GenZSL: Generative Zero-Shot Learning Via Inductive Variational Autoencoder

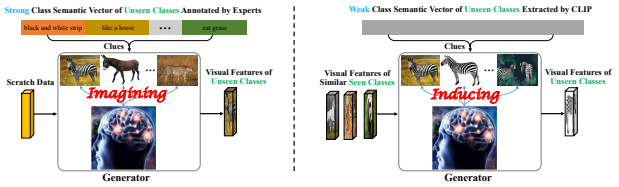
Shiming Chen [1] **Dingjie Fu** [2] **Salman Khan** [1 3] **Fahad Shahbaz Khan** [1 4]

Abstract

Remarkable progress in zero-shot learning (ZSL) has been achieved using generative models. However, existing generative ZSL methods merely generate (*imagine*) the visual features from scratch guided by the strong class semantic vectors annotated by experts, resulting in suboptimal generative performance and limited scene generalization. To address these and advance ZSL, we propose an inductive variational autoencoder for generative zero-shot learning, dubbed GenZSL. Mimicking human-level concept learning, GenZSL operates by *inducting* new class samples from similar seen classes using weak class semantic vectors derived from target class names (i.e., CLIP text embedding). To ensure the generation of informative samples for training an effective ZSL classifier, our GenZSL incorporates two key strategies. Firstly, it employs class diversity promotion to enhance the diversity of class semantic vectors. Secondly, it utilizes target class-guided information boosting criteria to optimize the model. Extensive experiments conducted on three popular benchmark datasets showcase the superiority and potential of our GenZSL with significant efficacy and efficiency over f-VAEGAN, e.g., 24.7% performance gains and more than 60× faster training speed on AWA2. Codes are available at https://github.com/shiming-chen/GenZSL.

1. Introduction

Zero-shot learning (ZSL) enables the recognition of unseen classes by transferring semantic knowledge from some seen classes to unseen ones (Palatucci et al., 2009; Lampert et al., 2009). Recently, generative models such as generative adversarial networks (GANs) (Goodfellow et al., 2014), variational autoencoders (VAEs) (Kingma & Welling, 2014), and normalizing flows (Dinh et al., 2017) have been successfully applied in ZSL, achieving significant performance improvements. These models synthesize images or visual features of unseen classes to alleviate the lack of samples for those classes (Arora et al., 2018; Xian et al., 2018; 2019b; Chen et al., 2021a; Narayan et al., 2020; Chen et al., 2021b).



(a) Existing Generative ZSL (b) Our GenZSL

Figure 1. Motivation illustration. (a) Existing generative ZSL methods merely generate (*imagine*) the visual features from scratch guided by the strong class semantic vectors, resulting in suboptimal generative performance and scene generalization. For example, the generator inevitably generates similar classes of "Zebra" or others, e.g., "Donkey". (b) Our GenZSL generates (*induces*) the reliable visual features of unseen classes from the similar seen classes with the clues of weak class semantic vector, e.g., from "Horse" to "Zebra".

Given that GAN architectures can generate higher-quality visual sample features, there's a growing trend in synthesizing features using GANs (Xian et al., 2018; 2019b; Chen et al., 2021a; Narayan et al., 2020; Yue et al., 2021; Chen et al., 2025). However, existing generative ZSL methods typically generate (*imagine*) visual features from scratch (e.g., Gaussian noises) guided by strong class semantic vectors annotated by expert (Xian et al., 2018; 2019b; Chen et al., 2021a; Narayan et al., 2020; Çetin et al., 2022; Chen et al., 2023a). This approach often fails to produce reliable feature samples and generalize to various scene tasks, as illustrated in Figure 1 (a). The shortcomings arise from: i) the generator learning from scratch without sufficient data to capture the high-dimensional data distribution, and ii) the reliance on strong class semantic vectors, which are time-consuming and labor-intensive to collect for various scene generations. Hence, there's a pressing need to explore novel

[1] Mohamed bin Zayed University of Artificial Intelligence [2]Huazhong University of Science and Technology [3]Australian National University [4]Linköping University. Correspondence to: Fahad Shahbaz Khan <fahad.khan@mbzuai.ac.ae; fahad.khan@liu.se>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

generative paradigms for ZSL.

Cognitive psychologist often frame the process of learning new concepts as "the problem of induction" (Carey, 1985; and, 2000). For instance, children typically induce novel concepts from a few familiar objects, guided by certain priors (Tenenbaum et al., 2011; Lake et al., 2015). Essentially, rich concepts can be induced "compositionally" from simpler primitives under a Bayesian criterion, and the model "learns to learn" by developing hierarchical priors that facilitate the learning of new concepts based on previous experiences with related concepts. These priors represent a learned inductive bias that abstracts the key regularities and dimensions of variation across both types of concepts and instances of a concept within a given domain. Following this paradigm, our objective is to devise a novel generative zero-shot learning (ZSL) model capable of generating (*inducing*) new/target classes based on samples from similar/referent seen classes. As illustrated in Figure 1 (b), our generative ZSL model can generate informative samples of new classes (e.g., "Zebra") by inducing them from referent seen classes (e.g., "Horse", "Tiger", and "Panda").

However, there are two challenges in targeting this goal. Firstly, addressing the issue of weak class semantic vectors. These vectors, extracted from sources like the CLIP text encoder (Radford et al., 2021), often lack specific class information, such as attributes, compared to vectors annotated by experts. As a result, they may not effectively guide generative methods. Furthermore, these vectors can be misaligned in the vision-language space. For instance, the text embedding of a class name might be close to embeddings of unrelated classes but distant from image embeddings (Hu et al., 2023; Tanwisuth et al., 2023; Khattak et al., 2023; Chen et al., 2024a). *How can we enhance the diversity of weak class semantic vectors to distinguish between various classes effectively, thereby avoiding the problem of generating visual features that are too similar to other classes?* Secondly, ensuring that a novel generative method evolves samples of referent classes into target classes with the guidance of weak class semantic vectors is equally challenging. This involves transforming samples of seen classes into samples that accurately represent unseen classes, guided only by the limited information provided by weak class semantic vectors. *How can we achieve this induction reliably and effectively within a generative ZSL framework?*

To guide the induction towards creating informative samples for training effective ZSL classifiers, we propose a novel inductive variational autoencoder for generative ZSL, namely **GenZSL**. GenZSL essentially considers two criteria, i.e., class diversity promotion and target class-guided information boosting. Specifically, we first deploy a class diversity promotion module to reduce redundant information from class semantic vectors by eliminating their major components. This process enables all class semantic vectors to become nearly perpendicular to each other but keep the origin relationships between all classes, thus enhancing the diversity among them. Then, we employ a semantically similar sample selection module to select the referent class samples for seen/unseen classes from seen classes. Finally, we design a target class-guided information boosting loss to guide the inductive variational autoencoder to synthesize the visual features belonging to target classes.

Our main contributions are summarized in the following: **i)** We propose an induction-based GenZSL for generative ZSL, which can synthesize the samples of unseen classes based on the weak class semantic vectors inducting from the similar seen classes. **ii)** We enable GenZSL to synthesize informative samples by designing the class diversity promotion, semantically similar sample selection, and inductive vatiational autoencoder modules. **iii)** We conduct extensive experiments on three wide-used ZSL benchmarks (e.g., CUB (Welinder et al., 2010), SUN (Patterson & Hays, 2012), and AWA2 (Xian et al., 2019a)), results demonstrate the significant efficacy and efficiency over the existing ZSL methods, e.g., 24.7% performance gains and more than $60\times$ faster training speed on AWA2. More importantly, our GenZSL can be flexibly extended on various scene tasks without the guidance of expert-annotated attributes.

