Deconstructed Generation-Based Zero-Shot Model

Dubing Chen[1], **Yuming Shen**[2], **Haofeng Zhang**[1*], **Philip H.S. Torr**[2]

[1] Nanjing University of Science and Technology
[2] University of Oxford
{db.chen, zhanghf}@njust.edu.cn, ymcidence@gmail.com, philip.torr@eng.ox.ac.uk

Abstract

Recent research on Generalized Zero-Shot Learning (GZSL) has focused primarily on generation-based methods. However, current literature has overlooked the fundamental principles of these methods and has made limited progress in a complex manner. In this paper, we aim to deconstruct the generator-classifier framework and provide guidance for its improvement and extension. We begin by breaking down the generator-learned unseen class distribution into class-level and instance-level distributions. Through our analysis of the role of these two types of distributions in solving the GZSL problem, we generalize the focus of the generation-based approach, emphasizing the importance of (i) attribute generalization in generator learning and (ii) independent classifier learning with partially biased data. We present a simple method based on this analysis that outperforms SotAs on four public GZSL datasets, demonstrating the validity of our deconstruction. Furthermore, our proposed method remains effective even without a generative model, representing a step towards simplifying the generator-classifier structure. Our code is available at https://github.com/cdb342/DGZ.

1 Introduction

Big data fuels the progress of deep learning, but obtaining specific data can sometimes prove difficult. In cases where specific data is not available, Zero-Shot Learning (ZSL) (Palatucci et al. 2009) can be used to recognize unseen data by utilizing the relationship between seen and unseen data. In general, ZSL seeks to recognize unseen data by exploiting the correlation between seen and unseen data. This correlation is established using semantic knowledge, which can be obtained through human annotations (Lampert, Nickisch, and Harmeling 2009) or word-to-vector approaches (Mikolov et al. 2013a). By using semantic descriptors, ZSL enables the transfer of information from seen to unseen domains. Generalized Zero-Shot Learning (GZSL) (Chao et al. 2016) expands on ZSL by including additional seen classes in the target decision domain, and it has received increasing attention from researchers.

Recently, generative models have been used in mainstream GZSL research to supplement information on unseen classes. A central hypothesis of generation-based GZSL methods is that the generated class-level and instance-level unseen distribution should match the real unseen distribution (Fig. 1). By generating pseudo-unseen instances, these methods enable classifier training to encompass unseen classes, resulting in a superior discrimination of unseen classes compared to their counterparts. Despite their success in enhancing GZSL performance, generation-based methods encounter various challenges in future extensions or developments. Firstly, the underlying reasons for the effectiveness of these approaches remain largely unexplored. Although certain literature suggests that improved discrimination (Wu et al. 2020) or diversity (Liu et al. 2021a) of generated samples contributes to enhanced GZSL performance, no theoretical or empirical evidence supports these performance gains. Secondly, training a generative model entails additional computational and complexity. In most generation-based methods, the primary time complexity arises from training the generative model.

To address these challenges, we conduct both an empirical and a theoretical investigation to uncover, understand, and extend generation-based methods. We begin by analyzing the role of instance-level distribution and class-level distribution. In doing so, we replace the generator-learned instance-level distribution with the Gaussian distribution and conclude its substitutability in improving GZSL performance. (Sec. 3.1). By decomposing the gradient of the cross-entropy loss, we further relate class- and instance-level distributions to unseen class discrimination and decision boundary formation (Sec. 3.2). Based on our analysis, we point out the core improvement direction for the generator-classifier framework. First, the key for the ZSL generator is attribute generalization, where we should focus on generalizing the attribute-conditioned image distribution learned from the seen data to unseen classes. Second, classifier learning is an independent task to learn from partially biased data. We summarize two principles for this task: mitigating the impact of pseudo samples on seen class boundaries during training and reducing the seen-unseen bias.

We finally propose a single baseline based on the idea of deconstruction. Our approach surpasses existing methods in performance, despite having lower complexity. Additionally, we replace the generative model with a one-to-one mapping network from attributes to the visual class centers. Our without-generator method retains most of the performance, which is a step towards simplifying the generator-classifier

*Corresponding author

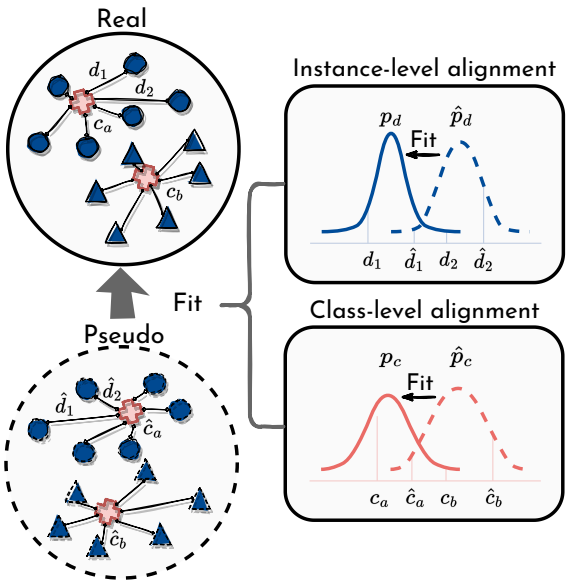

Figure 1: Illustration of two types of distributions learned by a generator: instance-level and class-level. c represents the potential class center, while d denotes an off-center position.

framework. Our main contributions include:

- We deconstruct the generator-classifier framework, using empirical and theoretical analysis to expose the core components of generator and classifier learning.
- We provide a guideline for optimizing the generator-classifier GZSL framework based on our deconstruction idea, which we use to derive a simple method.
- Without a complicated framework design, the proposed method achieves SotAs on four popular ZSL benchmark datasets. Additionally, our method can also be transferred to other generative methods, even a single attribute-vision center mapping net, bringing us closer to a streamlined generator-classifier framework.

