# Data Distribution Distilled Generative Model for Generalized Zero-Shot Recognition

**Yijie Wang[1], Mingjian Hong[1], Luwen Huangfu[2], Sheng Huang[1]***

[1]Chongqing University
[2]Fowler College of Business, San Diego State University
wangyj@stu.cqu.edu.cn, hmj@cqu.edu.cn, lhuangfu@sdsu.edu, huangsheng@cqu.edu.cn

## Abstract

In the realm of Zero-Shot Learning (ZSL), we address biases in Generalized Zero-Shot Learning (GZSL) models, which favor seen data. To counter this, we introduce an end-to-end generative GZSL framework called D$^3$GZSL. This framework respects seen and synthesized unseen data as in-distribution and out-of-distribution data, respectively, for a more balanced model. D$^3$GZSL comprises two core modules: in-distribution dual space distillation (ID$^2$SD) and out-of-distribution batch distillation (O$^2$DBD). ID$^2$SD aligns teacher-student outcomes in embedding and label spaces, enhancing learning coherence. O$^2$DBD introduces low-dimensional out-of-distribution representations per batch sample, capturing shared structures between seen and unseen categories. Our approach demonstrates its effectiveness across established GZSL benchmarks, seamlessly integrating into mainstream generative frameworks. Extensive experiments consistently showcase that D$^3$GZSL elevates the performance of existing generative GZSL methods, underscoring its potential to refine zero-shot learning practices.The code is available at: https://github.com/PJBQ/D3GZSL.git

## Introduction

Contemporary techniques for object classification (He et al. 2016) are primarily rooted in supervised learning, mandating a substantial pool of labeled data. Deep convolutional neural networks (Krizhevsky, Sutskever, and Hinton 2017; Tan and Le 2019; Xie et al. 2017) have elevated image classification performance within a predefined spectrum of categories, benefitting from an abundance of training samples. However, real-world classification often involves a distribution skewed toward certain categories, resulting in limited or even absent samples for others. The reliance of deep models on extensive data leads to suboptimal performance when labeled data is scarce (Wang et al. 2019). The emergence of zero-shot learning (ZSL) techniques (Lampert, Nickisch, and Harmeling 2009; Palatucci et al. 2009) is prompted by real-world demands and technological progress. ZSL endeavors to train a model capable of categorizing objects from unseen classes (target domain) by transferring knowledge gleaned from seen classes (source domain), employing semantic information as a channel between the two domains. Traditional ZSL approaches involve test sets exclusively composed of samples from unseen classes, an unrealistic scenario that inadequately mirrors actual recognition conditions. In practical applications, seen class data samples typically outnumber those from unseen classes, making it imperative to concurrently identify samples from both categories, rather than restricting classification to unseen class samples alone. This more practical setting is known as generalized zero-shot learning (GZSL).

Generative methodologies (Mishra et al. 2018; Verma et al. 2018; Kong et al. 2022) constitute a pivotal aspect of ZSL, enhancing its efficacy through data augmentation. By generating samples for unseen classes, a GZSL problem can be transformed into a traditional supervised learning problem. An illustrative example is f-CLSWGAN (Xian et al. 2018b), which utilizes the Wasserstein GAN (Arjovsky, Chintala, and Bottou 2017) and classification loss to generate CNN features endowed with robust discriminatory attributes. A complementary approach, f-VAEGAN-D2 (Xian et al. 2019), proposes a conditional generative model that combines the advantages of VAEs and GANs to generate more robust features. An additional contribution is made by CE-GZSL (Han et al. 2021), which harnesses both instance-level and class-level contrastive supervision to enhance the discriminative capabilities of the embedding space.

Despite the impressive achievements of current generative methodologies (Xian et al. 2018b; Han et al. 2021), a prevailing limitation surfaces. Most of these approaches, constrained by the scarcity of seen data, primarily concentrate on delineating the correlation between semantic knowledge and the available seen data. This dynamic naturally skews the generative model's inclination towards generating samples aligned with the seen data distribution, as depicted in Fig.1. Out-of-distribution (OOD) detection (Lee et al. 2018; Sun et al. 2022; Sun, Guo, and Li 2021) aims to identify data samples that are abnormal or substantially distinct from the other available samples. By reevaluating GZSL from the perspective of OOD detection, seen data can essentially be viewed as ID (In-Distribution) data, while the absent unseen data corresponds to OOD data. This perspective underscores that the bulk of mainstream generative techniques predominantly grapple with ID modeling, sidelining the crucial tasks of OOD modeling. Fortunately, some researchers

*Corresponding Author.

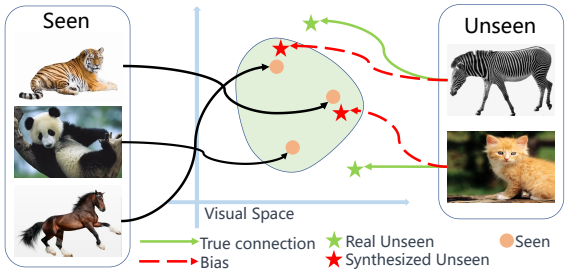


Figure 1: A schematic view of the bias concerning seen classes (source) in the visual space.

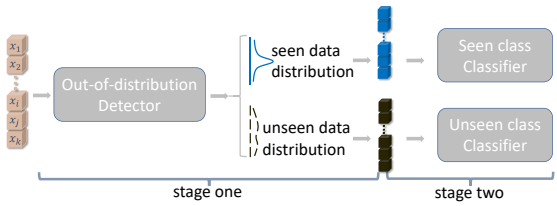


Figure 2: Two-stage classification method based on OOD detection. Stage one: OOD detector performs binary classification of the input data into seen and unseen categories. Stage two: Two expert classifiers separately classify the samples that the Out-Of-Distribution (OOD) detector identifies as seen and unseen categories.