2. Related Work

Zero-Shot Learning. Zero-shot learning is proposed to tackle the classification problem when some classes are unknown. To recognize the unseen classes, the side-information/semantic (e.g., attribute descriptions (Lampert et al., 2014), DNA information (Badirli et al., 2021)) is utilized to bridge the gap between seen and unseen classes. As such, the key task of ZSL is to conduct effective interactions between visual and semantic domains. Typically, there are two methodologies to target on this goal, i.e., embedding-based methods that learn visual→semantic mapping (Xian et al., 2016; Xu et al., 2020; Zhu et al., 2019; Wan et al., 2019; Han et al., 2022), and generative methods that learn semantic→visual mapping (Xian et al., 2019b; Chen et al., 2021a; Huynh & Elhamifar, 2020b; Çetin et al., 2022; Chen et al., 2020). Considering the semantic representations, embedding-based methods focus recently on learning the region-based visual features rather than the holistic visual features (Huynh & Elhamifar, 2020a; Xu et al., 2020; Chen et al., 2022a;b; 2024c). Since these methods learn the ZSL classifier only on seen classes, inevitably resulting in the models overfitting to seen classes. To tackle this challenge, generative ZSL methods employ the generative models (e.g., VAE, and GAN) to generate the unseen features for data augmentation, and thus ZSL is converted to a supervised classification task. As such, the generative ZSL methods

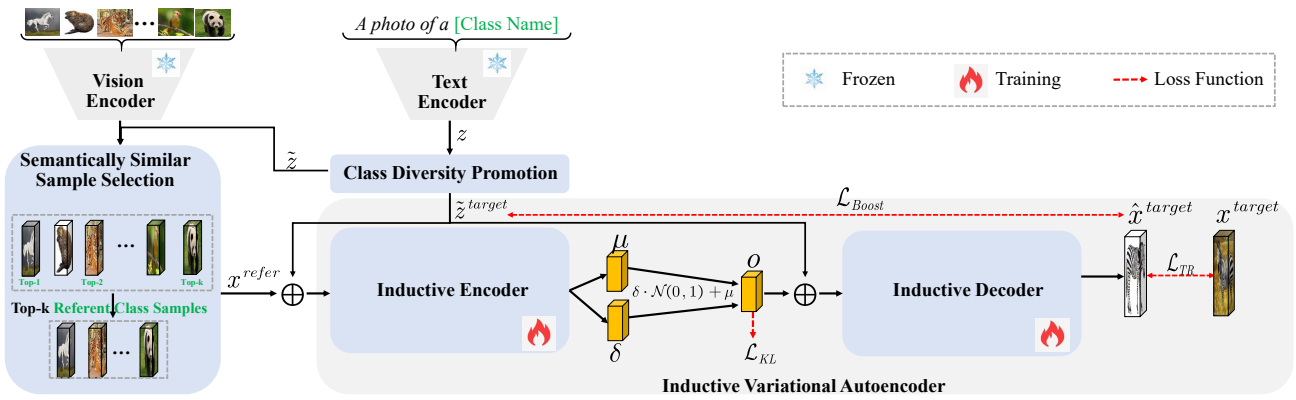


Figure 2. Pipeline of our GenZSL. GenZSL first takes class diversity promotion to reduce the redundant information from class semantic vectors, and to improve the identity for all class semantic vectors. Then, it employs a semantically similar sample selection module to select the top-k referent class from the seen classes for each target class as training inputs. Based on the referent samples, GenZSL learns an inductive variational autoencoder to create the new informative feature samples for unseen classes via induction optimized by target class-guided information boosting criteria.

have shown significant performance and become very popular recently. Furthermore, Li *et al.* (Li et al., 2023) introduces Stable Diffusion to perform zero-shot classification without any additional training by leveraging the ELBO as an approximate class-conditional log-likelihood.

However, existing generative ZSL methods simply imagine the visual feature from a Gaussian distribution with the guidance of a strong class semantic vector. Thus, they are limited in i) there lacks enough data for training a generative model to learn the high-dimension data distribution, resulting in undesirable generation performance; ii) they rely on the strong condition guidance (e.g., expert-annotated attributes) for synthesizing target classes, so they cannot easily generalize to various scenes. As such, we propose a novel generative method to create samples of unseen classes for advancing ZSL via induction rather than imagination.

Generative Model for Data Augmentation. Synthesizing new data using a generative model for data augmentation is a promising direction (Zhou et al., 2023; Jahanian et al., 2022). Many recent studies (Azizi et al., 2023; He et al., 2023) explored generative models to generate new data for model training. However, these methods fail to ensure that the synthesized data bring sufficient new information and accurate labels for the target small datasets. Because they imagine the new data from scratch (e.g., Gaussian distribution), which is infeasible with very limited/diverse training data. Zhang *et al.* (Zhang et al., 2023) introduce GIF to expanding small-scale datasets with guided imagination using pre-trained large-scale generative models, e.g., Stable Diffusion (Rombach et al., 2022) or DALL-E2 (Ramesh et al., 2022). Although GIF can expand a small dataset into a larger labeled one in a fully automatic manner without involving human annotators, it requires anchor samples for imagination. As such, these imagination-based generative models are not feasible for ZSL tasks. In contrast, we introduce a novel generative method to synthesize new informative data for ZSL via induction inspired by the human perception process (Carey, 1985; and, 2000).

3. Inductive Variational Autoencoder for ZSL

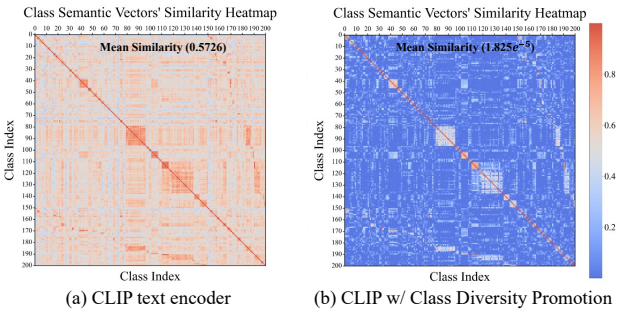
Problem Setting. The problem setting of ZSL and notations are defined in the following. Assume that data of seen classes $\mathcal{D}^s = \{(x_i^s, y_i^s)\}$ has C^s classes, where $x_i^s \in \mathcal{X}$ denotes the i-th visual feature, and $y_i^s \in \mathcal{Y}^s$ is the corresponding class label. $\mathcal{D}^s$ is further divided into training set $\mathcal{D}_{tr}^s$ and test set $\mathcal{D}_{te}^s$ following (Xian et al., 2019a). The unseen classes C^u has unlabeled samples $\mathcal{D}_{te}^u = \{(x_i^u, y_i^u)\}$, where $x_i^u \in \mathcal{X}$ are the visual samples of unseen classes, and $y_i^u \in \mathcal{Y}^u$ are the corresponding labels. Notably, $\mathcal{Y}^U \cap \mathcal{Y}^S = \varnothing$. A set of class semantic vectors of the class $c \in \mathcal{C}^s \cup \mathcal{C}^u = \mathcal{C}$ are extracted from CLIP text encoder, defined as z^c. In the conventional zero-shot learning (CZSL) setting, we learn a classifier only classifying unseen classes, i.e., $f_{CZSL} : \mathcal{X} \rightarrow \mathcal{Y}^U$, while we learn a classifier for both seen and unseen classes in the generalized zero-shot learning (GZSL) setting, i.e., $f_{GZSL} : \mathcal{X} \rightarrow \mathcal{Y}^U \cup \mathcal{Y}^S$.

Pipeline Overview. To enable the generative ZSL method to synthesize high-quality visual features with good scene generalization, we propose an inductive variational autoencoder for ZSL (namely GenZSL). Towards creating informative new samples for unseen classes, GenZSL considers two important criteria, i.e., class diversity promotion and target class-guided information boosting. As shown in Fig.

2, GenZSL first takes class diversity promotion to reduce the redundant information from class semantic vectors by removing their major components, enabling all class semantic vectors nearly perpendicular to each other. Based on the refined class semantic vectors, GenZSL employs a semantically similar sample selection module to select the top-k referent class from the seen classes for each target class. Subsequently, GenZSL learns the inductive variational autoencoder (IVAE) with the Kullback-Leibler divergence (KL) loss, target class reconstruction loss, and target class-guided information boosting loss, which ensures GenZSL inducts the target class samples from their similar class samples. After training, GenZSL takes IVAE to synthesize visual features of unseen classes to learn a supervised classifier.

3.1. Class Diversity Promotion

To avoid the ZSL model relying on the expert-annotated class semantic vectors, we adopt CLIP (Radford et al., 2021) text encoder to extract the class semantic vectors, i.e., text embedding of the class names. However, we observed that the CLIP text encoder fails to capture discriminative class information, especially on fine-grained datasets. As shown in Fig. 3(a), the class semantic vectors have high similarity with other classes, that is, all class semantic vectors are highly adjacent to ones of other classes. If we directly take such class semantic vectors as conditions to guide GenZSL, it inevitably causes the synthesized visual features confusion as the class semantic vectors with limited diversity.



(a) CLIP text encoder (b) CLIP w/ Class Diversity Promotion

Figure 3. Class semantic vectors' similarity heatmaps are extracted by CLIP text encoder and CLIP with class diversity promotion on the CUB dataset. The similarity heatmaps on SUN and AWA2 are presented in Appendix B.

