is the poor realism of the hallucinatory samples, thus many studies incorporate other techniques such as meta-learning [Yu *et al.*, 2020], representation learning [Li *et al.*, 2021][Chen *et al.*, 2021c][Chen *et al.*, 2021a][Han *et al.*, 2021][Kong *et al.*, 2022], etc. for joint training. Embedding techniques aim at projecting visual and attribute features to a certain space, from which the most similar semantic information is searched. In general, embedding techniques are categorized into three directions: visual-to-attribute space [Changpinyo *et al.*, 2016][Kodirov *et al.*, 2017][Liu *et al.*, 2020][Chen *et al.*, 2022a], attribute-to-visual space [Zhang *et al.*, 2017][Annadani and Biswas, 2018], and common space [Liu *et al.*, 2018][Jiang *et al.*, 2019]. Researchers in the first two directions invest considerable effort in designing robust mapping functions to cope with domain shift and out-of-distribution generalization problems. The third direction centers on finding a suitable semantic space. Class-level visual-attribute modeling lacks the fine-grained perceptual ability to respond to interactions between local visual features and subattributes.

2.2 Region-Wise Visual-Attribute Modeling

Region-wise modeling is a promising direction in embedding techniques. Unlike other embedding approaches, region-wise modeling focuses on the correlation between local information and subattributes to build more detailed mapping functions. Models based on attention mechanisms are the dominant means in this direction, motivated by training models to search for corresponding visual features based on semantic vectors. Recent approaches include feature-to-attribute attention networks [Xie *et al.*, 2019][Huynh and Elhamifar, 2020], bidirectional attention networks [Chen *et al.*, 2022b], and multi-attention networks [Zhu *et al.*, 2019]. In addition, some studies resort to prototype learning, where the goal is to explicitly learn the corresponding prototypical visual features of individual subattributes, thus aiding the model's judgment

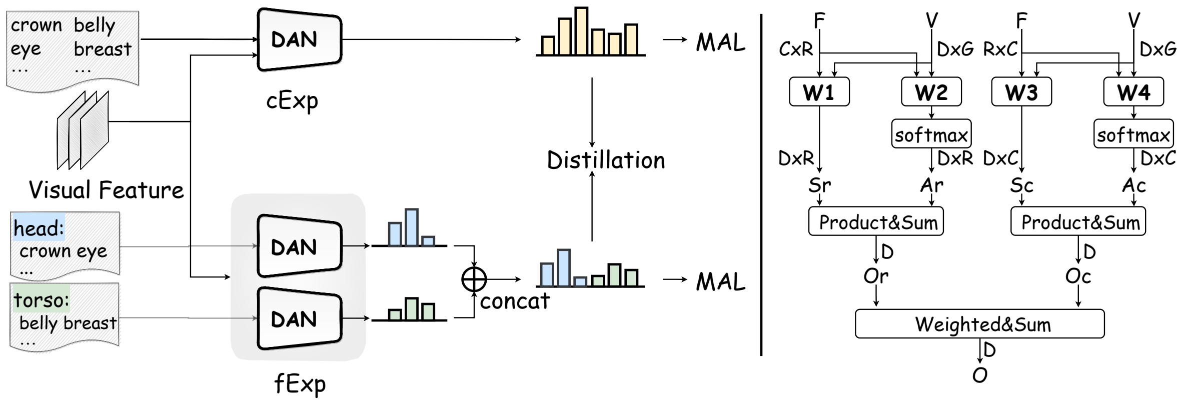


Figure 2: **Left:** *cExp* possesses the scope of a holistic attribute set, while *fExp* consists of multiple sub-networks, each of which is responsible for the prediction of only partial attributes. We concatenate all outputs of subnetworks as the final result of *fExp*. Then, distillation loss is implemented to facilitate joint learning. **Right:** The architecture of DAN.

[Xu *et al.*, 2020][Wang *et al.*, 2021]. Further, modeling the topological structure between regional features with the help of graph convolution techniques also yields promising results [Xie *et al.*, 2020][Guo *et al.*, 2023]. While the main idea of these approaches is to design appropriate attention networks or regularization functions, ignoring the inherent complexity of attribute annotations, we provide a new perspective to think about the visual-attribute modeling problem. In addition, existing region-attribute methods, although achieving good results, neglect the utilization of channel information, and we design a new attention network that utilizes both region and channel information.