2 Related Work

Zero-Shot Learning (ZSL) (Lampert, Nickisch, and Harmeling 2009; Farhadi et al. 2009) has been extensively studied in recent years, which requires knowledge transfer with the class-level edge information, *e.g.*, human-defined attributes (Farhadi et al. 2009; Parikh and Grauman 2011; Akata et al. 2015) and word vectors (Mikolov et al. 2013a,b). Traditional ZSL models (Akata et al. 2013; Frome et al. 2013) typically project the attribute and the visual feature to a common space. Lampert, Nickisch, and Harmeling (2013); Frome et al. (2013); Elhoseiny, Saleh, and Elgammal (2013) choose the attribute space as the common space. Some research afterward (Zhang, Xiang, and Gong 2017; Li, Min, and Fu 2019; Skorokhodov and Elhoseiny 2021) also embed attributes to visual space, or embed attributes and visual features to another space (Akata et al. 2015; Zhang and Saligrama 2015). These methods achieve good performance in the classic ZSL setting but meet a **seen-unseen bias problem** (*i.e.*, prediction results are biased towards seen classes) in **Generalized Zero-Shot Learning (GZSL)** (Chao et al. 2016; Xian, Schiele, and Akata 2017) which emphasizes seen-unseen discrimination.

Driven by the new technology in deep learning, some research enables deeper attribute-visual association with attribute attention (Zhu et al. 2019; Huynh and Elhamifar 2020; Xu et al. 2020; Liu et al. 2021c; Wang et al. 2021). Other methods introduce the out-of-distribution discrimination (Atzmon and Chechik 2019; Min et al. 2020; Chou, Lin, and Liu 2021), which decomposes the GZSL task into seen-unseen discrimination and inter-seen (or -unseen) discrimination. The most successful methods in GZSL build on the recent advent of generative models (Goodfellow et al. 2014; Kingma and Welling 2013), which have dominated recent ZSL research. The **generation-based methods** (Xian et al. 2018, 2019; Chen et al. 2022a) construct pseudo unseen samples to constrain the decision boundary, which form a better seen-unseen discrimination than their counterparts.

A large amount of literature aims at improving the generation-based framework. (Xian et al. 2019; Shen et al. 2020) focus their attention on new generative frameworks. (Verma, Brahma, and Rai 2020) explores the training method. These methods do not make full use of the prior information in the ZSL setting but seek breakthroughs from other fields. (Narayan et al. 2020) design a recurrent structure that utilizes the intermediate layers of the visual-to-attribute mapping network for a second generation. (Han, Fu, and Yang 2020; Han et al. 2021; Chen et al. 2021a,b; Kong et al. 2022) propose to transform the visual feature into an attribute-dependent space, the pseudo unseen samples generated in which contain less seen class bias information. The above-mentioned methods usually adopt complex strategies, which trade large time consumption for performance. In this paper, we explore the nature of the generation-based framework, surpassing current SotAs without complex design.

3 Generation-Based ZSL: A Deconstruction

Assume there are two disjoint class label sets $\mathcal{Y}^s$ and $\mathcal{Y}^u$ ($\mathcal{Y} = \mathcal{Y}^s \cup Y^u$), ZSL aims at recognizing samples belong to $\mathcal{Y}^u$ while only having access to samples with the labels in $\mathcal{Y}^s$ during training. Denote $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ as visual space and attribute space, respectively, where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{a} \in \mathcal{A}$ represent feature instances and their corresponding attributes (represented as column vectors) with dimensions d_x and d_a. Given the training set $\mathcal{D}^s = \{\mathbf{x}, y, \mathbf{a}_y | \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}^s, \mathbf{a}_y \in \mathcal{A}\}$, the goal of ZSL is to learn a classifier towards the unseen classes: $f_{zsl} : \mathcal{X} \rightarrow \mathcal{Y}^u$. GZSL extends this to classify samples belonging to either seen or unseen classes, *i.e.*, $f_{gzsl} : \mathcal{X} \rightarrow \mathcal{Y}$. We mainly discuss the challenges in the GZSL setting in this work.

In this paper, we focus on deconstructing the generator-classifier ZSL framework by understanding the behavior of the generator and the classifier. The framework involves training a conditional generator using visual-attribute pairs, followed by generating pseudo unseen samples using attributes from unseen classes. Finally, the ZSL or GZSL classifier is trained using the generated samples.

Method	DIST	T_1	A^u	A^s	H	CMMD
f-CLSWGAN	GEN	69.0	57.8	71.1	63.8	**0.0337**
	SVG	68.2	55.3	71.7	62.5	0.0341
	LVG	**69.7**	62.8	76.3	**68.9**	0.2523
	SCG	69.5	62.8	68.5	65.5	0.0339
CE-GZSL	GEN	69.8	63.5	77.5	69.8	**0.0071**
	SVG	69.4	60.1	78.2	68.0	0.0099
	LVG	66.0	47.9	72.7	57.7	0.2541
	SCG	**70.6**	63.2	78.9	**70.2**	**0.0071**

Table 1: Zero-Shot performance and CMMD *w.r.t.* different pseudo unseen distributions (DIST). GEN: Generated distribution; SVG: Small-variance Gaussian distribution; LVG: Large-variance Gaussian distribution; SCG: Statistical-covariance Gaussian distribution.

3.1 Empirical Analysis of the Generator-Learned Instance-Level Distribution

In generation-based methods, the generator is often relied upon to produce distributions for unseen classes. To better analyze the ZSL generator, we divide this distribution into two parts, as illustrated in Fig. 1: the *class-level distribution*, which determines how various unseen attributes are mapped to fit the real inter-class distribution in visual space, and the *instance-level distribution*, which deals with how generated samples of the same unseen attribute fit the real intra-class distribution. As the class-level distribution is fundamental to inter-class discrimination, our analysis will concentrate on exploring the generator-fitted instance-level distribution. Specifically, we will compare it to other human-defined distributions based on fitness (against real distribution) and Zero-Shot performance.