As such, we introduce class diversity promotion (CDP) to improve the diversity of class semantic vectors. CDP reduces the redundant information from class semantic vectors by removing their major components, enabling all class semantic vectors nearly perpendicular to each other but to keep the original class relationships. Specifically, we take Singular Value Decomposition to get the orthonormal basis of the span of class semantic vectors $Z = [z^1, z^2, \cdots, z^C]$, i.e., $U, S, V = svd(Z)$, where $U = [e^1, e^2, \cdots, e^C]$ is the orthonormal basis. As suggested in Principal Component Analysis, the first dimension e^1 of the outer-space basis U will be the major component, which overlaps on most class semantic vectors $[z^1, z^2, \cdots, z^C]$. We directly remove the major component e^1 to define the new projection matrix $P = U^{'}U^{'\top}$ with $U^{'} = [e^2, e^3, \cdots, e^C]$. Accordingly, we obtain the refined class semantic vectors, formulated as:

$$\tilde{Z} = P \cdot Z = \{\tilde{z}^1, \tilde{z}^2, \cdots, \tilde{z}^C\} \tag{1}$$

As shown in Fig. 3(b), we make the refined class semantic vectors nearly perpendicular to each other, such as the mean similarity between various classes drops from 0.5726 to $1.825e^{-5}$ on the CUB. Meanwhile, CDP preserves the original relationships of classes, and thus it will not destroy the class semantics. As such, the refined class semantic vectors will be the significant conditions for induction.

3.2. Semantically Similar Sample Selection

In this paper, we are interested in semantically similar samples as they can serve as reliable known data for inducing new samples of other similar classes. Specifically, we select the semantically similar samples in seen classes (defined a referent class samples) with respect to the target seen/unseen classes c^{target} during training/testing, respectively. According to the cosine similarity, we define similar samples as the referent ones whose class semantic vectors $\tilde{z}^{c^s}$ is top-k closed to the target class semantic vectors $\tilde{z}^{target}$, formulated as:

$$c^{refer} = \arg \max_{\text{top}-k(c^s)} \frac{\tilde{z}^{target} \times \tilde{z}^{c^s}}{\|\tilde{z}^{target}\| \cdot \|\tilde{z}^{c^s}\|}, \tag{2}$$

where k is the number of referent classes with respect to the corresponding target classes. Accordingly, we can obtain a set of referent samples $x^{refer} = x^{c^{refer}}$ of the target seen/unseen classes from seen classes for training/testing.

3.3. Inductive Variational Autoencoder

Network Components. Our GenZSL aims to generate informative new samples for novel classes by inducing from seen classes. To achieve this, we devise a novel generative model called the inductive variational autoencoder (IVAE). We formulate the induction of new samples for target classes $\hat{x}$ from reference samples x^{refer} as $\hat{x} = IVAE(x^{refer} + o, \tilde{z}^{target})$, where o represents the perturbation applied to x^{refer} to enable IVAE to variationally generate $\hat{x}$ distinct from x^{refer}.

Specifically, IVAE consists of an inductive encoder (IE) and an inductive decoder (ID). The IE and ID are the Multi-Layer Perceptron (MLP) networks. The IE encodes the referent samples x^{refer} into latent space o conditioned by the

target class semantic vectors $\tilde{z}^{target}$, i.e., $o = \delta \cdot \mathcal{N}(0,1) + \mu$, where $\mu, \delta = IE(x^{refer}, \tilde{z}^{target})$. Subsequently, The ID further comprises hidden layers with a progressively larger number of nodes that decode the latent features to be a reconstruction of the target classes samples x^{target} guided by $\tilde{z}^{target}$, formulated as $\hat{x} = ID(o, \tilde{z}^{target})$. This is different to VAE which ultimately reconstructs the data back to its original input x^{refer}.

Network Optimization. Similar to the conditional VAE (Sohn et al., 2015), our IVAE includes the KL loss $\mathcal{L}_{KL}$ and the target class reconstruction loss $\mathcal{L}_{TR}$, formulated as:

$$\begin{aligned}\mathcal{L}_{IVAE} &= \mathcal{L}_{KL} - \mathcal{L}_{TR} \\ &= KL(q(o \mid x, \tilde{z}^{target}) \| p(o \mid \tilde{z}^{target})) \\ &- \mathbb{E}_{q(o|x^{refer}, \tilde{z}^{target})}[\log p(x^{target} \mid o, \tilde{z}^{target})],\end{aligned} \tag{3}$$

where $q(o \mid x, \tilde{z}^{target})$ is modeled by $IE(x^{refer}, \tilde{z}^{target})$, $p(o \mid \tilde{z}^{target})$ is assumed to be $\mathcal{N}(0,1)$, and $p(x^{target} \mid o, \tilde{z}^{target})$ is represented by $ID(o, \tilde{z}^{target})$. Essentially, $\mathcal{L}_{TR}$ towards the target class-guided information boosting criteria in vision-level, encouraging IVAE to synthesize high-quality target class samples.

To ensure IVAE evolves the referent samples to belong to target classes, GenZSL further employs a target class-guided information boosting loss $\mathcal{L}_{Boost}$ for optimization. Considering CLIP's full prior knowledge, $\mathcal{L}_{Boost}$ aims to improve the information entropy between the synthesized visual features of target classes $\hat{x}^{target}$ and their corresponding class semantic vectors $\tilde{z}^{target}$, formulated as:

$$\mathcal{L}_{Boost} = -\frac{\exp\left(< \hat{x}^{target}, \tilde{z}^{target} > /\tau\right)}{\sum_{j=1}^{C^s} \exp\left(< \hat{x}^{target}, \tilde{z}^{target} > /\tau\right)}, \tag{4}$$

where τ is the temperature parameter and set to 0.07, $< \cdot, \cdot >$ denotes the similarity between the two elements. Indeed, $\mathcal{L}_{Boost}$ and $\mathcal{L}_{TR}$ cooperatively ensure IVAE to synthesize desirable target class samples from semantic- and vision-level, respectively.

As such, the total optimization loss function can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_{IVAE} + \lambda \mathcal{L}_{Boost}, \tag{5}$$

where λ is a weight to control the $\mathcal{L}_{Boost}$, enabling model optimization to be more effective.

3.4. ZSL Classification

After training, we first take the pre-trained IVAE to synthesize visual features for unseen classes:

$$\hat{x}^u = ID(o, \tilde{z}^{c^u}),$$
$$\text{where} \quad o = \delta \cdot \mathcal{N}(0,1) + \mu, and \quad \mu, \delta = IE(x^{refer}, \tilde{z}^{c^u}). \tag{6}$$

Different from the standard VAEs that synthesize samples from scratch (e.g., Gaussian noise), we synthesize the visual features of unseen classes inducting from referent seen class samples and take Gaussian noise as variations. As such, our GenZSL can more easily create informative new samples for unseen classes.

Then, we take the synthesized unseen visual features and the real visual features of seen classes $x^s \in \mathcal{D}^s_{tr}$ to learn a classifier (*e.g.*, softmax), *i.e.*, $f_{czsl} : \mathcal{X} \rightarrow \mathcal{Y}^s$ in the CZSL setting and $f_{gzsl} : \mathcal{X} \rightarrow \mathcal{Y}^s \cup \mathcal{Y}^u$ in the GZSL setting. Once the classifier is trained, we use the real sample in the test set $\mathcal{D}^u_{te}$ to test the model further. The details of the testing process are shown in Appendix A.

4. Experiments

Datasets. We evaluate our GenZSL on three well-known ZSL benchmark datasets, i.e., two fine-grained datasets (CUB (Welinder et al., 2010) and SUN (Patterson & Hays, 2012)) and one coarse-grained dataset (AWA2 (Xian et al., 2019a)). CUB has 11,788 images of 200 bird classes (seen/unseen classes = 150/50). SUN contains 14,340 images of 717 scene classes (seen/unseen classes = 645/72). AWA2 consists of 37,322 images of 50 animal classes (seen/unseen classes = 40/10).

Evaluation Protocols. During testing, we adopt the unified evaluation protocols following (Xian et al., 2019a). The top-1 accuracy of the unseen class (denoted as $\boldsymbol{acc}$) is used for evaluating the CZSL performance. In the GZSL setting, the top-1 accuracy on seen and unseen classes is adopted, denoted as $\boldsymbol{S}$ and $\boldsymbol{U}$, respectively. Meanwhile, their harmonic mean (defined as $\boldsymbol{H} = (2 \times \boldsymbol{S} \times \boldsymbol{U})/(\boldsymbol{S} + \boldsymbol{U})$) is a better protocols in the GZSL.