3 Methodology

3.1 Preliminary

Following previous studies [Chen *et al.*, 2022b][Li *et al.*, 2023], we adopt a fixed feature extractor, ResNet-101 [He *et al.*, 2016], to extract visual features. Suppose $\mathcal{D}^s = \{(F_i^s, Y_i^s)\}$ denotes the seen classes, where F_i^s is the visual feature and Y_i^s denotes its label. Note that $F \in \mathbb{R}^{C\times H\times W}$, where C, H, W are the channel number, height, and width, respectively. Similarly have $\mathcal{D}^u = \{(F_i^u, Y_i^u)\}$ to denote the unseen classes. Normally, the visual features of the unseen classes are not accessible during the training phase. Alternatively, we have the shared attribute $A \in \mathbb{R}^{K\times D}$, where K denotes the total number of categories, and D denotes the number of attributes. Also, we use the semantic vectors of each attribute learned by GloVe, denoted by $V \in \mathbb{R}^{D\times G}$, where G denotes the dimension of the vector.

3.2 Overview

Our approach is shown in Figure 2 (Left). First, we disassemble the attribute set into multiple clusters based on their characteristics. Then the attribute vectors and the visual feature are fed into *cExp* and *fExp* simultaneously. *cExp* directly computes the scores of all attributes on that visual feature, while the scores of *fExp* are obtained by combining the computation results of each subnetwork. We constrain the two to learn from each other using distillation loss. Meanwhile, we introduce DAN as the backbone and MAL as the optimization objective.

3.3 Dual Attention Network

Firstly we introduce the proposed novel backbone network, Dual Attention Network (DAN). Mining and constructing relations between visual features and attributes is crucial for zero-shot learning. Recently many works have been devoted to modeling the association between regions and attributes, such as attention-based approaches [Xie *et al.*, 2019][Huynh and Elhamifar, 2020][Chen *et al.*, 2022b] and prototype-based techniques [Xu *et al.*, 2020][Wang *et al.*, 2021]. However, these methods only focus on the semantic information of regions and ignore the role of channels. Therefore, DAN incorporates both the attention information of regions and channels to promote the efficacy of the model in utilizing visual features.

As shown in Figure 2 (Right), DAN contains two parallel components that model region-attribute and channel-attribute relations, respectively. We first introduce the region-attribute component. We have visual features $F \in \mathbb{R}^{C\times H\times W}$, which is flattened to $F \in \mathbb{R}^{C\times R}$, where $R = H \times W$ denotes the number of regions. Let $W_1, W_2 \in \mathbb{R}^{G\times C}$ denote two learnable matrices. W_1 maps the attribute vectors to the visual space and computes their similarity. The formula is expressed as:

$$S_r = VW_1F, \tag{1}$$

where $S_r \in \mathbb{R}^{D\times R}$ represents the score obtained for each attribute on each region. W_2 is in charge of computing the attention weights to encourage the model to focus on the region-attribute pairs with the highest similarity. The formula is expressed as:

$$A_r = \frac{VW_2F}{\sum_{r\in R} VW_2F_r}, \tag{2}$$

where $A_r \in \mathbb{R}^{D\times R}$ denote the normalized weight obtained by softmax. Then we naturally get the weighted matrix of scores, represented as:

$$O_r = \sum_R S_r \times A_r, \tag{3}$$

where $O_r \in \mathbb{R}^D$ represents the similarity score obtained for each attribute on a visual feature.

Next, we introduce the channel-attribute section, which has a similar principle. We have the scaled visual feature $F \in \mathbb{R}^{R\times C}$ and $W_3, W_4 \in \mathbb{R}^{G\times R}$. Then W_3 is charged with calculating the similarity score obtained by the attribute on each channel, formulated as:

$$S_c = VW_3F, \tag{4}$$

where $S_c \in \mathbb{R}^{D\times C}$. And W_4 computes its attention weights:

$$A_c = \frac{VW_4F}{\sum_{c\in C} VW_4F_c}, \tag{5}$$

where $A_c \in \mathbb{R}^{D\times C}$. Finally, we get the weighted score map:

$$O_c = \sum_C S_c \times A_c, \tag{6}$$

where $O_c \in \mathbb{R}^D$. We expect the final scores of attributes from different scale features to be consistent, i.e., semantic consistency. Therefore we employ $\mathcal{L}_{align}$, which contains a Jensen-Shannon Divergence (JSD) and a Mean Squared Error, to align the outputs of both, formulated as:

$$\mathcal{L}_{align} = \frac{1}{2}(\mathcal{L}_{KL}(O_r||O_c) + \mathcal{L}_{KL}(O_c||O_r)) + ||O_r - O_c||_2^2, \tag{7}$$

where $\mathcal{L}_{KL}$ denotes Kullback-Leibler Divergence. In the inference phase, we use the weighted sum of O_r and O_c as the final output, expressed as:

$$O = \lambda_{rc} \times O_r + (1 - \lambda_{rc}) \times O_c, \tag{8}$$

where λ_{rc} is a hyperparameter.