(Chen et al. 2020; Keshari, Singh, and Vatsa 2020; Mandal et al. 2019) have recognized this limitation and have introduced OOD techniques to facilitate the incorporation of OOD insights. For instance, (Mandal et al. 2019) combines OOD detection techniques with generative methods to address the challenges encountered in ZSL. As shown in Fig.2, the training process of these methods is not end-to-end, and is usually divided into two steps. A generative approach first trains to produce unseen samples, followed by training an OOD detector with both synthetic unseen and real seen samples. Simultaneously, two expert classifiers are trained separately on these generated unseen and authentic seen samples. However, this non-end-to-end training strategy overlooks the potential benefits of OOD detection insights during the optimization of the generative model, ultimately neglecting to model the data distribution across both seen and unseen classes. In the inference phase, the OOD detector is applied to distinguish seen classes instances from those of the unseen classes and the domain expert classifiers (seen/unseen) are used to individually classify data samples. Nonetheless, this two-stage classification approach can accrue errors, leading to suboptimal performance.

To address the above issues, we present a novel end-to-end knowledge distillation framework named Data Distribution Distilled Generative Zero-shot Learning ($D^3$GZSL) for mining and taking advantage of both ID and OOD knowledge of data, aiming to align the distribution of generated samples more closely with the distribution of actual samples. By training a unified classifier for classification, we successfully circumvent the issue of error accumulation in two-stage classification encountered in (Chen et al. 2020; Mandal et al. 2019). As shown in Fig.3, our framework comprises a Feature Generation(FG), In-Distribution Dual-Space Distillation($ID^2SD$), and Out-Of-Distribution Batch Distillation($O^2DBD$). The FG leverages the semantic relationships between seen and unseen classes to synthesize features for unseen classes. $ID^2SD$ is leveraged to align the outputs of teacher and student in both embedding and label spaces. Logits-layer distillation enables our target network to model the probability distribution of in-distribution data (seen class) from the expert network. The feature-based distillation excavates the correlation between samples of different categories. This dual-space approach could provide multifaceted guidance for the student model's learning process, thereby enhancing the precision of the distribution of seen category data. In $O^2DBD$, we incorporate a low-dimensional representation to encode OOD information for each sample within a batch. Subsequently, we model the correlations among these low-dimensional OOD representations to capture the potentially shared inherent structures between seen and unseen categories. This approach stimulates the model to grasp more nuanced, intricate features while simultaneously acquiring a diverse spectrum of features.

$D^3$GZSL can be applied to generally boost any generative GZSL models. The experiments on four benchmarks verify that $D^3$GZSL consistently improves the performances of three well known generative frameworks, such as a VAE-based model (Narayan et al. 2020) and two GAN-based models (Han et al. 2021; Xian et al. 2018b). Additionally, we have incorporated a novel generative GZSL method based on denoising diffusion model (Xiao, Kreis, and Vahdat 2021). The results also reveal that some generative GZSL approaches enhanced by our frameworks achieve promising performances compared with SOTA.

Our contributions encompass three key facets, each contributing uniquely to the advancement of GZSL:

(1) We introduce a novel generative framework for GZSL, named $D^3$GZSL, that operates in an end-to-end manner. Our framework incorporates the distilled knowledge emanating from both seen and unseen data distributions through the integration of OOD detection methodologies.

(2) Within our framework, $ID^2SD$ plays a crucial role in harmonizing the outputs of both teacher and student networks across embedding and label spaces. Furthermore, in $O^2DBD$, we introduce a low-dimensional OOD representation for within each batch, subsequently modeling their interrelations under the guidance of class labels.

(3) Notably, our methodology is adaptable and can be seamlessly integrated into mainstream generative frameworks such as GAN, VAE, and the diffusion model. This versatility allows our approach to enhance the capabilities of various existing generative GZSL methodologies.

## Related Work

**Generalized Zero-Shot Learning.** ZSL (Lampert, Nickisch, and Harmeling 2009; Palatucci et al. 2009; Xing et al. 2020; Huang, Lin, and Huangfu 2020) aims to train a model capable of classifying objects belonging to unseen classes (target domain) by leveraging knowledge acquired from seen classes (source domain), with the assistance of semantic information. Early ZSL (Romera-Paredes and Torr 2015; Kodirov, Xiang, and Gong 2017; Liu et al. 2021; Xu et al.