Setup. We conduct a comparison between the generator-fitted instance-level unseen distribution and three Gaussian distributions, which have independent small variance, independent large variance, and data-statistical covariance. Since typical zero-shot learning (ZSL) generators usually generate centralized distributions, we replace the instance-level distribution by shifting the centers of other distributions to the generated class centers. We then evaluate the Zero-Shot performance of these distributions and their discrepancy against real unseen distributions. The discrepancy is measured with Maximum Mean Discrepancy (MMD), which is a typical sample-based discrepancy measurement in research of domain adaptation (Long et al. 2015) and generative models (Tolstikhin et al. 2017). We calculate the MMD between the test unseen data and the experimental data for each class, and then take the average value to obtain the Centered MMD (CMMD) score:

$$\begin{aligned}\text{CMMD} = &\frac{1}{|\mathcal{Y}^u|}\sum_{c=1}^{|\mathcal{Y}^u|}\{\frac{1}{n_c\left(n_c-1\right)}\sum_{i,j=1,i\neq j}^{n_c}[\kappa\left(x_i^c,x_j^c\right)\\ &+\kappa\left(\widetilde{x}_i^c,\widetilde{x}_j^c\right)]-\frac{2}{n_c^2}\sum_{i,j=1}^{n_c}\kappa\left(x_i^c,\widetilde{x}_j^c\right)\},\end{aligned} \tag{1}$$

where x_i^c and $\widetilde{x}_i^c$ represent samples from class c in the test unseen and pseudo unseen sets, respectively. n_c denotes the sample number in class c, and $\kappa\left(\cdot,\cdot\right)$ is generally an arbitrary positive-definite reproducing kernel function. Note that the test data involved here is only for measuring the distribution discrepancy and is not used in training.

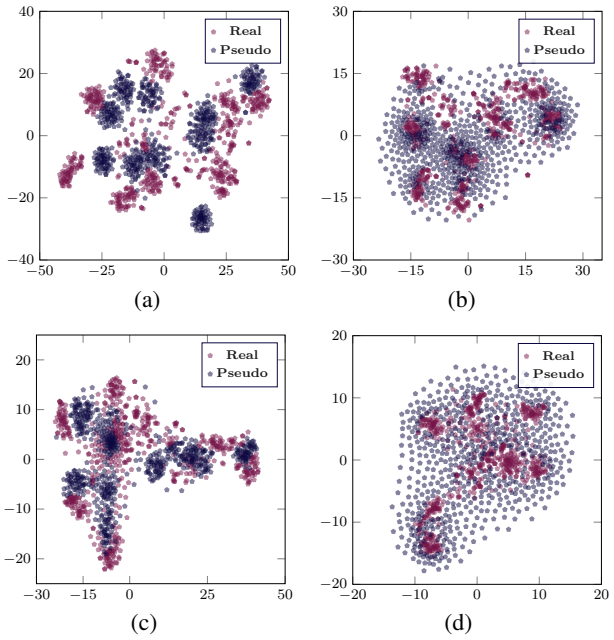

Figure 2: t-SNE comparison across various pseudo unseen distributions. (a) Generated with f-CLSWGAN; (b) Large-variance Gaussian distribution moved to the class center generated with f-CLSWGAN; (c) Generated with CE-GZSL; (d) Large-variance Gaussian distribution moved to the class center generated with CE-GZSL.

Results. We experiment with two classic generation-based methods, f-CLSWGAN (Xian et al. 2018) and CE-GZSL (Han et al. 2021), on AWA2 dataset (Lampert, Nickisch, and Harmeling 2013). The results presented in Tab. 1 led us to two main observations: (i) Gaussian distribution with statistical covariance produces similar results to the generated distribution in both methods; and (ii) the unrealistic unseen distribution negatively affects the performance of CE-GZSL but improves the performance of f-CLSWGAN. These observations prompted us to explore two questions: *(i) Can we generate only the class center instead of using a complex generative model? (ii) How does the large-variance Gaussian distribution affect Zero-Shot performance?* We answer the first question experimentally in Sec. 5, demonstrating that generating only class centers can still achieve reasonable Zero-Shot performance. To address the second question, we further investigate the role of pseudo unseen class samples in classifier training from a gradient perspective.

3.2 Impact of Pseudo Unseen Samples on Classifier Learning

We consider a linear classifier with weight parameters $\mathbf{W} \in \mathbb{R}^{|\mathcal{Y}|\times d_{\mathbf{x}}}$. With a slight abuse of notation, we subsequently use $(\mathbf{x}, y)$ to denote both real and generated data. In the

generation-based framework, the classifier is commonly trained using cross-entropy loss:

$$\mathcal{L}_{ce} = \frac{1}{n}\sum_{c=1}^{|\mathcal{Y}|}\sum_{i=1}^{n_c} -\log p_y(\mathbf{x}_i)$$
$$p_y(\mathbf{x}_i) = \frac{\exp(\langle \mathbf{W}_y, \mathbf{x}_i\rangle/\tau)}{\sum_{c=1}^{|\mathcal{Y}|}\exp(\langle \mathbf{W}_c, \mathbf{x}_i\rangle/\tau)}. \tag{2}$$

Here, $\langle\cdot,\cdot\rangle$ denotes the dot product, c is the index of the c-th row in $\mathbf{W}$, n is the total number of samples, n_c is the sample size in class c, and τ is the temperature parameter (Hinton, Vinyals, and Dean 2015).

Proposition 3.1. *Gradients of $\mathcal{L}_{ce}$ can be decomposed into two components that indicate moving towards the class center and constraining the decision boundary, respectively:*

$$-\frac{\partial_{\mathcal{L}_{ce}}}{\partial_{\mathbf{W}_k}} = \begin{cases} \frac{1}{n\tau}\sum_{i=1}^{n_k}\mathbf{x}_i \\ -\frac{1}{n\tau}\sum_{c=1}^{|\mathcal{Y}|}\sum_{j=1}^{n_c} p_k(\mathbf{x}_j)\mathbf{x}_j \end{cases}, \tag{3}$$

where $p_k(\cdot)$ has an analogous definition to Eq. (2), and $\mathbf{W}_k$ represents the classifier weight of the kth class.

The proofs of Proposition 3.1 is given in the appendix. According to Eq. (3), *the primary discriminant for unseen classes is determined by the fitness of the class-level distribution, while the instance-level pseudo unseen distribution controls the construction of decision boundaries.* Then we use Proposition 3.1 to analyze question *(i)* of Sec. 3.1. Specifically, we consider the seen-unseen bias problem where unseen class data is misidentified as seen class. A wider pseudo-unseen distribution promotes wider decision boundaries for unseen classes, which helps to mitigate the seen-unseen bias. As illustrated in Fig. 2, the large variance provides a wider pseudo unseen distribution for f-CLSWGAN that is still close to the real unseen distribution. In contrast, the feature distribution in CE-GZSL excessively deviates from the human-defined distribution as it uses a linear mapping on the original visual feature. From the perspective of decision boundaries, we can also understand the common strategy of sampling a large number of pseudo-unseen samples in classifier training (Xian et al. 2018; Han et al. 2021). An additional pseudo unseen datum $\mathbf{x}^u$ pulls class weight $\mathbf{W}^u$ towards the corresponding pseudo unseen distribution while pushing other class weights away, thus widening the unseen decision boundaries.