3.4 Dual Expert Distillation Network

Despite the fact that DAN enhances the modeling capability of the network, it is extremely challenging for a single model to simultaneously handle attributes with different semantic dimensions as well as visual features with different granularities. To this end, we propose the Dual Expert Distillation Network (DEDN) to alleviate the pressure on a single network (Figure 2 (left)). *cExp* is set up with a complete attribute-aware scope as in conventional practice. Specifically, the input of *cExp* is the semantic vectors of all attributes, and the output is the similarity scores of all attributes. Denote *cExp* by $\phi_{ec} = \{W_1^{ec}, W_2^{ec}, W_3^{ec}, W_4^{ec}\}$, the output is defined as:

$$O_{ec} = \phi_{ec}(V, F), \tag{9}$$

where $O_{ec} \in \mathbb{R}^D$ and $V \in \mathbb{R}^{D\times G}$.

fExp consists of multiple subnetworks, each focusing on a specific attribute cluster. At first, we elaborate on how the attribute clusters are divided. Since attribute annotations are manually labeled based on semantics, they are inherently clustered in nature. For example, in the SUN dataset [Patterson and Hays, 2012], the top 38 prompts are used to describe the scene function. Therefore, it is easy to perform the division by human operation, Chat-GPT [Radford *et al.*, 2018], or clustering algorithm. It requires a trivial amount of effort but is worth it.

Assuming that the attribute set is divided into Q disjoint clusters, i.e. $V = \{V_1 \in \mathbb{R}^{D_1\times G}, V_2 \in \mathbb{R}^{D_2\times G}, ..., V_Q \in \mathbb{R}^{D_Q\times G}\}$, where $D_1 + D_2 + ... + D_Q = D$. Accordingly, there are Q subnetworks for *fExp* to handle these attribute clusters one-to-one. Let $\phi_{ef} = \{\phi_{ef}^1, \phi_{ef}^2, ..., \phi_{ef}^Q\}$ denotes *fExp*, then the output is defined as:

$$O_{ef} = \phi_{ef}^1(V_1, F) \oplus \phi_{ef}^2(V_2, F) \oplus ... \oplus \phi_{ef}^Q(V_Q, F), \tag{10}$$

where $\oplus$ denotes concat operation.

After that, we calculate the score of each category for training and inference. Specifically, we compute the similarity with the output of the expert and the attributes of each category, defined as:

$$P_{ec} = O_{ec}A^\mathsf{T}, P_{ef} = O_{ef}A^\mathsf{T}, \tag{11}$$

where $P_{ec}, P_{ef} \in \mathbb{R}^K$. To facilitate cooperative learning between two expert networks, we introduce distillation loss to constrain their semantic consistency. Concretely, the distillation loss contains a Jensen-Shannon Divergence (JSD) and a Mean Squared Error, defined as:

$$\mathcal{L}_{distill} = \frac{1}{2}(\mathcal{L}_{KL}(P_{ec}||P_{ef}) + \mathcal{L}_{KL}(P_{ef}||P_{ec})) + ||P_{ec} - P_{ef}||_2^2. \tag{12}$$

3.5 Margin-Aware Loss

Once the category scores are obtained, the network is optimized by using the cross-entropy loss, which is formulated as:

$$\mathcal{L}_{ce} = -\log \frac{\exp(P_{ec}^y)}{\sum_{y_i}^K \exp(P_{ec}^{y_i})}, \tag{13}$$

where y is the ground truth. The loss of P_{ef} ditto. Note that we next narrate with P_{ec} only, and the principle is the same for P_{ef}.

Due to the lack of access to samples from the unseen classes during the training phase, the scores of the unseen classes are relatively low and thus cannot compete with the seen classes in GZSL. To address this problem, the common practice [Huynh and Elhamifar, 2020][Chen *et al.*, 2022b] is to add a margin to the scores:

$$PM_{ec} = [P_{ec}^1 - \epsilon, ..., P_{ec}^N - \epsilon, P_{ec}^{N+1} + \epsilon, ..., P_{ec}^K + \epsilon], \tag{14}$$

where ϵ is a constant, $P_{ec}^1 \sim P_{ec}^N$ are seen classes score, and $P_{ec}^{N+1} \sim P_{ec}^K$ are unseen classes score. However, this method leads to misclassification of seen classes that would otherwise be correctly predicted. In order to maintain the correctness of the predicted classes while enhancing the competitiveness of the unseen classes. We propose Margin-Aware Loss (MAL), which takes the form:

$$\mathcal{L}_{mal} = -\log \frac{\exp(P_{ec}^y - 2\epsilon)}{\exp(P_{ec}^y - 2\epsilon) + \sum_{y_i \neq y}^{\mathcal{S}} \exp(P_{ec}^{y_i} + \epsilon) + \sum^{\mathcal{U}} \exp(P_{ec}^{y_i})}, \tag{15}$$

where $\mathcal{S}, \mathcal{U}$ denote seen and unseen classes, respectively. In contrast to the cross-entropy loss, MAL reactivates the confidence of the predicted class to ensure that it stays ahead in the margin-processed scores, while suppressing the confidence of the other seen classes to ensure the competitiveness of the unseen classes.