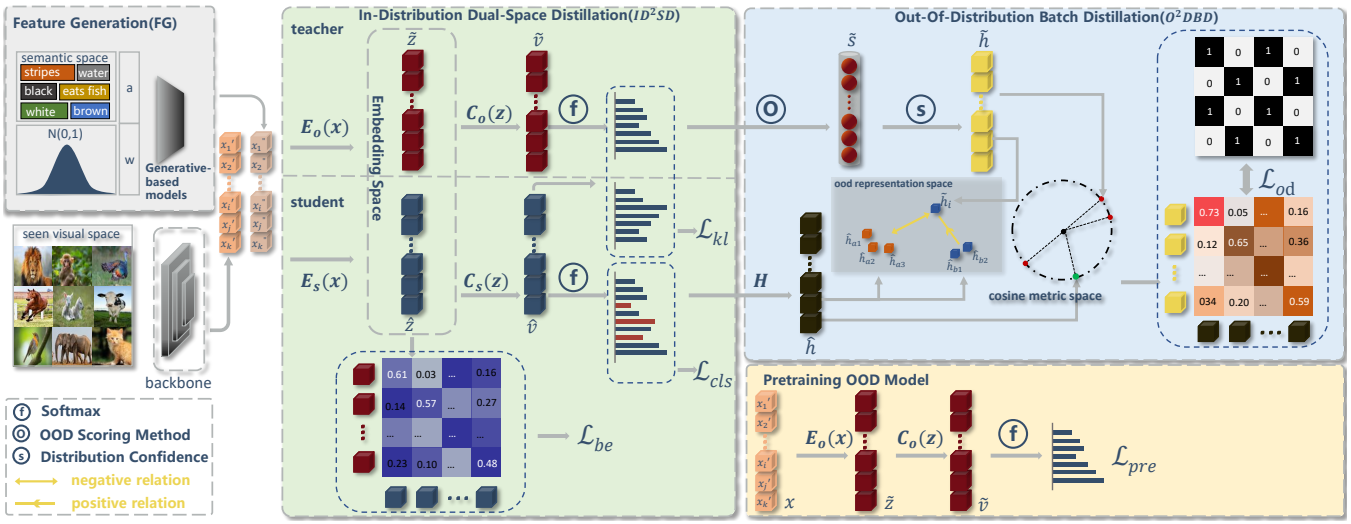


Figure 3: The structure of our D$^3$GZSL framework. The FG is our baseline model, which is a generative ZSL method. In ID$^2$SD, we learn two embedding function $E_o$ and $E_s$ that map the visual samples $x$ into the embedding space as $z = E(x)$. $C_o$ and $C_s$ are the classifier networks of the teacher and student architectures, respectively. $f$ is a softmax function. In O$^2$DBD, $O$ is OOD scoring method. $H$ is a mapping function that maps the softmax probability of student network to the OOD representation embedding space. $S$ is the transformation of out-of-distribution detection scores into OOD representation space.

2022) methods focused on learning an embedding space that connects the low-level visual features of seen classes to their corresponding semantic vectors. Recently, generative methods have focused on learning a model to generate images or visual features for unseen classes, drawing on samples from seen classes and semantic representations of both classes. (Xian et al. 2018b) developed a conditional WGAN model incorporating a classification loss to generate visual features for unseen classes, a model known as f-CLSWGAN. Subsequent to this, the conditional Generative Adversarial Network (CGAN) has been integrated with various strategies to produce discriminative visual features for the unseen classes. TFGNSCS (Lin et al. 2019) is an extended version of f-CLSWGAN, designed to take into account transfer information. By putting more emphasis on intra-class relationships but the inter-class structures, ICCE (Kong et al. 2022) can distinguish different classes with better generalization. However, these generative-based models often create highly unconstrained visual features for unseen classes, which can result in synthetic samples that deviate significantly from the actual distribution of real visual features. To address this issue, various strategies have been proposed, including calibrated stacking (Felix et al. 2019; Chao et al. 2016) and novelty detector (Atzmon and Chechik 2019; Min et al. 2020). In this paper, we address this issue by incorporating out-of-distribution detection techniques into the task of generalized zero-shot recognition.

**Out-of-Distribution Detection.** OOD detection (Hendrycks and Gimpel 2016; Hendrycks, Mazeika, and Dietterich 2018; Lee et al. 2017; Wang et al. 2022) is the task of detecting when a sample is drawn from a distribution different from the training data. Some techniques (Huang, Geng, and Li 2021; Liang, Li, and Srikant 2017; Liu et al. 2020; Hendrycks and Gimpel 2016; Lee et al. 2018; Sun, Guo, and Li 2021) concentrate on developing score functions for OOD detection during the inference stage and can be easily implemented without modifying the model parameters. The Maximum Softmax Probability (MSP) (Hendrycks and Gimpel 2016) approach uses the maximum prediction value from the model as the OOD score function. Moreover, (Liu et al. 2020) proposes to replace the softmax function with the energy functions for OOD detection. Recently, (Huang, Geng, and Li 2021) has been introduced as an improvement in OOD detection, where it uses the similarity between the model's predicted probability distribution and a uniform distribution to attain state-of-the-art results. Several pioneering works (Chen et al. 2020; Keshari, Singh, and Vatsa 2020; Mandal et al. 2019) have utilized OOD to solve zero-shot learning tasks. In (Chen et al. 2020), the boundary-based Out-of-Distribution (OOD) classifier learns a defined manifold for each seen class within a unit hyper-sphere, which serves as the latent space. Utilizing the boundaries of these manifolds along with their centers, unseen samples can be distinguished from the seen samples.

**Knowledge Distillation.** Knowledge distillation (Hinton, Vinyals, and Dean 2015) aims to train a more compact student network by replicating the behavior of a pretrained, complex teacher network. This knowledge can be response-based or feature-based. Response-based (Chen et al. 2017; Hinton, Vinyals, and Dean 2015) knowledge typically refers to the neural response of the teacher model's final output layer, such as logits and bounding box offsets in object detection tasks. Feature-based (Romero et al. 2014; Zagoruyko and Komodakis 2016; Chen et al. 2021) knowledge from intermediate layers serves as a valuable extension of response-based knowledge, particularly for training thinner and deeper networks. Specifically, (Zagoruyko and Komodakis 2016) derived an "attention map" from the original feature maps for conveying knowledge. Knowledge distilla-

tion is frequently utilized for model compression, fusion, or performance enhancement. In our approach, we amalgamate distillation learning with Out-of-Distribution detection techniques, aiming to align the distribution of generated samples more closely with the distribution of actual samples.

# Methodology

## Problem Statement

In GZSL, there are two separate sets of classes: $\mathcal{S}$ seen classes in $\mathcal{Y}_s$ and $\mathcal{U}$ unseen classes in $\mathcal{Y}_u$. We define a training set $\mathcal{D}_{tr} = \{(x_i, y_i)\}_{i=1}^{N}$, $N$ is the number of seen images. $x_i$ is a visual feature, $y_i$ is its class label in $\mathcal{Y}_s$, where we have $\mathcal{Y}_s \cap \mathcal{Y}_u = \varnothing$. The test set $\mathcal{D}_{te} = \{x_i, y_i\}_{i=N+1}^{N+M}$ holds $M$ unlabeled instances. The instances in $\mathcal{D}_{te}$ exclusively originate from both seen and unseen classes. Additionally, semantic embeddings (attributes) $\mathcal{A} = \{a_i\}_{i=1}^{S+U}$ for $S$ seen classes and $U$ unseen classes are also supplied. The attributes serve as the link between seen and unseen classes throughout the entire training process in GZSL settings.