In conclusion, in Sec. 3, we deconstruct and summarize the essential aspects of the generator and classifier in generation-based methods. Next, we will provide explicit optimization guidelines founded on the above analysis.

4 Generator-Classifier Learning under the Idea of Deconstruction

4.1 Learning Generator in Generalization View

In Sec. 3.1, we demonstrate that the generator-fitted instance-level unseen distribution is substitutable in Zero-Shot recognition. Therefore, we suggest focusing on optimizing the class-level distribution, which serves as the core to guide the gradient (Eq. (3)). To improve the class-level distribution, we provide insights from a generalization perspective. In typical supervised classification tasks, generalization refers to the learned conditional probability $q(y|\mathbf{x})$ from the empirical distribution $p(\mathbf{x}, y)$ fitting the test set. Inspired by this, we propose attribute generalization as the key to ZSL generator:

Proposition 4.1 (Key to ZSL generator)**.** *Attribute generalization in Zero-Shot generation is the conditional probability $p_g(\mathbf{x}|\mathbf{a})$ modeled on $p_r^s(\mathbf{x}, \mathbf{a}|\mathbf{a} \in \mathcal{A}^s)$ fitting $p_r^u(\mathbf{x}, \mathbf{a}|\mathbf{a} \in \mathcal{A}^u)$, where p_r^s and p_r^u are the real seen and unseen distributions, respectively.*

By converting a distributional learning problem into a generalization problem, we can handle it directly with existing tools. Leveraging the well-established research on generalization problems in supervised classification tasks, we investigate some existing overfitting suppression strategies such as L2 regularization, the Fast Gradient Method (Goodfellow, Shlens, and Szegedy 2014) (an adversarial training method), and attribute augmentation. These techniques lead to improvements in the original generator's Zero-Shot performance, as well as the CMMD value (Eq. (1)) against real unseen data. Please refer to the appendix for detailed results and additional experiments on attribute generalization.

4.2 Learning Classifier with Partly Biased Data

Due to the absence of unseen class data in ZSL setting, the generated unseen class data are bound to deviate from the real distribution, as shown in Fig. 2. Consequently, the main challenge in classifier learning is to capture the true decision boundary using partially biased data. However, data bias is unpredictable, and thus, it is essential for the classifier to adapt more toward the deterministic (*i.e.*, real seen) distribution and reduce the adverse effects of biased (*i.e.*, pseudo unseen class) distributions. Building upon the discussion in Sec. 3.2, we propose two principles for classifier design: (i) mitigating the impact of pseudo unseen samples on decision boundaries between seen classes during training, and (ii) reducing the seen-unseen bias.

4.3 A Simple Method over the Guidelines

We propose a simple method for verifying the validity of the above guidelines for generator-classifier learning. Our approach employs the widely-used (Gulrajani et al. 2017) as the generative model, which consists of a generator G and a discriminator D and is optimized by the following objective:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}\sim p_r}[D(\mathbf{x}, \mathbf{a})] - \mathbb{E}_{\widetilde{\mathbf{x}}}[D(\widetilde{\mathbf{x}}, \mathbf{a})]$$
$$-\lambda_0 \mathbb{E}_{\hat{\mathbf{x}}\sim p_{\hat{\mathbf{x}}}}[(\nabla_{\hat{x}}\|D(\hat{\mathbf{x}}, \mathbf{a})\|_2)^2 - 1],\ \widetilde{\mathbf{x}} = G(\mathbf{z}_0, \mathbf{a}), \tag{4}$$

where p_r denotes the real distribution of $\mathbf{x}$, $\mathbf{z}_0 \in \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\hat{\mathbf{x}} = \alpha\mathbf{x} + (1-\alpha)\widetilde{\mathbf{x}}$ with $\alpha \sim U(0,1)$ is for calculating the gradient penalty and λ_0 is a hyper-parameter.

We augment the attribute with Gaussian noise to enhance the attribute generalization (Proposition 4.1), *i.e.*,

$$G(\mathbf{z}_0, \mathbf{a}) \rightarrow G(\mathbf{z}_0, \mathbf{a} + \mathbf{z}_1), \tag{5}$$

where $\mathbf{z}_1 \in \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$, and σ decides the standard deviation of the augmenting distribution. The reason for attribute augmentation is detailed in the appendix.

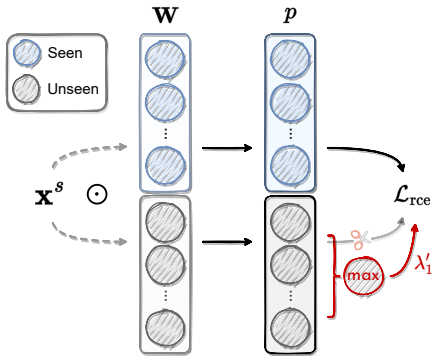

Figure 3: Illustration of the revised cross-entropy loss (Eq. (8)), where $\odot$ denotes the dot product. Per seen class sample, only the unseen class weight that gives it the largest activation is involved in the calculation. The calculating of the seen class weights remains unchanged.

During the classifier training phase, we follow *principle (i)* (Sec. 4.2) and begin by representing the unseen class corresponding terms in loss function as an increment (on the cross-entropy with seen class only), *i.e.*,

$$\begin{aligned}\mathcal{L}_{ce} = \frac{1}{n}[\sum_{c^s=1}^{|\mathcal{Y}^s|}\sum_{i}^{n_{c^s}} -\log\frac{p_y(\mathbf{x}_i)}{\hat{p}^s(\mathbf{x}_i)+\lambda_1\hat{p}^u(\mathbf{x}_i)} \\ +\lambda_2\sum_{c^u=1}^{|\mathcal{Y}^u|}\sum_{j}^{n_{c^u}} -\log p_y(\mathbf{x}_j)],\ \hat{p}^{\cdot}(\mathbf{x}_i)=\sum_{c=1}^{|\mathcal{Y}^{\cdot}|}p_c(\mathbf{x}_i),\end{aligned} \tag{6}$$

where $p_c(\cdot)$ is defined in Eq. (2), (3). We introduce two parameters, λ_1 and $\lambda 2$, to weight the generalized incremental forms. When λ_1 and λ_2 are set to zero, it indicates that the added pseudo unseen samples do not affect the seen class decision boundaries, and *principle (i)* can be achieved by selecting small values for λ_1 and λ_2.