METHOD	ROUTE	CUB T	CUB U	CUB S	CUB H	SUN T	SUN U	SUN S	SUN H	AWA2 T	AWA2 U	AWA2 S	AWA2 H
f-CLSWGAN	Gen.	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	Gen.	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
TF-VAEGAN	Gen.	64.9	52.8	64.7	58.1	66.0	45.6	**40.7**	43.0	72.2	59.8	75.1	66.6
E-PGN	Gen.	72.4	52.0	61.1	56.2	-	-	-	-	73.4	52.6	83.5	64.6
CADA-VAE	Gen.	59.8	51.6	53.5	52.4	61.7	47.2	35.7	40.6	63.0	55.8	75.0	63.9
FREE	Gen.	-	55.7	59.9	57.7	-	47.4	37.2	41.7	-	60.4	75.4	67.1
SDGZSL	Gen.	75.5	59.9	66.4	63.0	62.4	48.2	36.1	41.3	72.1	64.6	73.6	68.8
CE-GZSL	Gen.	77.5	63.9	66.8	65.3	63.3	48.8	38.6	43.1	70.4	63.1	78.6	70.0
VS-Boost	Gen.	**79.8**	68.0	68.7	68.4	62.4	49.2	37.4	42.5	-	67.9	81.6	**74.1**
SGMA	Emb.†	71.0	36.7	71.3	48.5	-	-	-	-	68.8	37.6	87.1	52.5
AREN	Emb.†	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	Emb.	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	Emb.†	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	Emb.	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
DCN	Emb.	56.2	28.4	60.7	38.7	61.8	25.5	37.0	30.2	65.2	25.5	84.2	39.1
HSVA	Emb.	62.8	52.7	58.3	55.3	63.8	48.6	39.0	43.3	-	59.3	76.6	66.8
MSDN	Emb.†	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3	70.1	62.0	74.5	67.7
DEDN(**Ours**)	Emb.	77.4	**70.9**	70.0	**70.4**	**67.4**	**54.7**	36.0	**43.5**	**75.8**	**68.0**	76.5	72.0

Table 1: Comparison with state-of-the-art methods (%). *Gen.* denotes generative method and *Emb.* denotes embedding method. † denotes the region-attribute modeling method. The best and second-best results are highlighted in **blue** and underlined, respectively.

3.6 Summarize

During training, the basic training loss of *cExp* stems from the classification and the alignment loss and is expressed as:

$$\mathcal{L}_{ec} = \mathcal{L}_{mal}^{ec} + \beta \mathcal{L}_{align}^{ec}, \quad (16)$$

where β is a hyperparameter. Similarly, we have the basic training loss of *fExp*:

$$\mathcal{L}_{ef} = \mathcal{L}_{mal}^{ef} + \beta \mathcal{L}_{align}^{ef}. \quad (17)$$

Then the final loss is obtained from the combination of basic losses and distillation loss, denoted as:

$$\mathcal{L}_{DEDN} = \mathcal{L}_{ec} + \mathcal{L}_{ef} + \gamma \mathcal{L}_{distill}, \quad (18)$$

where γ is a hyperparameter.

In the inference phase, the recommendations of the two experts are combined and used for final judgment. The predicted result is expressed as:

$$\arg\max \lambda_e \times P_{ec} + (1 - \lambda_e) \times P_{ef}, \quad (19)$$

where λ_e is a hyperparameter.

4 Experiments

Datasets. We conduct extensive experiments on three benchmark datasets to verify the effectiveness of the method, including CUB (Caltech UCSD Birds 200) [Wah *et al.*, 2011], SUN (SUN Attribute) [Patterson and Hays, 2012], and AWA2 (Animals with Attributes 2) [Xian *et al.*, 2017]. We split all datasets following [Xian *et al.*, 2017]. CUB comprises 200 bird species totaling 11,788 image samples, of which 50 categories are planned as unseen classes. We use class attributes for fair comparison, which contain 312 subattributes. SUN has a sample of 717 different scenes totaling 14,340 images, where 72 categories are unseen classes. Attribute annotations are 102-dimensional. AWA2 includes 50 classes of assorted animals totaling 37,322 samples, of which 10 categories are considered unseen classes. Its number of attributes is 85.

Evaluation Protocols. We perform experiments in both the Zero-Shot learning (ZSL) and Generalized Zero-Shot learning (GZSL) settings. For ZSL, we employ top-1 accuracy to evaluate the performance of the model, denoted as **T**. For GZSL, we record the accuracy for both seen classes, and unseen classes, denoted as **S**, and **U**, respectively. We also record the harmonic mean **H**, which is computed as, $\mathrm{H} = (2 \times \mathrm{S} \times \mathrm{U})/(\mathrm{S} + \mathrm{U})$.