## D$^3$-GZSL Framework

As illustrated in Fig.3, we present a data distribution distillation framework named D$^3$GZSL to generally boost generative models for achieving better GZSL performance. D$^3$GZSL deems the real seen data and generated unseen data as the ID data and OOD data respectively. It consists of FG, ID$^2$SD and O$^2$DBD. FG module is used to generate the features for unseen classes. Here, it can be replaced any generative models. ID$^2$SD aims to utilize the advantages of logits-layer and feature-based distillation to model the data distribution of seen classes in a detailed and accurate manner. O$^2$DBD conducts an analysis from the OOD perspective with the objective of capturing the potentially shared inherent structures between both seen and unseen categories. In this section, we will introduce these three modules in detail

**Feature Generation(FG).** This module often consists of a generative GZSL approach. Among these models, we have chosen four baseline models, including a VAE-based model (Narayan et al. 2020) and two GAN-based models (Han et al. 2021; Xian et al. 2018b). Furthermore, (Xiao, Kreis, and Vahdat 2021) introduced a novel denoising diffusion GAN, in which the denoising distributions are modeled with conditional GAN. We incorporate their method(Xiao, Kreis, and Vahdat 2021) into zero-shot learning as one of our baseline models. Building on this, We significantly modified the original image generation model for feature generation in GZSL tasks, detailed in the **supplementary materials**. We use $\mathcal{L}_{gen}$ to represent the loss of these generative GZSL methods. We utilize $G$ to symbolize generative models. $G$ is capable of transforming the semantic embedding $a$ and normal sampling $w$ into visual features $x$. We employ $x''$ to denote the generated features, thus expressing

$$x'' = G(a, w). \tag{1}$$

**In-Distribution Dual-Space Distillation (ID$^2$SD).** Our objective is to develop a generic classifier capable of distinguishing between both seen and unseen categories. We are able to reliably train a teacher network, utilizing only authentic samples. We adopt the teacher-student network framework of dual-space distillation to build our ID$^2$SD model. The knowledge obtained from our reliable labels is distilled alongside the aspects corresponding to our unified classifier. Similarly, at the feature level, we possess a trustworthy feature extractor $E_o$, with the aspiration that $E_s$ within our student network exhibits comparable capabilities. In fact, the distilled knowledge contains not only feature information but also mutual relations of data samples. Instead of measuring the similarity between the features directly through methods like MSE, we explore the correlation of samples within a batch matrix. Our methodology consists of employing the embedding function $E_o$ along with the classifier $C_o$ to construct the teacher network. The student network architecture mirrors that of the teacher network, albeit with a distinction in that the classifier $C_s$ utilizes different dimensions. $C_s$ encompasses unified classifiers for both seen and unseen categories. $x$ symbolizes the sample feature, constituting the input of our model, wherein $x'$ stands for the real sample feature, and $x''$ denotes the generated sample feature. Where $\tilde{z} = E_o(x)$ and $\hat{z} = E_s(x)$, $z$ forms the features within our embedding space. $v$ represents the logical layer output of the network, expressed as $\tilde{v} = C_o(z)$ and $\hat{v} = C_s(z)$. We employ $\phi$ to denote the composite function of $E_o$ and $C_o$, and use $\psi$ to symbolize the integration of $E_s$ and $C_s$, defining our teacher-student network as follows:

$$\begin{aligned} \tilde{v} &= \phi(x), \\ \hat{v} &= \psi(x). \end{aligned} \tag{2}$$

**Batch-Wise ID Embedding Identical Loss.** We contemplate the uniformity of samples, along with the interrelation between sample features. We initially construct the batch embedding similarity matrix $A$, wherein its element $a_{ij}$ as the cosine similarity between $\tilde{z}$ and $\hat{z}$, $a_{ij} = \frac{\hat{z}_i^T \tilde{z}_j}{\|\hat{z}_i\|_2 \|\tilde{z}_j\|_2}$. Our similarity matrix $A$ encodes the correlation between samples. Following (Kong et al. 2022), we formulate a ground truth, designating 1 for samples of identical classes and 0 for those of disparate classes. Thus, we metamorphose our task into a binary classification predicament. We constrain our loss calculations to utilize sample features exclusively from the seen category. Here, $N_a$ represents the number of samples belonging to the seen category within a batch. Our objective function is designed as follows:

$$\begin{aligned} \mathcal{L}_{be} = \frac{1}{N_a \times N_a} \sum_{i=1}^{N_a} \sum_{j=1}^{N_a} \big( & \mathbb{1}_{y_i = y_j} \log(\sigma(a_{ij})) \\ & + \mathbb{1}_{y_i \neq y_j} \log(1 - \sigma(a_{ij})) \big), \end{aligned} \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function.

**Instance-Wise ID Logit Identical Loss.** The purpose of instance-wise ID logit identical loss is to align the outputs of the student and teacher networks in label space. We initially extract the dimension corresponding to the seen category in $\hat{v}$, employing $\ddot{v}$ as a representation. The $L_2$ normalized logits within the teacher and student network is articulated as

follows: $v_o = \frac{\tilde{v}}{\|\tilde{v}\|_2}$ and $v_s = \frac{\ddot{v}}{\|\ddot{v}\|_2}$. The teacher and student networks' probability distributions are obtained as follows:

$$p_o{}^{(k)} = \frac{\exp\left(v_o{}^{(k)}\right)}{\sum_{i=1}^{S} \exp\left(v_o{}^{(i)}\right)}, k = 1, 2, ...S,$$
$$p_s{}^{(k)} = \frac{\exp\left(v_s{}^{(k)}\right)}{\sum_{i=1}^{S} \exp\left(v_s{}^{(i)}\right)}, k = 1, 2, ...S, \tag{4}$$

where $S$ is the seen class number, $k$ is the class index. In pursuit of ensuring that projections from the identical seen class yield the same predicted probability, we introduce the normalized probability distillation loss, where $D_{KL}(p_o \| p_s)$ denotes the $KL$ divergence between $p_o$ and $p_s$.

$$\mathcal{L}_{kl}(p_s, p_o) = D_{KL}(p_o \| p_s) = \sum_{k=1}^{K} p_o{}^{(k)} \log(\frac{p_o{}^{(k)}}{p_s{}^{(k)}}), \tag{5}$$