Then we express the gradient of $\mathcal{L}_{ce}$ with respect to the weights of an unseen class, $\mathbf{W}_u$, as

$$\begin{aligned}-\frac{\partial_{\mathcal{L}_{ce}}}{\partial_{\mathbf{W}_u}} = \frac{\lambda_2}{n\tau}(\sum_{i=1}^{n_u}\mathbf{x}_i - \sum_{c^u=1}^{|\mathcal{Y}^u|}\sum_{j=1}^{n_{c^u}}p_u(\mathbf{x}_j)\mathbf{x}_j) \\ -\frac{1}{n\tau}\sum_{c^s=1}^{|\mathcal{Y}^s|}\sum_{k=1}^{n_{c^s}}\frac{\lambda_1 p_u(\mathbf{x}_k)}{\hat{p}^s(\mathbf{x}_k)+\lambda_1\hat{p}^u(\mathbf{x}_k)}\mathbf{x}_k.\end{aligned} \tag{7}$$

Here, a small λ_1 makes the seen data have little effect on the decision boundaries of unseen classes, while λ_2 determines the extent to which the loss function focuses on inter-unseen-class decision boundaries. This provides a direction to mitigate the seen-unseen bias, *i.e.*, the *principle (ii)*.

In summary, selecting a small value for λ_1 and an appropriate value for λ_2 aligns with the two guiding principles for classifier design. As λ_2 has the same optimization direction as the generation number of pseudo-unseen samples, we remove it by fixing it to 1. We assign different values of λ_1 to each unseen class based on their optimization difficulty. Empirically, we only set non-zero values for the hardest class, and only if it exceeds the true class score, as illustrated in Fig. 3. The revised cross-entropy formula is presented as:

$$\begin{aligned}\mathcal{L}_{rce} = \frac{1}{n}\sum_{c^s=1}^{|\mathcal{Y}^s|}\sum_{i}^{n_{c^s}} -\log\frac{p_y(\mathbf{x}_i)}{\hat{p}^s(\mathbf{x}_i)+\lambda_1' p_m^u(\mathbf{x}_i)} \\ +\sum_{c^u=1}^{|\mathcal{Y}^u|}\sum_{j}^{n_{c^u}} -\log p_y(\mathbf{x}_j),\ p_m^u(\mathbf{x}_i)=\max\{p_c(\mathbf{x}_i)|c\in\mathcal{Y}^u\},\end{aligned} \tag{8}$$

where $\lambda_1' = \lambda_1\mathbb{1}[p_m(\mathbf{x}_i > p_y(\mathbf{x}_i)]$, and $\mathbb{1}[\cdot]$ is the indicator function. The classifier trained with an appropriate value of λ_1 exhibits stronger inter-seen class discriminability and smaller seen-unseen bias, as demonstrated in Fig. 4 (c), (d). Finally, we constrain the classifier weights with the attributes using a mapping network $M(\cdot)$, *i.e.*,

$$\mathbf{W}_c := M(\mathbf{a}_c), c\in\mathcal{Y}^s\cup\mathcal{Y}^u, \tag{9}$$

which replaces the weights in Eq. (2). We also normalize the elements before feeding them into the dot product, which is a common strategy in ZSL. After training, a datum $\mathbf{x}$ is classified as the class with the attribute exhibiting the greatest similarity to it, *i.e.*,

$$\hat{y} = \arg\max_c \langle\frac{M(\mathbf{a}_c)}{||M(\mathbf{a}_c)||_2}, \frac{\mathbf{x}}{||\mathbf{x}||_2}\rangle, \tag{10}$$

where $||\cdot||_2$ denotes the l_2 norm.

5 Experiments

Benchmark Datasets. We conduct GZSL experiments on four public ZSL datasets. Animals with Attributes 2 (AWA2) (Lampert, Nickisch, and Harmeling 2013) contains 50 animal species and 85 attribute annotations, accounting 37,322 samples. Attribute Pascal and Yahoo (APY) (Farhadi et al. 2009) includes 32 classes of 15,339 samples and 64 attributes. Caltech-UCSD Birds-200-2011 (CUB) (Wah et al. 2011) consists of 11,788 samples with 200 bird species, annotated by 312 attributes. SUN Attribute (SUN) (Patterson and Hays 2012) carries 14,340 images from 717 different scenario-style with 102 attributes. We split the data into seen and unseen classes according to the common benchmark procedure in Xian, Schiele, and Akata (2017).

Representation. Most experiments are performed with the 2048-dimensional visual features extracted from the pre-trained ResNet101 (He et al. 2016), following Xian, Schiele, and Akata (2017). We also compare the GZSL performance on the fine-tuned data that we take from Chen et al. (2021b). For class representations (*i.e.*, attributes), we adopt the artificial attribute annotations that come with the datasets for AWA2, APY, and SUN, and employ the 1024-dimensional character-based CNN-RNN features (Reed et al. 2016) generated from textual descriptions for CUB.

Evaluation Metric. We calculate the average per-class top-1 accuracy among the unseen and seen classes respectively, denoted as A^u and A^s, then their harmonic mean H is employed as the measurement of GZSL. The classic ZSL is evaluated with per-class averaged top-1 accuracy on unseen classes (Xian, Schiele, and Akata 2017).