Implementation Details. For a fair comparison, we use the fixed ResNet-101 [He *et al.*, 2016] without finetune as the feature extractor. We set the batch size to 50 and the learning rate to 0.0001. The RMSProp optimizer with the momentum set as 0.9 and weight decay set as 1e-4 is employed. For hyperparameters, $[\beta, \gamma]$ are fixed to [0.001, 0.1]. We empirically set $[\lambda_{rc}, \lambda_e]$ to [0.8, 0.9] for CUB, [0.95, 0.3] for SUN, [0.8, 0.5] for AWA2. Subsequent experimental analyses show that the performance of our method has **low sensitivity** to hyperparameters. For attribute clusters, we classify attribute sets according to their characteristics (Table 2).

CUB #Des.	CUB #Num.	SUN #Des.	SUN #Num.	AWA2 #Des.	AWA2 #Num.
head	112	function	38	texture	18
torso	87	instance	27	organ	14
wing	24	environ.	17	environ.	13
tail	40	light	20	abstract	40
leg	15				
whole	34				

Table 2: Manual division of attribute clusters. *Des.* (description) indicates the criteria for classification. *Num.* (number) is the size of the attribute cluster. *environ*: environment.

4.1 Compared With State-of-the-Arts

To evaluate the performance of the proposed method, we compare it with the state-of-the-art various methods. Generative methods: f-CLSWGAN (CVPR ′18) [Xian *et al.*, 2018], f-VAEGAN-D2 (CVPR ′19) [Xian *et al.*, 2019], TF-VAEGAN (ECCV ′20) [Narayan *et al.*, 2020], E-PGN

SETTING	CUB T	CUB U	CUB S	CUB H	SUN T	SUN U	SUN S	SUN H	AWA2 T	AWA2 U	AWA2 S	AWA2 H
cExp w/o $\mathcal{L}_{distill}$	74.6	62.4	71.4	66.6	64.0	41.6	35.7	38.4	71.1	62.8	78.8	69.9
fExp w/o $\mathcal{L}_{distill}$	75.5	68.1	67.9	68.0	64.0	42.8	35.5	38.7	71.1	62.9	79.1	70.1
DEDN w/o $\mathcal{L}_{distill}$	75.7	66.7	70.7	68.6	65.2	47.3	35.0	40.3	72.1	63.8	79.3	70.7
DAN w/o CA$*$	77.0	58.7	73.6	65.3	65.8	48.5	34.6	40.4	74.6	61.7	**79.8**	69.6
DEDN w/o $\mathcal{L}_{mal}$	75.8	**73.2**	62.5	67.4	66.0	**56.5**	34.3	42.7	73.1	66.5	72.4	69.3
DAN w/o $\mathcal{L}_{align}$	**77.6**	63.3	**72.8**	67.7	65.5	47.5	35.3	40.5	74.6	64.8	76.8	70.3
DEDN(full)	77.4	70.9	70.0	**70.4**	**67.4**	54.7	**36.0**	**43.5**	**75.8**	**68.0**	76.5	**72.0**

Table 3: Ablation Study (%). *w/o* denotes remove the module. *CA$*$* denotes channel attention. The best result is highlighted in **bold**.

(CVPR ′20) [Yu *et al.*, 2020], CADA-VAE (CVPR ′19) [Schonfeld *et al.*, 2019], FREE (ICCV ′21) [Chen *et al.*, 2021a], SDGZSL (ICCV ′21) [Chen *et al.*, 2021c], CE-GZSL (CVPR ′21) [Han *et al.*, 2021], VS-Boost (IJCAI ′23) [Li *et al.*, 2023]; Embedding methos: LFGAA (ICCV ′19) [Liu *et al.*, 2019], APN (NeurIPS ′20) [Xu *et al.*, 2020], DCN (NeurIPS ′18) [Liu *et al.*, 2018], HSVA (NeurIPS ′21) [Chen *et al.*, 2021b]; Region-Attribute modeling: SGMA (NeurIPS ′19) [Zhu *et al.*, 2019], AREN (CVPR ′19) [Xie *et al.*, 2019], DAZLE (CVPR ′20) [Huynh and Elhamifar, 2020], MSDN (CVPR ′22) [Chen *et al.*, 2022b].