Implementation Details. We use the training splits proposed in (Xian et al., 2018). Meanwhile, the visual features with 512 dimensions are extracted from the CLIP vision encoder (Radford et al., 2021). The IE and ID are the MLP networks. The specific network settings are $fc(512) - fc(1024) - fc(2048) - ReLu$ and $fc(512) - fc(1024) - fc(2048) - ReLu - fc(512)$ for IE and ID, respectively. We synthesize 1600, 800, and 5000 features per unseen class to train the classifier for CUB, SUN, and AWA2 datasets, respectively. We empirically set the loss weight λ as 0.1 for CUB and AWA2, and 0.001 for SUN. The top-2 similar classes serve as the referent classes for inductions on all datasets. Furthermore, to enlarge the reference of the referent samples for effective model training, we take mixup technique (Zhang et al., 2018) to randomly fuse the samples of various referent classes for data augmentation, i.e., $x^{refer} = 0.8 \cdot x^{c^{top-1}} + 0.2 \cdot x^{c^{top-2}}$. All experiments are performed on a single NVIDIA RTX 3090

Table 1. State-of-the-art comparisons for generative ZSL methods on CUB, SUN, and AWA2 under the GZSL settings. The best and second-best results are marked in **Red** and **Blue**, respectively. † denotes methods use CLIP visual features.

Methods	CUB			SUN			AWA2		
	U	S	H	U	S	H	U	S	H
CADA-VAE (Schönfeld et al., 2019)	51.6	53.5	52.4	47.2	35.7	40.6	55.8	75.0	63.9
f-VAEGAN (Xian et al., 2019b)	48.7	58.0	52.9	45.1	38.0	41.3	57.6	70.6	63.5
LisGAN (Li et al., 2019)	46.5	57.9	51.6	42.9	37.8	40.2	52.6	76.3	62.3
LsrGAN (Vyas et al., 2020)	48.1	59.1	53.0	44.8	37.7	40.9	54.6	74.6	63.0
IZF-NBC (Shen et al., 2020)	44.2	56.3	49.5	–	–	–	58.1	76.0	65.9
AGZSL (Chou et al., 2021)	48.3	58.9	53.1	29.9	40.2	34.3	65.1	78.9	71.3
HSVA (Chen et al., 2021b)	52.7	58.3	55.3	48.6	39.0	43.3	59.3	76.6	66.8
ICCE (Kong et al., 2022)	–	–	–	–	–	–	**65.3**	82.3	72.8
SCE-GZSL (Han et al., 2022)	–	–	–	45.9	41.7	43.7	64.3	77.5	70.3
FREE+ESZSL (Çetin et al., 2022)	51.6	60.4	55.7	48.2	36.5	41.5	51.3	78.0	61.8
CLSWGAN + DSP (Chen et al., 2023a)	51.4	63.8	56.9	48.3	**43.0**	**45.5**	60.0	86.0	70.7
ViFR (Chen et al., 2025)	**57.8**	62.7	**60.1**	**48.8**	35.2	40.9	58.4	81.4	68.0
f-VAEGAN† (Xian et al., 2019b)	22.5	**82.2**	35.3	–	–	–	61.2	**95.9**	**74.7**
TF-VAEGAN† (Narayan et al., 2020)	21.1	**84.4**	34.0	–	–	–	43.7	**96.3**	60.1
GenZSL	**53.5**	61.9	**57.4**	**50.6**	**43.8**	**47.0**	**86.1**	88.7	**87.4**

with 24G memory. We employ Pytorch to implement our experiments.

Table 2. State-of-the-art comparisons for ZSL methods on SUN and AWA2 under the CZSL setting. Embedding-based methods are categorized as †, and generative methods are categorized as ‡. $*$ denotes ZSL methods using the ViT visual features. The best and second-best results are marked in **Red** and **Blue**, respectively.

	Methods	SUN	AWA2
		acc	acc
†	APN (Xu et al., 2020)	62.6	66.8
	DAZLE (Huynh & Elhamifar, 2020a)	59.4	67.9
	GEM-ZSL (Liu et al., 2021)	62.8	67.3
	TransZero (Chen et al., 2022a)	65.6	70.1
	MSDN (Chen et al., 2022b)	65.8	70.1
	ICIS (Christensen et al., 2023)	51.8	64.6
	DUET* (Chen et al., 2023b)	64.4	69.9
	I2MVFormer-Wiki* (Naeem et al., 2023)	–	79.6
	HAS (Chen et al., 2023c)	63.2	71.4
	I2DFormer+* (Naeem et al., 2024)	–	77.3
	EG-GZSL (Chen et al., 2024d)	69.5	77.6
	ZSLViT* (Chen et al., 2024c)	68.3	70.7
	CVsC* (Chen et al., 2024b)	**71.5**	73.1
‡	CLSWGAN (Xian et al., 2018)	60.8	68.2
	f-VAEGAN (Xian et al., 2019b)	64.7	71.1
	CADA-VAE (Schönfeld et al., 2019)	61.7	63.0
	LisGAN (Li et al., 2019)	61.7	70.6
	IZF-NBC (Shen et al., 2020)	63.0	71.9
	LsrGAN (Vyas et al., 2020)	62.5	66.4
	HSVA (Chen et al., 2021b)	63.8	70.6
	GG (Cavazza et al., 2023)	62.7	70.1
	f-VAEGAN+DSP (Chen et al., 2023a)	68.6	71.6
	VADS (Hou et al., 2024)	–	**82.5**
	GenZSL (Ours)	**73.5**	**92.2**

4.1. Comparisons with State-of-the-Art Methods

We first compare our GenZSL with the various imagination-based generative ZSL methods (e.g., VAE, GAN, VAEGAN, and normalizing flow) under the GZSL setting. Table 1 shows the evaluation results on three datasets. Our GenZSL consistently achieves the best results with the $\boldsymbol{H}$ of 47.0% and 87.4% on SUN and AWA2, and second best results with the $\boldsymbol{H}$ of 57.4% on CUB, respectively. Notably, our GenZSL relies solely on weak class semantic vectors, while the compared methods utilize strong ones annotated by experts. Furthermore, when imagination-based generative ZSL methods (e.g., f-VAEGAN (Xian et al., 2019b) and TF-VAEGAN(Narayan et al., 2020)) using CLIP visual and semantic features, GenZSL still obtains improvements of $\boldsymbol{H} = 22.1\%/1.5\%/16.7\%$ on CUB/SUN/AWA2, respectively. This indicates that GenZSL is more adaptable to generalizing across various scenes. These results consistently demonstrate our induction-based GenZSL is a desirable generative paradigm for ZSL.

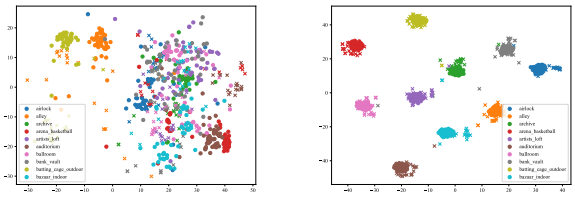
We also take our GenZSL to compare with the state-of-the-art ZSL methods under the CZSL setting, including the embedding-based methods and generative methods. Results are shown in Table 2. Compared to the embedding-based methods, our GenZSL consistently achieves the best performance on SUN and AWA2. When taking our GenZSL to compare with the imagination-based generative methods, GenZSL performs best results of $\boldsymbol{acc}$=73.5% and $\boldsymbol{acc}$=92.2% on SUN and AWA2, respectively. Notably, our GenZSL obtains the performance gains by 20.3% at least on AWA2 over the imagination-based generative ZSL methods. Compared with other ViT visual features based methods (Naeem et al., 2024; Chen et al., 2024c;b), our GenZSL still obtains competitive performance gains. These competitive results demonstrate the superiority and potential of our induction-based generative method, which significantly synthesizes informative new samples for unseen classes.

Table 3. Results of ablation study for our GenZSL on CUB and AWA2.