Implementation Details. The method proposed in Sec. 4 consists of three modules implemented with multi-layer perceptrons. The Generator G carries two hidden layers with

	Method	Source	AWA2			CUB			SUN			APY		
			A^u	A^s	H	A^u	A^s	H	A^u	A^s	H	A^u	A^s	H
†	Chou et al.	ICLR 2021	65.1	78.9	71.3	41.4	49.7	45.2	29.9	40.2	34.3	35.1	65.5	45.7
	SDGZSL	ICCV 2021b	64.6	73.6	68.8	59.9	66.4	63.0	48.2	36.1	41.3	38.0	57.4	45.7
	GCM-CF	CVPR 2021	60.4	75.1	67.0	61.0	59.7	60.3	47.9	37.8	42.2	37.1	56.8	44.9
	CE-GZSL	CVPR 2021	63.1	78.6	70.0	63.9	66.8	65.3	48.8	38.6	43.1	-	-	-
	SE-GZSL	AAAI 2022	59.9	80.7	68.8	53.1	60.3	56.4	45.8	40.7	43.1	-	-	-
	ICCE	CVPR 2022	65.3	82.3	72.8	67.3	65.5	66.4	-	-	-	45.2	46.3	45.7
	ZLA	IJCAI 2022a	65.4	82.2	72.8	73.0	64.8	68.7	50.1	38.0	43.2	40.2	53.8	46.0
	DGZ	**Proposed**	67.4	81.0	**73.6**	70.1	68.3	**69.2**	48.6	39.4	**43.5**	37.7	64.9	**47.7**
	DGZ w/o GM		65.9	78.2	71.5	71.4	64.8	68.0	49.9	37.6	42.8	38.0	63.5	47.6
‡	TF-VAEGAN*	ECCV 2020	55.5	83.6	66.7	63.8	79.3	70.7	41.8	51.9	46.3	-	-	-
	Chou et al.*	ICLR 2021	69.0	86.5	76.8	69.2	76.4	72.6	50.5	43.1	**46.5**	36.2	58.6	44.8
	GEM-ZSL	CVPR 2021c	64.8	77.5	70.6	64.8	77.1	70.4	38.1	35.7	36.9	-	-	-
	SDGZSL*	ICCV 2021b	69.6	78.2	73.7	73.0	77.5	75.1	51.1	40.2	45.0	39.1	60.7	47.5
	DPPN	NeurIPS 2021	63.1	86.8	73.1	70.2	77.1	73.5	47.9	35.8	41.0	40.0	61.2	48.4
	TransZero	AAAI 2022b	61.3	82.3	70.2	69.3	68.3	68.8	52.6	33.4	40.8	-	-	-
	MSDN	CVPR 2022c	62.0	74.5	67.7	68.7	67.5	68.1	52.2	34.2	41.3	-	-	-
	DGZ*	**Proposed**	71.7	83.7	**77.2**	76.9	77.7	77.3	49.4	43.5	46.3	37.1	79.3	**50.5**
	DGZ* w/o GM		67.2	85.7	75.4	77.4	78.0	**77.7**	50.4	39.8	44.5	38.5	67.4	49.0

Table 2: GZSL performance comparison with state of the arts. † denotes generative methods based on the common image feature proposed in Xian, Schiele, and Akata (2017). ‡ denotes allowing fine-tuning the feature extraction backbone, and * represents generative methods based on features extracted from the fine-tuned backbone. A^u and A^s are per-class accuracy scores (%) on seen and unseen test sets. H is their harmonic mean. The best results are shown in bold, with second place underlined.

4096 and 2048 dimensions. The Discriminator D contains one 4096-D hidden layer, and the mapping net M includes a 1024-D hidden layer. All the hidden layers are activated by Leaky-ReLU. We follow Xian et al. (2018) to set other hyper-parameters of WGAN-GP. In addition, we put 512 for the (mini) batch size and adopt Adam (Kingma and Ba 2015) as the optimizer with a learning rate of 1.0×10^{-4}.

5.1 Comparison with SotAs

We evaluate the proposed method by comparing its GZSL results with the current SotAs, as shown in Tab. 2. Notably, our results on common image features outperform SotAs in all four datasets. Moreover, our fine-tuned feature results ranked first on three datasets and second only to Chou, Lin, and Liu (2021) on SUN dataset. It is important to highlight that our approach is simple and does not require complex designs. Yet, it outperforms other complex approaches such as Chou, Lin, and Liu (2021), which uses the out-of-distribution discrimination method, and Han et al. (2021); Kong et al. (2022), which rely on instance discrimination, both leading to significant time consumption.

We also report the results without a generative model. The pseudo unseen distribution is constructed as mixed Gaussian distribution with the covariance as the statistics of the training set. A one-to-one mapping net (from attributes to visual class centers) estimates its mean (detailed in the appendix). In this baseline, our method still achieves comparable performance with current SotAs. It demonstrates the plug-in capability of the proposed classifier learning strategy, even in the case of no generator. It is also an attempt to simplify the generator-classifier framework.

Ablation	AWA2			CUB		
	A^u	A^s	H	A^u	A^s	H
i) w/o ATA	66.4	77.2	71.4	72.2	66.1	69.0
ii) w/o CR	39.8	89.4	55.1	58.3	70.9	64.0
iii) w/o M	64.0	79.4	70.9	70.7	57.8	63.6
iv) w/o CR&M	34.7	90.0	50.0	44.7	70.2	54.7
v) DIS $\rightarrow$ SCG	67.5	78.0	72.4	68.3	67.7	68.0
vi) DIS $\rightarrow$ GC+SCG	65.9	78.2	71.5	71.4	64.8	68.0
Full Model	67.4	81.0	**73.6**	70.1	68.3	**69.2**

Table 3: Ablation study results on AWA2 and CUB. The baselines are constructed by ablating some key modules. ATA: Attribute augmentation; CR: Classifier revision; M: Mapping net; SCG: Statistical-covariance Gaussian distribution; GC: Direct generating the class center.

5.2 Ablation Study

Baselines. To validate the effect of each component, we conduct an ablation study on AWA2 and CUB, with the following baselines: **i)** Setting σ to 0. **ii)** Training the classifier with vanilla cross-entropy. **iii)** Removing the mapping net (Eq. (9)). **iv)** Combination of ii) and iii). **v)** Replacing the WGAN-generated distribution with the statistical-covariance Gaussian distribution (same to Sec. 3.1). **vi)** On the basis of v), directly estimating the mean of the distribution by mapping from the attributes (same to Sec. 5.1).