The experimental results are shown in Table 1. Our method achieves the best performance in seven metrics and second place in one metric. For Generalized Zero-Shot Learning (GZSL), we beat VS-Boost by 2% in the H-score of CUB, a fine-grained bird dataset whose attribute annotations possess explicit correspondences to visual features. It demonstrates the superiority of the proposed method for fine-grained modeling. On the SUN and AWA2 datasets, we obtain the best and second-best results in H-score, respectively. These two datasets have fewer attributes and contain complex semantic dimensions, including abstract, concrete, etc. The experimental results demonstrate the effectiveness of the proposed method in deconstructing complex tasks to alleviate the modeling pressure of a single network. In addition, the U-scores of our method on all three datasets are well ahead of the others, demonstrating that the proposed method effectively captures the relationship between attributes and visuals to generalize to unseen classes.

For Zero-Shot Learning (ZSL), we achieve the highest top-1 accuracy on the SUN and AWA2 datasets, as well as competitive performance on CUB. Specifically, our method outperforms TF-VAEGAN by 1.4% on the SUN dataset. On AWA2, we have a 2.4% lead relative to the second-place E-PGN. The experimental results validate the superiority of the proposed method. Notably, our method achieves far better results than existing region-attribute modeling methods in both ZSL and GZSL settings, which implies the potential of attribute intrinsic asymmetry and channel information is not fully exploited.

4.2 Ablation Study

To evaluate the role of each module, we perform a series of ablation experiments. The results of the experiments are shown in Table 3. Comprehensively, removing any of the modules leads to different degrees of performance degradation, verifying the rationality and necessity of the design of each module. Concretely, it is observed that the performance of *cExp* is slightly lower than that of *fExp* without the distillation loss constraint, which indicates the potential research value of the inherent asymmetry of the attributes. Meanwhile, without distillation, the performance of DEDN is higher than both *cExp* and *fExp*, demonstrating the complementary properties of the dual experts. In addition, it is worth noting that DAN removing the channel attention results in a substantial performance degradation, demonstrating the importance of channel information. Moreover, the role of $\mathcal{L}_{mal}$ in balancing the confidence of unseen and seen classes can be observed from the metrics **U** and **S**. When $\mathcal{L}_{mal}$ is removed, the metric **U** increases dramatically while **S** decreases dramatically. Finally, the results also demonstrate the importance of $\mathcal{L}_{align}$ for constraining semantic consistency.

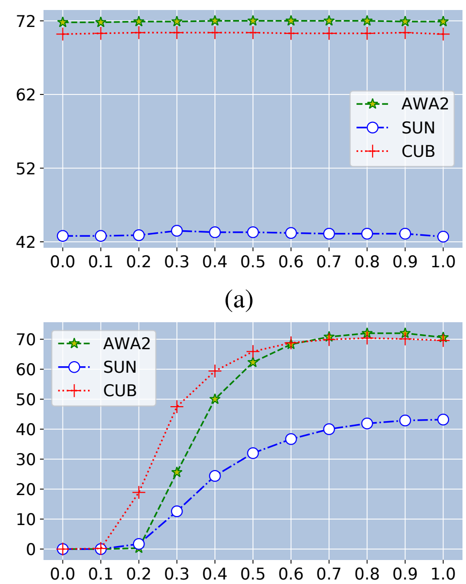


Figure 3: (a) Sensitivity to λ_e. (b) Sensitivity to λ_{rc}. The harmonic mean (H) is reported. (better viewer in color).

4.3 Empirical Analysis

The Influence of Parameters λ_e and λ_{rc}

We launch a series of empirical analyses, including evaluating the impact of parameters λ_e and λ_{rc} on the final performance. Figure 3 (a) illustrates the sensitivity of the harmonic mean for each dataset with respect to parameter λ_e. It can be observed that the influence of parameter a is extremely small.

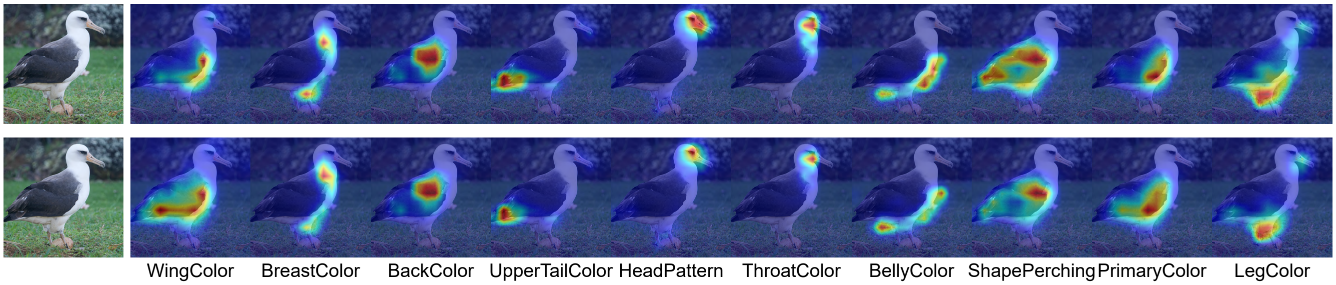


Figure 4: Visualization of the attention heat maps. The first row represents the heat maps of *cExp*, and the second row denotes the heat maps of *fExp* (better viewer in color).