Methods	CUB				AWA2			
	CZSL	GZSL			CZSL	GZSL		
	acc	U	S	H	acc	U	S	H
GenZSL w/o CDP	60.9	48.2	64.6	55.2	90.7	82.3	87.9	85.0
GenZSL w/o Selection	62.5	48.0	67.0	55.9	91.1	84.2	86.4	85.3
GenZSL w/o $\mathcal{L}_{TR}$	48.3	20.1	37.5	26.2	87.5	39.9	83.1	53.9
GenZSL w/o $\mathcal{L}_{Boost}$	61.1	47.7	66.4	55.5	90.5	75.3	91.4	82.6
GenZSL w/o CDP&$\mathcal{L}_{Boost}$	60.0	42.5	69.3	52.7	87.7	89.0	75.3	81.6
GenZSL (full)	**63.3**	53.5	61.9	**57.4**	**92.2**	86.1	88.7	**87.4**

Table 4. Results of various models using weak class semantic vectors as side-information on CUB.

Methods	CUB		
	U	S	H
CLIP (Radford et al., 2021)	55.2	54.8	55.0
CoOp (Zhou et al., 2021)	49.2	63.8	55.6
CoOp + SHIP (Wang et al., 2023)	55.3	58.9	57.1
f-VAEGAN (Xian et al., 2019b)	22.5	82.2	35.3
TF-VAEGAN (Narayan et al., 2020)	21.1	84.4	34.0
GenZSL (Ours)	53.5	61.9	57.4



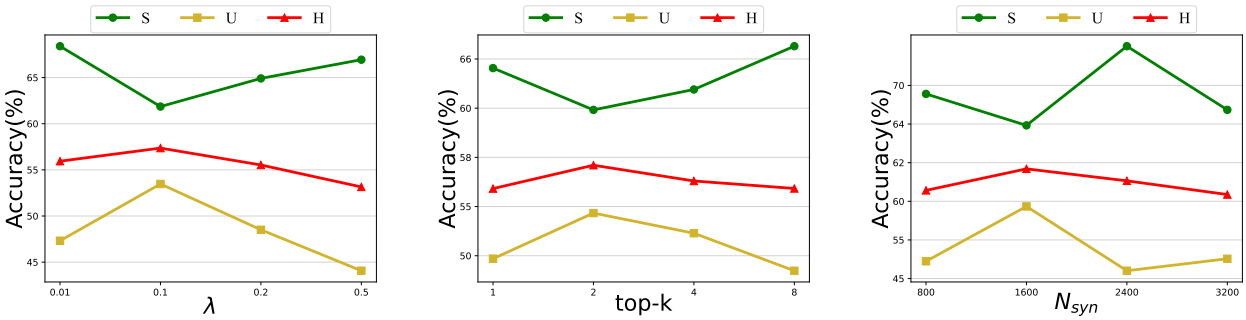
(a) f-VAEGAN (b) Our GenZSL

Figure 4. Qualitative evaluation with t-SNE visualization. The sample features from f-VAEGAN (Xian et al., 2019b) are shown on the left, and from our GenZSL are shown on the right. We use 10 colors to denote randomly selected 10 classes from SUN. The "×" and "o" are denoted as the real and synthesized sample features, respectively. The synthesized sample features and the real features distribute differently on the left while distributing similarly on the right. The visualization on the CUB and AWA2 is shown in Appendix D.

4.2. Ablation Study

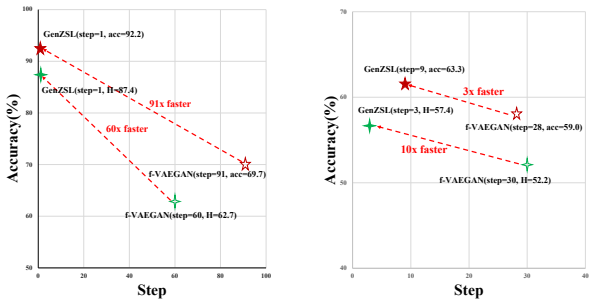
Various Model Components of Our GenZSL. To gain further insights into GenZSL, we conducted ablation studies to evaluate the effect of various model components, specifically class diversity promotion (CDP), semantically similar sample selection, target class reconstruction loss $\mathcal{L}_{TR}$, and target class-guided information boosting loss $\mathcal{L}_{Boosting}$, on the CUB and AWA2 datasets. The ablation results are summarized in Table 3. When GenZSL lacks CDP to consider class diversity criteria, there is a notable degradation in performance. This is attributed to the inability of class semantic vectors extracted from the CLIP text encoder to capture discriminative class information, resulting in weak diversity among class semantic vectors. Moreover, if GenZSL does not incorporate $\mathcal{L}_{TR}$ for target class information boosting, there is a significant drop in performance, with the harmonic mean decreasing by 30.8% and 33.5% on CUB and AWA2, respectively. These findings underscore the importance of $\mathcal{L}_{TR}$ as a fundamental loss for target class-guided information boosting, ensuring that our IVAE accurately induces referent samples to target class samples. Furthermore, $\mathcal{L}_{Boosting}$ enhances the induction process at the semantic level, complementing $\mathcal{L}_{TR}$. Semantically similar sample selection can slightly improve the performances of GenZSL, this means GenZSL is relatively robust in various source samples for model induction. Overall, these results demonstrate the effects of various components of GenZSL and underscore the significance of the two criteria for induction.

Various Models with Weak Class Semantic Vectors. We conducted a comparative analysis of various models utilizing weak class semantic vectors extracted from the CLIP text encoder. The results are presented in Table 4. Compared to large-scale visual-language methods (e.g., CLIP (Radford et al., 2021) and CoOp (Zhou et al., 2021)), our GenZSL demonstrates substantial improvements, indicating the effectiveness of our inductive generative paradigm as a desirable ZSL model. When imagination-based generative ZSL methods (e.g., f-VAEGAN (Xian et al., 2019b) and TF-VAEGAN (Narayan et al., 2020)) utilize weak class semantic vectors as side information, GenZSL achieves significant performance gains, with a minimum increase of 22.1% in harmonic mean over these methods. Additionally, we observed that when imagination-based generative ZSL methods use weak class semantic vectors, their performances experience more significant drops compared to when they utilize strong class semantic vectors. For instance, the harmonic mean of f-VAEGAN decreases from 52.9% to 35.3%. These findings highlight the superiority of our induction-based generative method over imagination-based approaches in ZSL, as it can synthesize high-quality sample features for unseen classes



(a) Varying λ effect (b) Varying top-k effect (c) Varying N_{syn} effect

Figure 5. Hyper-parameter analysis. We show the performance variations on CUB by adjusting the value of loss weight λ in (a), the number of the top referent classes top-k in (b), and the number of synthesized samples of each unseen class N_{syn} in (c).



(a) AWA2 (b) CUB

Figure 6. Induction-based *vs* Imagination-based methods on AWA2 and CUB .

with feasible scene generalization. Moreover, our work bridges the gap between large-scale visual-language ZSL methods and classical ZSL methods, leveraging the advantages of both approaches to achieve improved performance in ZSL tasks. More discussions are in Appendix C.

4.3. Qualitative Evaluation

We conducted a qualitative evaluation to intuitively showcase the performance of imagination-based generative ZSL methods (e.g., f-VAEGAN (Xian et al., 2019b)) and our induction-based approach (GenZSL). The t-SNE visualization (Maaten & Hinton, 2008) of real and synthesized sample features is presented in Fig. 4. We randomly selected 10 classes from SUN and visualized the sample features generated by f-VAEGAN and GenZSL. Fig. 4(a) illustrates that sample features synthesized by f-VAEGAN and real features exhibit significant differences, indicating that the synthesized visual features may not facilitate reliable classification for ZSL. In contrast, Fig. 4(b) demonstrates that our GenZSL synthesizes informative samples for unseen classes that closely match real sample features. This visualization confirms that GenZSL is a desirable generative ZSL model, and the induction-based generative paradigm holds value for ZSL tasks.

4.4. Induction *vs* Imagination

We analyze the efficiency and efficacy of induction-based generative ZSL (e.g., our GenZSL) and imagination-based generative ZSL (e.g., f-VAEGAN (Xian et al., 2019b), which is a most typical generative ZSL method) on AWA2 and CUB. Results are shown in Fig. 6. We find that our GenZSL eases the optimization by providing faster convergence at the early stage, while f-VAEGAN towards convergence slowly. For example, GenZSL achieves the best GZSL performance with a remarkable $60\times$ and $10\times$ acceleration in training speed than f-VAEGAN on AWA2 and CUB, respectively. Meanwhile, our GenZSL obtains better performance both in the GZSL and CZSL settings. These demonstrate the efficiency and efficacy of our GenZSL and the great potential of the induction-based generative paradigm.