Results. Tab. 3 depicts the results of this experiment. **Baseline i)** shows that the fewer effects of attribute augmentation on the fine-grained dataset CUB than on the coarse-grained dataset AWA2. This is mainly due to the fine-grained dataset's inherently smaller domain shift problem, causing less gain from a targeted approach. Meanwhile, for

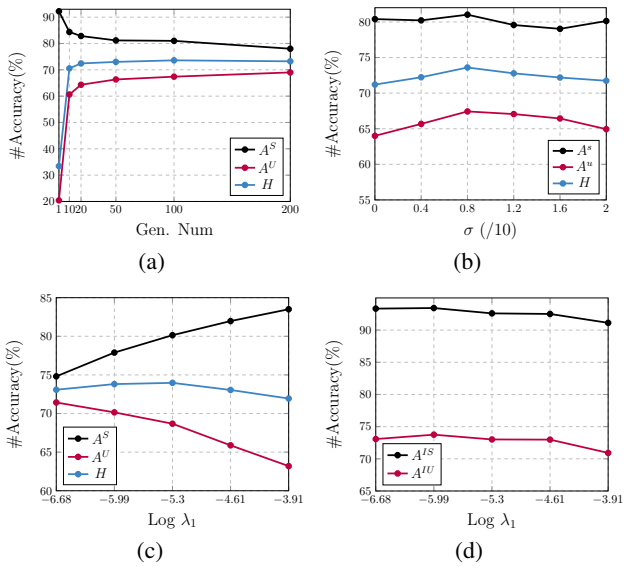

Figure 4: (a), (b), (c) GZSL performance *w.r.t.* the generation number per unseen class, σ, and λ_1. (d) Intra-discriminability of seen and unseen classes *w.r.t.* λ_1, where A^{is} and A^{iu} represent the intra- seen or unseen classes accuracy. The experiments are conducted on the AWA2 dataset.

the same reason, classifier revision plays a bigger role for AWA2 than for CUB **(baseline ii), iv))**. **Baseline iii), iv)** reflect the importance of the mapping net, which establishes implicit semantic connections between classifier weights. Overall, due to its intractability, attribute generalization enhancement brings fewer performance gains than classifier revision. **Baseline v) and vi)** compare the ways to obtain the mean of Gaussian distribution. **Baseline v)** averages the WGAN-generated samples for the mean of each class, which yields better performance than directly mapping attributes to the class mean **(baseline vi))**. This is probably because the instance-level modeling extracts more distribution information and better generalizes to unseen class attributes. More details and analysis are Provided in the appendix.

5.3 Hyper-parameters

The final objective involves four main hyperparameters: σ, τ, λ_1, and the generated number per unseen class. We set τ to 0.04, following Skorokhodov and Elhoseiny (2021); Chen et al. (2022a). We then analyze the influence of the other three parameters empirically. As shown in Fig. 4 (b), A^u and H have the same trend when σ varies, whose curves rise first and then fall as σ becomes larger. A big σ leads to performance degradation because a large variance of noise intuitively makes the attribute input of the generator lose inter-class discriminability. A small λ_1 mitigates the seen-unseen bias in Fig. 4 (c). Moreover, a suitable generated number creates the best performance, as shown in Fig. 4 (a), and the number is much smaller than the existing generation-based methods (100 *vs.* 2400 in (Han et al. 2021) and 4600 in (Chen et al. 2021a)). This demonstrates the joint effect of the number of generations and λ_1 as we stated in Sec. 4. We also report the effect of λ_1 on the intra-seen class discriminability in Fig. 4 (d), showing a downward trend when λ_1 increases within a certain range. We empirically generate 50 samples per unseen class in CUB, SUN, and APY, and 100 for AWA2 in all experiments. We put λ_1 to 4, 0.8, 0.04, and 0.005 for the above datasets. σ is set to 0.08 on all datasets.

Method	AWA2	CUB	SUN	APY
TCN (2019)	71.2	59.5	61.5	38.9
TF-VAEGAN (2020)	72.2	64.9	**66.0**	-
Chou et al. (2021)	73.8	57.2	63.3	41.0
IPN (2021b)	**74.4**	59.6	-	42.3
CE-GZSL (2021)	70.4	77.5	63.3	-
SDGZSL (2021b)	72.1	75.5	-	45.4
DGZ	74.0	**80.1**	65.4	**46.6**

Table 4: Discriminability on unseen classes, evaluated by ZSL performance (%) (compared with SotAs). Note that our classifier is trained toward the GZSL setting.

5.4 Discriminability on Unseen Classes

As shown in Tab. 4, we analyze the discriminability of the trained GZSL classifier among unseen classes, quantified by ZSL accuracy. Despite not being specifically designed for the ZSL setting, our model still achieves comparable results to SotA ZSL methods. This is primarily due to improvements in attribute generalization ability and the intrinsic semantic association of classifier weights carried from attribute mapping.

6 Conclusion

In this paper, we deconstruct the generator-classifier Zero-Shot Learning framework. We begin by decomposing the unseen class distribution learned by the generator into class- and instance-level distribution. Then we empirically analyze the learning center of the generator and the role of these two distributions in classifier learning. Specifically, we emphasize attribute generalization in generator training and regard classifier training as an independent task to learn from partially biased data. Based on these points, we propose a simple method that outperforms current SotAs in performance without a complex design, demonstrating the effectiveness of the proposed guideline. Additionally, we evaluate the transferability of the proposed method and find that it can achieve SotA even when replacing the generative model with a class center mapping net. We acknowledge that our analysis is primarily empirical and lacks mathematical discussion. We will explore the generation-based framework more thoroughly from a theoretical standpoint and continue to simplify it in future work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61872187, 62077023, 62072246), the Natural Science Foundation of Jiangsu Province (BK20201306), and the "111 Program" (B13022).

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In *CVPR*, 819–826.

Akata, Z.; Reed, S.; Walter, D.; Lee, H.; and Schiele, B. 2015. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2927–2936.

Atzmon, Y.; and Chechik, G. 2019. Adaptive confidence smoothing for generalized zero-shot learning. In *CVPR*, 11671–11680.

Chao, W.-L.; Changpinyo, S.; Gong, B.; and Sha, F. 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, 52–68.

Chen, D.; Shen, Y.; Zhang, H.; and Torr, P. H. 2022a. Zero-Shot Logit Adjustment. In Raedt, L. D., ed., *IJCAI*, 813–819. International Joint Conferences on Artificial Intelligence Organization.

Chen, S.; Hong, Z.; Liu, Y.; Xie, G.-S.; Sun, B.; Li, H.; Peng, Q.; Lu, K.; and You, X. 2022b. TransZero: Attribute-guided Transformer for Zero-Shot Learning. In *AAAI*.