Of particular note, when λ_e is set to 1 or 0, it indicates that only the *cExp* or *fExp* after distillation learning is used for the inference phase. It implies that by mutual distillation learning, each of the two experts learns the strengths of the other, thereby reaching an agreement. Figure 3 (b) illustrates the impact of λ_{rc}. It can be seen that setting λ_{rc} above 0.7 stabilizes the performance. Optimization is achieved when it is set between 0.7 and 0.9.

The Influence of Different Clustering Algorithms

We further evaluate the impact of the clustering algorithm on performance. In Table 2, we have explained that attribute clusters are obtained by humans to classify the attribute sets based on their characteristics. In this subsection, we use the K-Means algorithm for attribute clustering as a comparison to evaluate the performance. The experimental results are shown in Figure 5 (a), where the harmonic mean (H) and top-1 accuracy (T) are reported. From the figure, it can be seen that the K-Means algorithm is slightly poorer compared to human classification, but a good result is also achieved. It again shows that the idea of dividing the attribute set into different clusters holds great promise.

The Influence of the Number of Attribute Clusters

We evaluate the impact of the number of attribute clusters on performance. The attributes of CUB, SUN, and AWA2 are classified into 6, 4, and 4 categories, respectively (Table 2). In this subsection, we halve the categories, i.e., the numbers of attribute clusters for CUB, SUN, and AWA2 are 3, 2, and 2. The experimental results are shown in Figure 5 (b), where *half* denotes that the cluster number is halved. We can see that *half* leads to a reduction of **H** by 0.6%, 1.0%, and 6.8%, respectively, and a reduction of **T** by 0.7%, 0.2%, and 11%, respectively. The results show that detailed attribute classification facilitates the model in capturing more fine-grained information and thus improves the performance.

Visual Analysis of Attention

We perform a visual analysis of the attention of the two experts, and the schematic is shown in Figure 4. It can be observed that *cExp* has a better localization for some global attributes, such as *HeadPatternMaler, BellyColorGrey, ShapePerchingLike*. Meanwhile, *fExp* has more detailed and precise localization for some local attributes, such as *UpperTailColorGrey, ThroatColorGrey, LegColorWhite*. The two experts collaborate and learn in a complementary way to improve together, which leads to better performance.

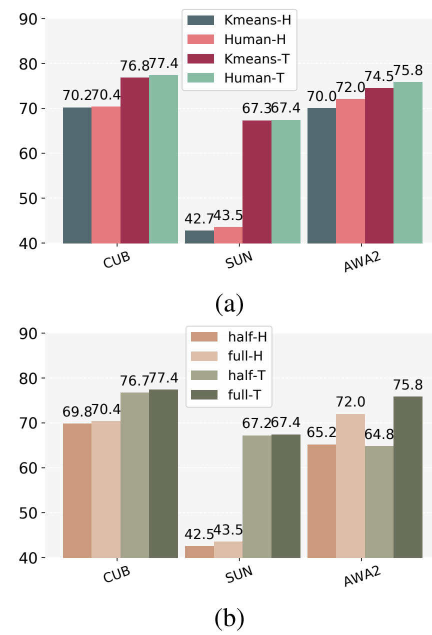


Figure 5: (a) Comparison with Kmeans. (b) Impact of the number of attribute clusters. The harmonic mean (H) and top-1 accuracy (T) are reported (better viewer in color).

5 Conclusion

In this paper, we analyze that the intrinsic asymmetry of attributes is one of the important bottlenecks constraining existing generalized zero-shot learning approaches and propose a simple yet effective framework named DEDN to address this issue. DEDN consists of two expert networks, one with complete attribute-domain perception to harmonize the global correlation confidence and the other consisting of multiple subnetworks, each focusing on a specific attribute domain to capture fine-grained association information. Meanwhile, we introduce DAN as a strong backbone, a novel attention network that incorporates both region and channel knowledge. Moreover, we present a new loss named MAL to train the network. Numerous experiments demonstrate the significant superiority of the proposed approach.

Acknowledgments

Jingcai Guo and Rao Zhijie contributed equally to this work. This research was supported by funding from the Hong Kong RGC General Research Fund (No. 152211/23E and 15216424/24E), the National Natural Science Foundation of China (No. 62102327), and PolyU Internal Fund (No. P0043932). It was also supported by the NVIDIA AI Technology Center (NVAITC).

References

[Annadani and Biswas, 2018] Yashas Annadani and Soma Biswas. Preserving semantic relations for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7603–7612, 2018.

[Changpinyo *et al.*, 2016] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336, 2016.

[Chao *et al.*, 2016] 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 *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 52–68. Springer, 2016.