4.5. Hyper-Parameter Analysis.

We analyze the effects of different hyper-parameters of our GenZSL on the CUB dataset. These hyper-parameters include the loss weight λ in Eq. 5, the number of the top referent classes top-k, and the number of synthesized samples for each unseen class N_{syn}. Fig. 5 shows the CZSL and GZSL performances using different hyper-parameters. In (a), the results indicate that GenZSL is robust to varying values of λ and achieves good performance when λ is relatively small (i.e., $\lambda = 0.1$). This is because $\mathcal{L}_{Boost}$ is a semantic-level toward target class-guided information boosting criteria, which is a supplement to the vision-level one (e.g., $\mathcal{L}_{LR}$). In (b), we evaluate the top similar classes as referent classes varying $k = \{1, 2, 4, 8\}$. We find that our GenZSL uses the $top-2$ referent classes to obtain better performance, which brings the mixup technique for data

augmentation. In (c), our GenZSL is shown robust to N_{syn} when it is not set in a large number. The N_{syn} can be set as 1600 to balance between the data amount and the ZSL performance. Overall, Fig. 5 shows that our GenZSL is robust to overcome hyper-parameter variations. The hyper-parameter analysis on SUN and AWA2 are presented in Appendix E. Accordingly, we empirically set these hyper-parameters $\{\lambda, k, N_{syn}\}$ as $\{0.1, 2, 1600\}$, $\{0.001, 2, 800\}$ and $\{0.1, 2, 5000\}$ for CUB, SUN and AWA2, respectively.

5. Conclusion

We propose an inductive variational autoencoder as a new generative model for zero-shot learning, namely GenZSL. Inspired by human perception, GenZSL operates on an induction-based approach to synthesize informative and high-quality sample features for unseen classes. To achieve this, we introduce class diversity promotion to enhance the diversity and discrimination of class semantic vectors. Additionally, we design two losses targeting the criteria of target class-guided information boosting to optimize the model. Through qualitative and quantitative analyses, we demonstrate that GenZSL consistently outperforms existing generative ZSL methods in terms of efficacy and efficiency.

Impact Statement

Our induction-based generative method 1) offers new insights into ZSL and other generation tasks, 2) aligns with vision-language models (e.g., CLIP) to enable attribute-free generalization, which paves the way for further advancements in ZSL, 3) bridges the gap between classical ZSL method (e.g., generative ZSL) and VLM-based methods (e.g., CLIP).

References

and, S. C. The origin of concepts. *Journal of Cognition and Development*, 1:37–41, 2000. 2, 3

Arora, G., Verma, V., Mishra, A., and Rai, P. Generalized zero-shot learning via synthesized examples. In *CVPR*, pp. 4281–4289, 2018. 1

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., and Fleet, D. J. Synthetic data from diffusion models improves imagenet classification. *arXiv print arXiv:2304.08466*, 2023. 3

Badirli, S., Akata, Z., Mohler, G. O., Picard, C., and Dundar, M. Fine-grained zero-shot learning with dna as side information. In *NeurIPS*, 2021. 2

Carey, S. Conceptual change in childhood. *MIT Press*, 1985. 2, 3

Cavazza, J., Murino, V., and Bue, A. D. No adversaries to zero-shot learning: Distilling an ensemble of gaussian feature generators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:12167–12178, 2023. 6

Chen, H., Liu, Y., Ma, Y., Zheng, N., and Yu, X. TPR: topology-preserving reservoirs for generalized zero-shot learning. In *NeurIPS*, 2024a. 2

Chen, S., Wang, W., Xia, B., Peng, Q., You, X., Zheng, F., and Shao, L. Free: Feature refinement for generalized zero-shot learning. In *ICCV*, 2021a. 1, 2

Chen, S., Xie, G.-S., Yang Liu, Y., Peng, Q., Sun, B., Li, H., You, X., and Shao, L. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. In *NeurIPS*, 2021b. 1, 6

Chen, S., Hong, Z., Liu, Y., Xie, G.-S., Sun, B., Li, H., Peng, Q., Lu, K., and You, X. Transzero: Attribute-guided transformer for zero-shot learning. In *AAAI*, 2022a. 2, 6

Chen, S., Hong, Z., Xie, G.-S., Yang, W., Peng, Q., Wang, K., Zhao, J., and You, X. Msdn: Mutually semantic distillation network for zero-shot learning. In *CVPR*, 2022b. 2, 6

Chen, S., Hou, W. Q., Hong, Z., Ding, X., Song, Y., You, X., Liu, T., and Zhang, K. Evolving semantic prototype improves generative zero-shot learning. In *ICML*, 2023a. 1, 6

Chen, S., Fu, D., Chen, S., Ye, S., Hou, W., and You, X. Causal visual-semantic correlation for zero-shot learning. In *ACM MM*, pp. 4246–4255, 2024b. 6

Chen, S., Hou, W. Q., Khan, S. H., and Khan, F. S. Progressive semantic-guided vision transformer for zero-shot learning. In *CVPR*, 2024c. 2, 6

Chen, S., Hong, Z., You, X., and Shao, L. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2025. 1, 6

Chen, X., Deng, X., Lan, Y., Long, Y., Weng, J., Liu, Z., and Tian, Q. Explanatory object part aggregation for zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(2):851–868, 2024d. 6

Chen, Z., Wang, S., Li, J., and Huang, Z. Rethinking generative zero-shot learning: An ensemble learning perspective for recognising visual patches. In *ACM MM*, 2020. 2

Chen, Z., Huang, Y., Chen, J., Geng, Y., Zhang, W., Fang, Y., Pan, J. Z., and Chen, H. DUET: cross-modal semantic grounding for contrastive zero-shot learning. In *AAAI*, pp. 405–413, 2023b. 6

Chen, Z., Zhang, P., Li, J., Wang, S., and Huang, Z. Zero-shot learning by harnessing adversarial samples. In *ACM MM*, pp. 4138–4146, 2023c. 6

Chou, Y.-Y., Lin, H.-T., and Liu, T.-L. Adaptive and generative zero-shot learning. In *ICLR*, 2021. 6

Christensen, A., Mancini, M., Koepke, A. S., Winther, O., and Akata, Z. Image-free classifier injection for zero-shot classification. In *ICCV*, pp. 19026–19035, 2023. 6

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. In *ICLR*, 2017. 1

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *NeurIPS*, 2014. 1

Han, Z., Fu, Z., Chen, S., and Yang, J. Semantic contrastive embedding for generalized zero-shot learning. *International Journal of Computer Vision*, 130(11):2606–2622, 2022. 2, 6

He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P. H. S., Bai, S., and Qi, X. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023. 3

Hou, W., Chen, S., Chen, S., Hong, Z., Wang, Y., Feng, X., Khan, S. H., Khan, F. S., and You, X. Visual-augmented dynamic semantic prototype for generative zero-shot learning. In *CVPR*, pp. 23627–23637, 2024. 6

Hu, X., Zhang, K., Xia, L., Chen, A., Luo, J., Sun, Y., Wang, K. M., Qiao, N., Zeng, X., Sun, M., Kuo, C.-H., and Nevatia, R. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *WACV*, pp. 2982–2991, 2023. 2

Huynh, D. and Elhamifar, E. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, pp. 4482–4492, 2020a. 2, 6

Huynh, D. T. and Elhamifar, E. Compositional zero-shot learning via fine-grained dense feature composition. In *NeurIPS*, 2020b. 2

Jahanian, A., Puig, X., Tian, Y., and Isola, P. Generative models as a data source for multiview representation learning. In *ICLR*, 2022. 3

Khattak, M. U., Wasim, S. T., Naseer, M., Khan, S., Yang, M., and Khan, F. S. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, pp. 15144–15154, 2023. 2

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2014. 1

Kong, X., Gao, Z., Li, X., Hong, M., Liu, J., Wang, C., Xie, Y., and Qu, Y. En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning. In *CVPR*, pp. 9296–9305, 2022. 6

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350:1332–1338, 2015. 2

Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pp. 951–958, 2009. 1

Lampert, C. H., Nickisch, H., and Harmeling, S. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:453–465, 2014. 2

Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., and Pathak, D. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, pp. 2206–2217, 2023. 3

Li, J., Jing, M., Lu, K., Ding, Z., Zhu, L., and Huang, Z. Leveraging the invariant side of generative zero-shot learning. In *CVPR*, pp. 7394–7403, 2019. 6

Liu, Y., Zhou, L., Bai, X., Huang, Y., Gu, L., Zhou, J., and Harada, T. Goal-oriented gaze estimation for zero-shot learning. In *CVPR*, pp. 3794–3803, 2021. 6