**Classification Loss.** Our framework constitutes an end-to-end trainable network, and the trained student network can be directly applied to GZSL inference. Beyond the dual-space distillation loss, we conduct supervised learning on the student network. The network's input encompasses the real features of the seen samples as well as the generated features of the unseen samples. The cross-entropy loss is employed to supervise the classifier with the class labels:

$$\mathcal{L}_{cls}(v, y) = -\sum_{i=1}^{S+U} y^{(i)} \log \frac{\exp(v^{(i)}/\tau_s)}{\sum_{k=1}^{S+U} \exp(v^{(k)}/\tau_s)}, \tag{6}$$

where $\tau_s > 0$ is the temperature parameter, $S$ and $U$ respectively represent the number of seen classes and the number of unseen classes. $y^{(i)}$ represents the i-th element in the true probability distribution, and $v^{(i)}$ constitutes the i-th element in the predicted vector.

**Total Loss of ID$^2$SD.** The final optimization objective of our ID$^2$SD is formulated as:

$$\mathcal{L}_{id} = \mathbb{E}[\mathcal{L}_{be}] + \mathbb{E}[\mathcal{L}_{kl}(p_s, p_o)] + \mathbb{E}[\mathcal{L}_{cls}(v, y)]. \tag{7}$$

**Out-of-Distribution Batch Distillation (O$^2$DBD).** In O$^2$DBD, we introduce a low-dimensional representation to encode OOD information for each sample within a batch. And then we model the correlations among these low-dimensional OOD representations. When unseen data serves as input, the teacher network generates a uniform distribution of equal probabilities across all seen categories, resulting in maximum entropy and inhibiting the model's ability to identify the input sample. Leveraging this distinctive feature of the OOD detection, we have constructed an out-of-distribution batch distillation network specifically for GZSL.

**OOD Logits.** The score-based strategy is commonly used by many notable OOD detection methods. The basic idea of this type of method is to use the model's output scores to determine whether an input comes from a known (i.e., seen during training) data distribution. The model calculates a score for each input sample, indicating its confidence in the sample's category. A key step in this method is determining a threshold $\gamma$ for the obtained scores, which is used to distinguish between ID and OOD data. Diverging from traditional OOD detection methods, we have crafted an approach for confidence estimation, without the need to manually set a threshold value $\gamma$. We obtain a low-level representation $\tilde{h}$, which encodes both the ID and OOD information. $\tilde{h}$ consists of two elements. The first element is the ID confidence $c$, and the second element is the OOD confidence $\hat{c}$. Through an OOD scoring method, we derive an OOD detection score $\tilde{s}$ (e.g., $\tilde{s} = \max(\text{softmax}(F_o(x_i)))$ (Hendrycks and Gimpel 2016)). We employ a learnable sigmoid function $\epsilon$ to compress the score $\tilde{s}$ into the range between 0 and 1, namely $c = \epsilon(\tilde{s})$. We regard this as a probabilistic scenario, $\hat{c} = 1 - c$. Then we learn a mapping function $H$ in the student network to map the softmax probability of student network to the OOD representation embedding space.

$$\hat{h} = H(\psi(x)). \tag{8}$$

**Batch-Wise OOD Logits Identical Loss.** We employ the identical loss function as expressed in Eq.3, utilizing the confidence label $\tilde{h}$, to construct supervised learning in a low-dimensional space that encodes both ID and OOD information. We first create an OOD representation similarity matrix $B$, whose element $b_{ij}$ is derived by calculating the cosine similarity between $\tilde{h}$ and $\hat{h}$, $b_{ij} = \frac{\hat{h}_i^T \tilde{h}_j}{\|\hat{h}_i\|_2 \|\tilde{h}_j\|_2}$. Our aim is to augment the confidence of the same class within the student and teacher networks while minimizing the confidence of disparate classes. The deviation from Eq.3 lies in that our input includes both seen and unseen samples, and $N_b$ symbolizes the OOD representation numbers within a batch. Our objective function is designed as follows:

$$\mathcal{L}_{od} = \frac{1}{N_b \times N_b} \sum_{i=1}^{N_b} \sum_{j=1}^{N_b} (\mathbb{1}_{y_i = y_j} \log(\sigma(b_{ij})) + \mathbb{1}_{y_i \neq y_j} \log(1 - \sigma(b_{ij}))). \tag{9}$$

## Model Optimization

**Training.** Our framework first undergoes a preprocessing phase in which the teacher model is trained on real seen samples. The teacher model then extracts knowledge, either in the form of logits or as intermediate features. This extracted knowledge is subsequently used to guide the training of the student model during the distillation process. The cross-entropy loss in the preprocessing phase is as follows:

$$\mathcal{L}_{pre}(x, y) = -\sum_{i=1}^{S} y^{(i)} \log \frac{\exp(\phi(x)^{(i)})/\tau_o)}{\sum_{k=1}^{S} \exp(\phi(x)^{(k)})/\tau_o)}. \tag{10}$$

After completing the pre-training of the OOD detection model, we jointly train the FG, ID$^2$SD and O$^2$DBD end-to-end. We utilize the real seen samples $x'$ and the unseen samples $x''$ generated by the FG as inputs for the ID$^2$SD module. Then, we calculate the OOD confidence labels from the output of the teacher network, and map the softmax probability of student network to the OOD representation embedding space. Thus, the total loss of D$^3$GZSL is formulated as:

$$\min_{G, E_s, C_s, H} \mathcal{L}_{gen} + \lambda(\mathcal{L}_{id} + \mathcal{L}_{od}), \tag{11}$$