[Chen *et al.*, 2021a] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 122–131, 2021.

[Chen *et al.*, 2021b] Shiming Chen, Guosen Xie, Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Advances in Neural Information Processing Systems*, 34:16622–16634, 2021.

[Chen *et al.*, 2021c] Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, and Zheng Zhang. Semantics disentangling for generalized zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8712–8720, 2021.

[Chen *et al.*, 2022a] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. Transzero: Attribute-guided transformer for zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 330–338, 2022.

[Chen *et al.*, 2022b] Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao, and Xinge You. Msdn: Mutually semantic distillation network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7612–7621, 2022.

[Felix *et al.*, 2018] Rafael Felix, Ian Reid, Gustavo Carneiro, et al. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision*, pages 21–37, 2018.

[Guo *et al.*, 2023] Jingcai Guo, Song Guo, Qihua Zhou, Ziming Liu, Xiaocheng Lu, and Fushuo Huo. Graph knows unknowns: Reformulate zero-shot learning as sample-level graph recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7775–7783, 2023.

[Han *et al.*, 2021] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2371–2381, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Huynh and Elhamifar, 2020] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4483–4493, 2020.

[Jiang *et al.*, 2019] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Transferable contrastive network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9765–9774, 2019.

[Keshari *et al.*, 2020] Rohit Keshari, Richa Singh, and Mayank Vatsa. Generalized zero-shot learning via over-complete distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13300–13308, 2020.

[Kodirov *et al.*, 2017] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3174–3183, 2017.

[Kong *et al.*, 2022] Xia Kong, Zuodong Gao, Xiaofan Li, Ming Hong, Jun Liu, Chengjie Wang, Yuan Xie, and Yanyun Qu. En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9306–9315, 2022.

[Lampert *et al.*, 2009] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009.

[Li *et al.*, 2019] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7402–7411, 2019.

[Li *et al.*, 2021] Xiangyu Li, Zhe Xu, Kun Wei, and Cheng Deng. Generalized zero-shot learning via disentangled representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1966–1974, 2021.

[Li *et al.*, 2023] Xiaofan Li, Yachao Zhang, Shiran Bian, Yanyun Qu, Yuan Xie, Zhongchao Shi, and Jianping Fan. Vs-boost: Boosting visual-semantic association for generalized zero-shot learning. In *International Joint Conference on Artificial Intelligence*, 2023.

[Liu *et al.*, 2018] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration network. *Advances in Neural Information Processing Systems*, 31, 2018.

[Liu *et al.*, 2019] Yang Liu, Jishun Guo, Deng Cai, and Xiaofei He. Attribute attention for semantic disambiguation in zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6698–6707, 2019.

[Liu *et al.*, 2020] Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Attribute propagation network for graph zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4868–4875, 2020.

[Narayan *et al.*, 2020] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 479–495. Springer, 2020.

[Patterson and Hays, 2012] 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. IEEE, 2012.

[Radford *et al.*, 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[Schonfeld *et al.*, 2019] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019.

[Verma *et al.*, 2018] 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*, pages 4281–4289, 2018.

[Vyas *et al.*, 2020] Maunil R Vyas, Hemanth Venkateswara, and Sethuraman Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 70–86. Springer, 2020.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[Wang *et al.*, 2021] Chaoqun Wang, Shaobo Min, Xuejin Chen, Xiaoyan Sun, and Houqiang Li. Dual progressive prototype network for generalized zero-shot learning. *Advances in Neural Information Processing Systems*, 34:2936–2948, 2021.

[Xian *et al.*, 2017] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017.

[Xian *et al.*, 2018] 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.

[Xian *et al.*, 2019] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10275–10284, 2019.

[Xie *et al.*, 2019] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9384–9393, 2019.

[Xie *et al.*, 2020] Guo-Sen Xie, Li Liu, Fan Zhu, Fang Zhao, Zheng Zhang, Yazhou Yao, Jie Qin, and Ling Shao. Region graph embedding network for zero-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 562–580. Springer, 2020.

[Xie *et al.*, 2022] Guo-Sen Xie, Xu-Yao Zhang, Tian-Zhu Xiang, Fang Zhao, Zheng Zhang, Ling Shao, and Xuelong Li. Leveraging balanced semantic embedding for generative zero-shot learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Xu *et al.*, 2020] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33:21969–21980, 2020.

[Yu *et al.*, 2020] Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang. Episode-based prototype generating network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14035–14044, 2020.

[Zhang *et al.*, 2017] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2021–2030, 2017.

[Zhu *et al.*, 2019] Yizhe Zhu, Jianwen Xie, Zhiqiang Tang, Xi Peng, and Ahmed Elgammal. Semantic-guided multi-attention localization for zero-shot learning. *Advances in Neural Information Processing Systems*, 32, 2019.