Maaten, L. V. D. and Hinton, G. E. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 8

Naeem, M. F., Khan, M. G. Z. A., Xian, Y., Afzal, M. Z., Stricker, D., Gool, L. V., and Tombari, F. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *CVPR*, pp. 15169–15179, 2023. 6

Naeem, M. F., Xian, Y., Gool, L. V., and Tombari, F. I2dformer+: Learning image to document summary attention for zero-shot image classification. *International Journal of Computer Vision*, 132(9):3806–3822, 2024. 6

Narayan, S., Gupta, A., Khan, F., Snoek, C. G. M., and Shao, L. Latent embedding feedback and discriminative features for zero-shot classification. In *ECCV*, 2020. 1, 6, 7

Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. Zero-shot learning with semantic output codes. In *NeurIPS*, pp. 1410–1418, 2009. 1

Patterson, G. and Hays, J. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pp. 2751–2758, 2012. 2, 5

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 4, 5, 7

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv: 2204.06125*, 2022. 3

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685, 2022. 3

Schönfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., and Akata, Z. Generalized zero- and few-shot learning via aligned variational autoencoders. In *CVPR*, pp. 8239–8247, 2019. 6

Shen, Y., Qin, J., and Huang, L. Invertible zero-shot recognition flows. In *ECCV*, 2020. 6

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. In *NeuIPS*, 2015. 5

Tanwisuth, K., Zhang, S., Zheng, H., He, P., and Zhou, M. Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *ICML*, 2023. 2

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331:1279–1285, 2011. 2

Vyas, M. R., Venkateswara, H., and Panchanathan, S. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *ECCV*, 2020. 6

Wan, Z., Chen, D., Li, Y., Yan, X., Zhang, J., Yu, Y., and Liao, J. Transductive zero-shot learning with visual structure constraint. In *NeurIPS*, 2019. 2

Wang, Z., Liang, J., He, R., Xu, N., Wang, Z., and Tan, T.-P. Improving zero-shot generalization for clip with synthesized prompts. In *CVPR*, 2023. 7

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S. J., and Perona, P. Caltech-ucsd birds 200. *Technical Report CNS-TR-2010-001, Caltech,*, 2010. 2, 5

Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., and Schiele, B. Latent embeddings for zero-shot classification. In *CVPR*, pp. 69–77, 2016. 2

Xian, Y., Lorenz, T., Schiele, B., and Akata, Z. Feature generating networks for zero-shot learning. In *CVPR*, pp. 5542–5551, 2018. 1, 5, 6

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2251–2265, 2019a. 2, 3, 5

Xian, Y., Sharma, S., Schiele, B., and Akata, Z. F-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, pp. 10267–10276, 2019b. 1, 2, 6, 7, 8

Xu, W., Xian, Y., Wang, J., Schiele, B., and Akata, Z. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020. 2, 6

Yue, Z., Wang, T., Zhang, H., Sun, Q., and Hua, X. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, 2021. 1

Zhang, H., Cissé, M., Dauphin, Y., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 5

Zhang, Y., Zhou, D., Hooi, B., Wang, K., and Feng, J. Expanding small-scale datasets with guided imagination. In *NeurIPS*, 2023. 3

Zhou, D., Wang, K., Gu, J., Peng, X., Lian, D., Zhang, Y., You, Y., and Feng, J. Dataset quantization. In *ICCV*, pp. 17159–17170, 2023. 3

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337 – 2348, 2021. 7

Zhu, Y., Xie, J., Tang, Z., Peng, X., and Elgammal, A. Semantic-guided multi-attention localization for zero-shot learning. In *NeurIPS*, 2019. 2

Çetin, S., Baran, O. B., and Cinbis, R. G. Closed-form sample probing for learning generative models in zero-shot learning. In *ICLR*, 2022. 1, 2, 6

Appendix organization:

- Appendix A: Testing process of GenZSL.
- Appendix B: Class semantic vectors' similarity heatmaps.
- Appendix C: Generative ZSL with weak class semantic vectors.
- Appendix D: t-SNE visualization on CUB and AWA2.
- Appendix E: Hyper-parameter analysis on SUN and AWA2.

A. Testing Process of GenZSL

We present the testing process of GenZSL in Fig. 7. Different to the standard VAE that samples the new data from Gaussian noise, our GenZSL inducts the informative new sample features for unseen classes from the similar seen classes and takes Gaussian noises to enable IVAE to synthesize variable and diverse samples. Then, we take the synthesized unseen class samples $\hat{x}^u$ to learn a supervised classifier (e.g., softmax), which is used for ZSL evaluation further.

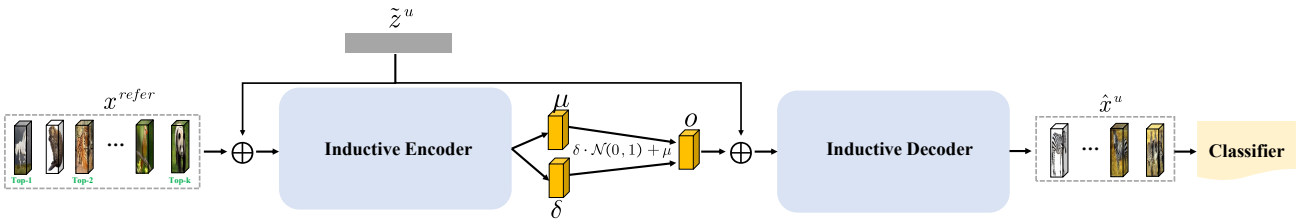
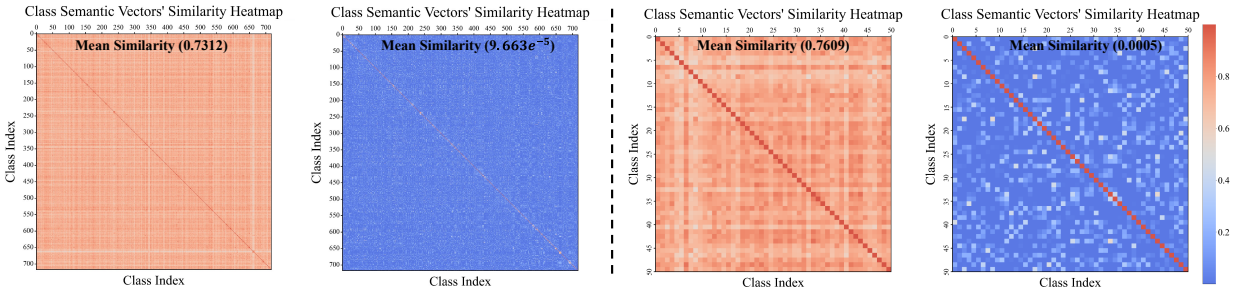


Figure 7. Testing process of GenZSL.

B. Class Semantic Vectors' Similarity Heatmaps

We show the lass semantic vectors' similarity heatmaps of SUN and AWA2 in Fig. 8. Results show that our CDP effectively improves the discrimination and diversity for class semantic vectors, avoiding the confusion of synthesized visual features between various classes. For example, the mean similarity of class semantic vectors on AWA2 is reduced from 0.7609 to 0.0005. As such, the class semantic vectors served as a distinct conditions for effective generation. Furthermore, we note that due to the number of classes in SUN is more than AWA2, the impact of self-similarity of classes in AWA2 is more heavier. This is why the mean similarity in AWA2 (coarse-grained dataset) is larger than SUN (fine-grained dataset).



(a) CLIP text encoder (b) CLIP w/ CDP (c) CLIP text encoder (d) CLIP w/ CDP

Figure 8. Class semantic vectors' similarity heatmaps are extracted by CLIP text encoder and CLIP with class diversity promotion on SUN (a,b) and AWA2 (c,d).

C. Generative ZSL Methods with Weak Class Semantic Vectors

We provide the results of imagination-based ZSL (e.g., f-VAEGAN) and induction-based generative ZSL (e.g., GenZSL) using weak class semantic vectors (e.g., CLIP text embeddings of class names) on SUN and AWA2. Results are shown in Table 5. We find that i) the performances of f-VAEGAN drop heavily on SUN ($\boldsymbol{acc}: 64.7\% \rightarrow 45.2\%$; $\boldsymbol{H}: 41.3\% \rightarrow 33.3\%$) and AWA2 ($\boldsymbol{acc}: 71.1\% \rightarrow 67.1\%$; $\boldsymbol{H}: 63.5\% \rightarrow 59.8\%$) when it uses the weak class semantic vector rather than the strong one (e.g., expert-annotated attributes); ii) our GenZSL achieves significant performance gains over f-VAEGAN. These demonstrate that induction-based generative model is more feasible for ZSL than the imagination-based ones.