| Method | AWA1 | | | AWA2 | | | CUB | | | FLO | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | U | S | H | U | S | H | U | S | H | U | S | H |
| cycle-CLSWGAN (Felix et al. 2018) | 56.9 | 64.0 | 60.2 | - | - | - | 45.7 | 61.0 | 52.3 | 59.2 | 72.5 | 65.1 |
| CADA-VAE (Schonfeld et al. 2019) | 57.3 | 72.8 | 64.1 | 55.8 | 75.0 | 63.9 | 51.6 | 53.5 | 52.4 | - | - | - |
| LisGAN (Li et al. 2019) | 52.6 | 76.3 | 62.3 | - | - | - | 46.5 | 57.9 | 51.6 | 57.7 | 83.8 | 68.3 |
| IZF (Shen et al. 2020) | 61.3 | **80.5** | 69.6 | 60.6 | 77.5 | 68.0 | 52.7 | 68.0 | 59.4 | - | - | - |
| SE-GZSL (Kim, Shim, and Shim 2022) | 61.3 | 76.7 | 68.1 | 59.9 | 80.7 | 68.8 | 53.1 | 60.3 | 56.4 | - | - | - |
| TDCSS (Feng et al. 2022) | 54.4 | 69.8 | 60.9 | 59.2 | 74.9 | 66.1 | 44.2 | 62.8 | 51.9 | 54.1 | 85.1 | 66.2 |
| DUET (Chen et al. 2023) | - | - | - | 63.7 | **84.7** | **72.7** | 62.9 | **72.8** | 67.5 | - | - | - |
| GKU (Guo et al. 2023) | - | - | - | - | - | - | 52.3 | 71.1 | 60.3 | - | - | - |
| f-CLSWGAN (Xian et al. 2018b) | 57.9 | 61.4 | 59.6 | - | - | - | 43.7 | 57.7 | 49.7 | 59.0 | 73.8 | 65.6 |
| **f-CLSWGAN+D$^3$GZSL** | 57.1 | 69.8 | 62.8 | - | - | - | 52.3 | 61.5 | 56.5 | 61.1 | **86.7** | 71.7 |
| TF-VAEGAN (Narayan et al. 2020) | - | - | - | 59.8 | 75.1 | 66.6 | 52.8 | 64.7 | 58.1 | 62.5 | 84.1 | 71.7 |
| **TF-VAEGAN+D$^3$GZSL** | - | - | - | 60.2 | 74.9 | 66.8 | 57.3 | 64.5 | 60.7 | 65.6 | 81.4 | 72.7 |
| DDGAN | 58.1 | 63.5 | 60.7 | 61.7 | 68.4 | 64.9 | 47.5 | 61.5 | 53.6 | 61.1 | 85.2 | 71.2 |
| **DDGAN+D$^3$GZSL** | 59.5 | 68.3 | 63.6 | 62.9 | 67.7 | 65.2 | 54.2 | 59.7 | 56.8 | 63.4 | 82.1 | 71.5 |
| CE-GZSL (Han et al. 2021) | 65.3 | 73.4 | 69.1 | 63.1 | 78.6 | 70.0 | 63.9 | 66.8 | 65.3 | **69.0** | 78.7 | 73.5 |
| **CE-GZSL+D$^3$GZSL** | **65.7** | 76.2 | **70.5** | **64.6** | 76.7 | 70.1 | **66.7** | 69.1 | **67.8** | 68.6 | 80.9 | **74.2** |

Table 1: Comparisons with state-of-the-art GZSL methods and baseline generative ZSL methods. D$^3$GZSL represents the use of our D$^3$GZSL framework based on the baseline model. $U$ and $S$ are the Top-1 accuracy of the unseen and seen classes, respectively. $H$ is the harmonic mean of $U$ and $S$. The best and second best results are marked in bold and underline, respectively.

where $\lambda$ is the hyper-parameters indicating the effect of $\mathcal{L}_{id}$ and $\mathcal{L}_{od}$ towards the generator.

**Inference.** We no longer train a separate classifier for classification. Once the training is completed, we map the $\mathcal{D}_{te} = \{x_i, y_i\}_{i=N+1}^{N+M}$ to the embedding space using the embedding function $E_s$ of the student network. Then, we employ the classifier $C_s$ to predict the class label $\hat{y}$:

$$\hat{y} = \arg\max_i \frac{\exp(\psi(x)^{(i)})}{\sum_{k=1}^{S+U} \exp(\psi(x)^{(k)})}. \tag{12}$$

# Experiment

**Datasets.** We perform experiments on four ZSL benchmark datasets that are widely used: the Animals with Attributes1&2 (AWA1 (Lampert, Nickisch, and Harmeling 2013) & AWA2 (Xian et al. 2018a)) dataset, Caltech-UCSD Birds-200-2011 (CUB (Wah et al. 2011)) dataset, and Oxford Flowers (FLO (Nilsback and Zisserman 2008)) dataset.

**Evaluation Protocols.** We evaluate the top-1 accuracy separately on both seen classes ($S$) and unseen classes ($U$) in the generalized zero-shot learning (GZSL). We also use the harmonic mean of these two accuracies ($H = (2 \times S \times U)/(S + U)$) as a performance measure for GZSL.

**Implementation Details.** We set the embedding dimension $z$ to 2048 on all datasets. The classifier $C_s$ outputs logits on all classes, and the classifier $C_o$ outputs logits on seen classes. The projector $H$ maps softmax probabilities into a two-dimensional space that encodes both ID and OOD information. The input noise dimension $w$ in the generator is equal to that of the corresponding attributes. In batch distillation, instances of the same class within a batch serve as positive samples for each other, while those of different class are treated as negative samples. Here are some of the parameter settings when employing f-CLSWGAN as the baseline model. We set batch size of 4096 for AWA1, 256 for CUB, 512 for FLO. The number of generated samples for each unseen category is as follows: 200 for AWA1, 5 for CUB, and 30 for FLO. We empirically set the loss weights $\lambda = 0.0001$ for AWA1, CUB. We set $\lambda = 0.001$ for FLO.

## Comparisons with Previous Methods

In Table 1, we applied our framework to three previous baseline methods and the DDGAN which is a new generative method of ZSL we introduced to demonstrate the improvement of our framework on diffusion models. The results show that we achieved improvements on the AWA1, AWA2, CUB, and FLO datasets. The most significant improvement was observed with the f-CLSWGAN method, with increases of 3.2% on the AWA1 dataset, 6.8% on the CUB dataset, and 6.1% on the FLO dataset. The best-performing dataset was CUB, with improvements of 6.8%, 2.6%, 2.5%, and 3.2% on the four baseline methods, respectively.