Chen, S.; Hong, Z.; Xie, G.-S.; Yang, W.; Peng, Q.; Wang, K.; Zhao, J.; and You, X. 2022c. MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning. In *CVPR*, 7612–7621.

Chen, S.; Wang, W.; Xia, B.; Peng, Q.; You, X.; Zheng, F.; and Shao, L. 2021a. FREE: Feature Refinement for Generalized Zero-Shot Learning. In *ICCV*.

Chen, Z.; Luo, Y.; Qiu, R.; Huang, Z.; Li, J.; and Zhang, Z. 2021b. Semantics Disentangling for Generalized Zero-shot Learning. In *ICCV*.

Chou, Y.-Y.; Lin, H.-T.; and Liu, T.-L. 2021. Adaptive and generative zero-shot learning. In *ICLR*.

Elhoseiny, M.; Saleh, B.; and Elgammal, A. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, 2584–2591.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *CVPR*, 1778–1785.

Frome, A.; Corrado, G.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2121–2129.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. arXiv:"1412.6572".

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. 2017. Improved training of wasserstein gans. In *NeurIPS*.

Han, Z.; Fu, Z.; Chen, S.; and Yang, J. 2021. Contrastive Embedding for Generalized Zero-Shot Learning. In *CVPR*, 2371–2381.

Han, Z.; Fu, Z.; and Yang, J. 2020. Learning the redundancy-free features for generalized zero-shot object recognition. In *CVPR*, 12865–12874.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. In *NeurIPS*.

Huynh, D.; and Elhamifar, E. 2020. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, 4483–4493.

Jiang, H.; Wang, R.; Shan, S.; and Chen, X. 2019. Transferable contrastive network for generalized zero-shot learning. In *ICCV*, 9765–9774.

Kim, J.; Shim, K.; and Shim, B. 2022. Semantic feature extraction for generalized zero-shot learning. In *AAAI*, 1166–1173.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. In *ICLR*.

Kong, X.; Gao, Z.; Li, X.; Hong, M.; Liu, J.; Wang, C.; Xie, Y.; and Qu, Y. 2022. En-Compactness: Self-Distillation Embedding & Contrastive Generation for Generalized Zero-Shot Learning. In *CVPR*, 9306–9315.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 951–958.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 453–465.

Li, K.; Min, M. R.; and Fu, Y. 2019. Rethinking zero-shot learning: A conditional visual classification perspective. In *ICCV*, 3583–3592.

Liu, J.; Bai, H.; Zhang, H.; and Liu, L. 2021a. Near-Real Feature Generative Network for Generalized Zero-Shot Learning. In *ICME*, 1–6.

Liu, L.; Zhou, T.; Long, G.; Jiang, J.; Dong, X.; and Zhang, C. 2021b. Isometric propagation network for generalized zero-shot learning. In *ICLR*.

Liu, Y.; Zhou, L.; Bai, X.; Huang, Y.; Gu, L.; Zhou, J.; and Harada, T. 2021c. Goal-oriented gaze estimation for zero-shot learning. In *CVPR*, 3794–3803.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *ICML*, 97–105. PMLR.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. In *ICLR Work-shop Papers*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *NeurIPS*.

Min, S.; Yao, H.; Xie, H.; Wang, C.; Zha, Z.-J.; and Zhang, Y. 2020. Domain-aware visual bias eliminating for generalized zero-shot learning. In *CVPR*, 12664–12673.

Narayan, S.; Gupta, A.; Khan, F. S.; Snoek, C. G.; and Shao, L. 2020. Latent embedding feedback and discriminative features for zero-shot classification. In *ECCV*, 479–495.

Palatucci, M. M.; Pomerleau, D. A.; Hinton, G. E.; and Mitchell, T. 2009. Zero-shot learning with semantic output codes. In *NeurIPS*. Carnegie Mellon University.

Parikh, D.; and Grauman, K. 2011. Relative attributes. In *ICCV*, 503–510. IEEE.

Patterson, G.; and Hays, J. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2751–2758.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 49–58.

Shen, Y.; Qin, J.; Huang, L.; Liu, L.; Zhu, F.; and Shao, L. 2020. Invertible zero-shot recognition flows. In *ECCV*, 614–631.

Skorokhodov, I.; and Elhoseiny, M. 2021. Class Normalization for (Continual)? Generalized Zero-Shot Learning. In *ICLR*.

Tolstikhin, I.; Bousquet, O.; Gelly, S.; and Schoelkopf, B. 2017. Wasserstein auto-encoders. In *ICLR*.

Verma, V. K.; Brahma, D.; and Rai, P. 2020. Meta-learning for generalized zero-shot learning. In *AAAI*, 6062–6069.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. Technical report, california institute of technology.

Wang, C.; Min, S.; Chen, X.; Sun, X.; and Li, H. 2021. Dual Progressive Prototype Network for Generalized Zero-Shot Learning. In *NeurIPS*, 2936–2948.

Wu, J.; Zhang, T.; Zha, Z.-J.; Luo, J.; Zhang, Y.; and Wu, F. 2020. Self-supervised domain-aware generative network for generalized zero-shot learning. In *CVPR*, 12767–12776.

Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018. Feature generating networks for zero-shot learning. In *CVPR*, 5542–5551.

Xian, Y.; Schiele, B.; and Akata, Z. 2017. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 4582–4591.

Xian, Y.; Sharma, S.; Schiele, B.; and Akata, Z. 2019. f-gan-d2: A feature generating framework for any-shot learning. In *CVPR*, 10275–10284.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2020. Attribute prototype network for zero-shot learning. In *NeurIPS*, 21969–21980.

Yue, Z.; Wang, T.; Sun, Q.; Hua, X.-S.; and Zhang, H. 2021. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, 15404–15414.

Zhang, L.; Xiang, T.; and Gong, S. 2017. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2021–2030.

Zhang, Z.; and Saligrama, V. 2015. Zero-shot learning via semantic similarity embedding. In *ICCV*, 4166–4174.

Zhu, Y.; Xie, J.; Tang, Z.; Peng, X.; and Elgammal, A. 2019. Semantic-guided multi-attention localization for zero-shot learning. In *NeurIPS*.