Table 5. Results of various generative ZSL methods with weak class semantic vectors on SUN and AWA2.

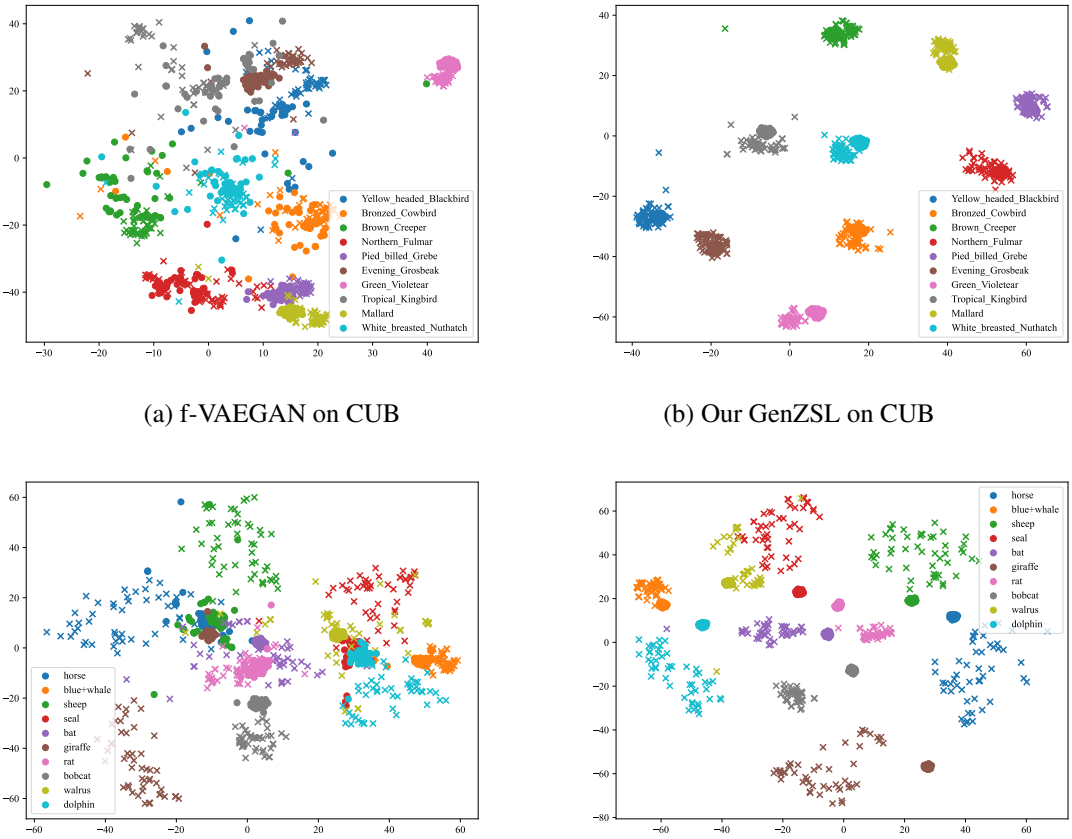
Methods	**SUN**				**AWA2**			
	CZSL	GZSL			CZSL	GZSL		
	acc	U	S	H	acc	U	S	H
f-VAEGAN (strong)	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
f-VEAGAN (weak)	45.2	32.4	34.3	33.3	67.0	43.3	83.2	59.8
GenZSL (weak)	73.5	50.6	43.8	47.0	92.2	86.1	88.7	87.4

D. t-SNE Visualization on CUB and AWA2

As shown in Fig. 9, t-SNE visualizations of visual features learned by the f-VAEGAN and our GenZSL on CUB (a,b) and AWA2 (c,d). Analogously, the visual features generated by f-VAEGAN are also far away from their corresponding real ones, and the discrimination of these real/synthesized visual features is undesirable. In contrast, our GenZSL synthesize visual features close to their corresponding real ones. As such, our GenZSL significantly improves the performances of f-VAEGAN on CUB and AWA2. This demonstrates that GenZSL is a effective generative ZSL model.

E. Hyper-Parameter Analysis on SUN and AWA2

We analyze the effects of different hyper-parameters of our GenZSL on SUN and AWA2 datasets. These hyper-parameters include the loss weight λ in Eq. 5 the number of the top referent classes top-k, and the number of synthesized samples for each unseen class N_{syn}. Fig. 10 shows the GZSL performances of using different hyper-parameters. We observe that our GenZSL is robust and easy to train. We empirically set these hyper-parameters $\{\lambda, k, N_{syn}\}$ as $\{0.001, 2, 800\}$ and $\{0.1, 2, 5000\}$ for SUN and AWA2, respectively.



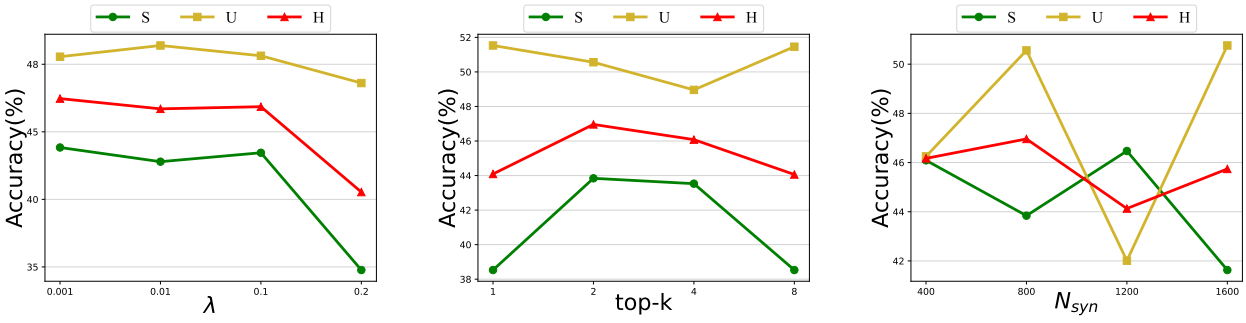
(a) f-VAEGAN on CUB

(b) Our GenZSL on CUB

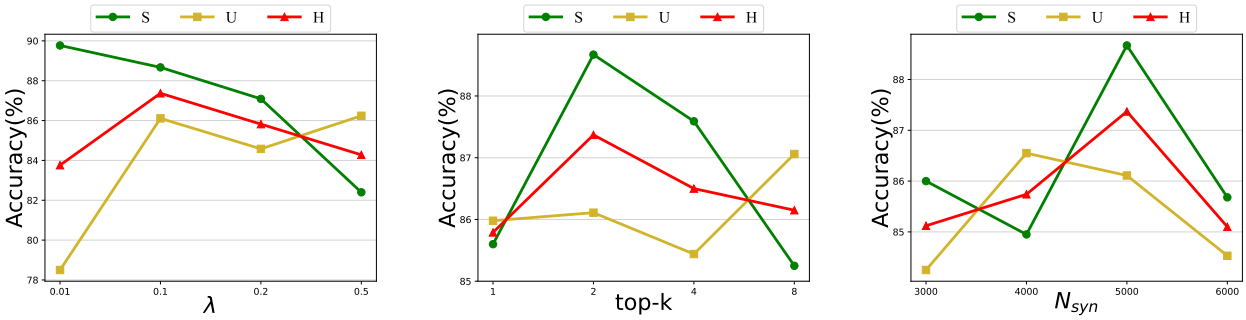
(c) f-VAEGAN on AWA2

(d) Our GenZSL on AWA2

Figure 9. Qualitative evaluation with t-SNE visualization. The sample features from f-VAEGAN are shown on the left, and from our GenZSL are shown on the right. We use 10 colors to denote randomly selected 10 classes from CUB (a,b) and AWA2 (c,d). The "×" and "○" are denoted as the real and synthesized sample features, respectively. The synthesized sample features and the real features distribute differently on the left while distributing similarly on the right.



(a) Varying λ effect (b) Varying top-k effect (c) Varying N_{syn} effect



(d) Varying λ effect (e) Varying top-k effect (f) Varying N_{syn} effect

Figure 10. Hyper-parameter analysis. We show the performance variations loss weight λ, the number of the top referent classes top-k, and the number of synthesized samples of each unseen class N_{syn} on SUN (a,b,c) and AWA2 (d,e,f).