Within the CE-GZSL baseline, our $H$ metric delivered the top performance on the AWA1, CUB, and FLO datasets, while ranking second on the AWA2 dataset, only surpassed by DUET (Chen et al. 2023). Strikingly, when compared to the $S$ metric, our $U$ metric demonstrated a substantial improvement, achieving the highest scores on AWA1, AWA2 and CUB, with 65.7%, 64.6% and 66.7% respectively. It also secured the second-highest performance on the FLO dataset, recording 68.6%. The experimental results demonstrate that aligning the distribution of generated samples with that of real samples through out-of-distribution detection is an effective method to address seen bias.

## Ablation Study

**Training Strategy Analysis.** In this section, we compare the outcomes of three different experimental groups. Table 2 shows the comparison results. The first comprises our proposed one-stage end-to-end training method. The second involves a Two-Stage (TS) classification method based on OOD detection. The third represents an idealized version of the Two-Stage (IV-TS) classification method based on OOD detection. This idealized experiment is designed to simulate

| Method | IV-TS | TS | Ours | AWA1 | | | AWA2 | | | CUB | | | FLO | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | U | S | H | U | S | H | U | S | H | U | S | H |
| f-CLSWGAN | ✓ | | | 67.4 | 88.9 | - | - | - | - | 56.1 | 70.0 | - | 65.2 | 87.4 | - |
| | | ✓ | | **58.9** | **71.8** | **64.7** | - | - | - | 39.7 | 46.1 | 42.7 | 60.6 | 62.3 | 61.4 |
| | | | ✓ | 57.1 | 69.8 | 62.8 | - | - | - | **52.3** | **61.5** | **56.5** | **61.1** | **86.7** | **71.7** |
| CE-GZSL | ✓ | | | 69.2 | 88.3 | - | 69.7 | 91.8 | - | 78.7 | 73.5 | - | 68.1 | 89.4 | - |
| | | ✓ | | 57.5 | 72.3 | 64.1 | 55.5 | 76.7 | 64.4 | 61.1 | 42.8 | 50.3 | 58.3 | 70.8 | 64.0 |
| | | | ✓ | **65.7** | **76.2** | **70.5** | **64.6** | 76.7 | **70.1** | **66.7** | **69.1** | **67.8** | 68.6 | **80.9** | **74.2** |

Table 2: The performance comparison of our proposed D$^3$GZSL framework (our), the two-stage classification method based on OOD detection (TS) and idealized version of the two-stage classification method based on OOD detection (IV-TS).

| Datasets | Baseline | ID$^2$SD | O$^2$DBD | U | S | H |
|---|---|---|---|---|---|---|
| CUB | ✓ | × | × | 50.4 | 59.8 | 54.7 |
| | ✓ | ✓ | × | 49.4 | **63.8** | 55.6 |
| | ✓ | × | ✓ | 51.3 | 59.6 | 55.1 |
| | ✓ | ✓ | ✓ | **52.3** | 61.5 | **56.5** |
| FLO | ✓ | × | × | 58.6 | 83.3 | 68.8 |
| | ✓ | ✓ | × | **61.8** | 84.3 | 71.3 |
| | ✓ | × | ✓ | 59.9 | 86.3 | 70.7 |
| | ✓ | ✓ | ✓ | 61.1 | **86.7** | **71.7** |

Table 3: The baseline model includes the FG and the classification loss $\mathcal{L}_{cls}$. ID$^2$SD indicates the use of $\mathcal{L}_{be}$ and $\mathcal{L}_{kl}$ losses, while O$^2$DBD represents the use of $\mathcal{L}_{od}$ loss.

| Method | CUB | | | FLO | | |
|---|---|---|---|---|---|---|
| | U | S | H | U | S | H |
| Baseline | 43.7 | 57.7 | 49.7 | 59.0 | 73.8 | 65.6 |
| Energy | 49.1 | **62.3** | 54.9 | 60.3 | **87.2** | 71.3 |
| Softmax | **52.3** | 61.5 | **56.5** | **61.1** | 86.7 | **71.7** |

Table 4: The performance of our framework is demonstrated under various OOD detection methods, employing f-CLSWGAN as the baseline model in our framework.

the performance of seen and unseen expert classifiers under conditions where the process of OOD detection can classify seen and unseen samples with complete accuracy. In an ideal scenario, the results of the domain expert classifiers surpass those of our method. In fact, the OOD detection in TS cannot achieve a 100% accuracy rate. Some data that is not within the training distribution is mistakenly assigned to the expert classifiers for classification. This results in the performance of the TS approach being inferior to that of our proposed method because of the compounding of errors from OOD detection and the expert classifiers (error accumulation).

**Component Analysis.** Here, we set up an ablation study to examine the impact of various components on our D$^3$GZSL framework. The baseline model includes the FG and the classification loss $\mathcal{L}_{cls}$. We conducted three sets of experiments on two benchmark datasets to validate the individual and combined effects of our ID$^2$SD and O$^2$DBD modules. Through the experimental results presented in Table 3, we can draw the following conclusions: (1) Employing modules ID$^2$SD and O$^2$DBD separately has led to an enhancement in our performance over the baseline method. (2) When the two modules operate in conjunction, there is a marked enhancement in performance on the $H$ metric. This suggests that our framework is effective in reducing the discrepancy between the distribution of generated samples and the distribution of real samples. It accomplishes this by optimizing both the in-distribution and out-of-distribution aspects, thereby creating a more cohesive alignment between the two distributions.

**OOD Scoring Strategy Analysis.** In this paper, we experimented with two different architectures to verify the impact of using different OOD detection methods on our framework. We conducted experiments using two methods, Softmax score (Hendrycks and Gimpel 2016) and Energy (Liu et al. 2020), on the f-CLSWGAN baseline model. The experimental results in Table 4 showed that no matter which architecture was used, the performance was significantly improved. This provides strong evidence for our framework, demonstrating its effective adaptability and scalability, and its ability to be successfully applied to different architectures.

## Conclusion

In this paper, we introduce a generative GZSL framework(D$^3$GZSL) that combines OOD detection and knowledge distillation technologies. Our D$^3$GZSL leverages the OOD detection model to distill the student model, effectively aligning the distribution of generated samples more closely with the distribution of actual samples. By training a unified classifier as the final GZSL classifier, our framework addresses the issue of accumulated error stemming from two-stage classification in previous ZSL methods based on OOD detection. Empirical validation through comprehensive experiments demonstrates that our hybrid D$^3$GZSL framework consistently enhances the performance of existing generative GZSL approaches. This novel approach not only leverages cutting-edge techniques but also addresses the limitations of traditional ZSL methodologies, propelling the field towards more precise and reliable zero-shot learning outcomes.

## Acknowledgements

Reported research is partly supported by the National Natural Science Foundation of China under Grant 62176030, and the Natural Science Foundation of Chongqing under Grant cstc2021jcyj-msxmX0568.

## References


Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Atzmon, Y.; and Chechik, G. 2019. Adaptive confidence smoothing for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 52–68. Springer.

Chen, D.; Mei, J.-P.; Zhang, Y.; Wang, C.; Wang, Z.; Feng, Y.; and Chen, C. 2021. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7028–7036.

Chen, G.; Choi, W.; Yu, X.; Han, T.; and Chandraker, M. 2017. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30.

Chen, X.; Lan, X.; Sun, F.; and Zheng, N. 2020. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, 572–588. Springer.

Chen, Z.; Huang, Y.; Chen, J.; Geng, Y.; Zhang, W.; Fang, Y.; Pan, J. Z.; and Chen, H. 2023. Duet: Cross-modal semantic grounding for contrastive zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 405–413.

Felix, R.; Reid, I.; Carneiro, G.; et al. 2018. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 21–37.

Felix, R.; Sasdelli, M.; Reid, I.; and Carneiro, G. 2019. Multi-modal ensemble classification for generalized zero shot learning. *arXiv preprint arXiv:1901.04623*.

Feng, Y.; Huang, X.; Yang, P.; Yu, J.; and Sang, J. 2022. Non-generative generalized zero-shot learning via task-correlated disentanglement and controllable samples synthesis. In *CVPR*.

Guo, J.; Guo, S.; Zhou, Q.; Liu, Z.; Lu, X.; and Huo, F. 2023. Graph knows unknowns: Reformulate zero-shot learning as sample-level graph recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7775–7783.

Han, Z.; Fu, Z.; Chen, S.; and Yang, J. 2021. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2371–2381.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, R.; Geng, A.; and Li, Y. 2021. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34: 677–689.

Huang, S.; Lin, J.; and Huangfu, L. 2020. Class-prototype discriminative network for generalized zero-shot learning. *IEEE Signal Processing Letters*, 27: 301–305.

Keshari, R.; Singh, R.; and Vatsa, M. 2020. Generalized zero-shot learning via over-complete distribution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13300–13308.

Kim, J.; Shim, K.; and Shim, B. 2022. Semantic feature extraction for generalized zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1166–1173.

Kodirov, E.; Xiang, T.; and Gong, S. 2017. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3174–3183.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9306–9315.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, 951–958. IEEE.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3): 453–465.

Lee, K.; Lee, H.; Lee, K.; and Shin, J. 2017. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Li, J.; Jing, M.; Lu, K.; Ding, Z.; Zhu, L.; and Huang, Z. 2019. Leveraging the invariant side of generative zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7402–7411.

Liang, S.; Li, Y.; and Srikant, R. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

Lin, G.; Chen, W.; Liao, K.; Kang, X.; and Fan, C. 2019. Transfer feature generating networks with semantic classes structure for zero-shot learning. *IEEE Access*, 7: 176470–176483.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475.

Liu, Y.; Zhou, L.; Bai, X.; Huang, Y.; Gu, L.; Zhou, J.; and Harada, T. 2021. Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3794–3803.

Mandal, D.; Narayan, S.; Dwivedi, S. K.; Gupta, V.; Ahmed, S.; Khan, F. S.; and Shao, L. 2019. Out-of-distribution detection for generalized zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9985–9993.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12664–12673.

Mishra, A.; Krishna Reddy, S.; Mittal, A.; and Murthy, H. A. 2018. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2188–2196.

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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, 479–495. Springer.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729. IEEE.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. *Advances in neural information processing systems*, 22.

Romera-Paredes, B.; and Torr, P. 2015. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, 2152–2161. PMLR.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Shen, Y.; Qin, J.; Huang, L.; Liu, L.; Zhu, F.; and Shao, L. 2020. Invertible zero-shot recognition flows. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, 614–631. Springer.

Sun, Y.; Guo, C.; and Li, Y. 2021. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34: 144–157.

Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 20827–20840. PMLR.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Verma, V. K.; Arora, G.; Mishra, A.; and Rai, P. 2018. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4281–4289.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, L.; Huang, S.; Huangfu, L.; Liu, B.; and Zhang, X. 2022. Multi-label out-of-distribution detection via exploiting sparsity and co-occurrence of labels. *Image and Vision Computing*, 126: 104548.

Wang, W.; Zheng, V. W.; Yu, H.; and Miao, C. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2): 1–37.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018a. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9): 2251–2265.

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

Xian, Y.; Sharma, S.; Schiele, B.; and Akata, Z. 2019. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10275–10284.

Xiao, Z.; Kreis, K.; and Vahdat, A. 2021. Tackling the generative learning trilemma with denoising diffusion GANs. *arXiv preprint arXiv:2112.07804*.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Xing, Y.; Huang, S.; Huangfu, L.; Chen, F.; and Ge, Y. 2020. Robust bidirectional generative network for generalized zero-shot learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2022. Vgse: Visually-grounded semantic embeddings for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9316–9325.

Zagoruyko, S.; and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.