VS-Boost: Boosting Visual-Semantic Association for Generalized Zero-Shot Learning

Xiaofan Li[1], **Yachao Zhang**[2*], **Shiran Bian**[1], **Yanyun Qu**[1*], **Yuan Xie**[3], **Zhongchao Shi**[4] and **Jianping Fan**[4]
[1]School of Informatics, Xiamen University, Fujian, China
[2] Tsinghua University, Shenzhen, China
[3]School of Computer Science and Technology, East China Normal University, Shanghai, China
[4]Lenovo Research, Beijing, China
lxfun@stu.xmu.edu.cn,yachaozhang@sz.tsinghua.edu.cn, yyqu@xmu.edu.cn

Abstract

Unlike conventional zero-shot learning (CZSL) which only focuses on the recognition of unseen classes by using the classifier trained on seen classes and semantic embeddings, generalized zero-shot learning (GZSL) aims at recognizing both the seen and unseen classes, so it is more challenging due to the extreme training imbalance. Recently, some feature generation methods introduce metric learning to enhance the discriminability of visual features. Although these methods achieve good results, they focus only on metric learning in the visual feature space to enhance features and ignore the association between the feature space and the semantic space. Since the GZSL method uses semantics as prior knowledge to migrate visual knowledge to unseen classes, the consistency between visual space and semantic space is critical. To this end, we propose relational metric learning which can relate the metrics in the two spaces and make the distribution of the two spaces more consistent. Based on the generation method and relational metric learning, we proposed a novel GZSL method, termed VS-Boost, which can effectively boost the association between vision and semantics. The experimental results demonstrate that our method is effective and achieves significant gains on five benchmark datasets compared with the state-of-the-art methods.

1 Introduction

Recently, zero-shot learning has made great progress and attracted increasing attention. Conventional zero-shot learning (CZSL) [Lampert *et al.*, 2013] aims to only recognize objects of unseen classes through a classifier learned from seen classes and semantic embeddings *e.g.* attributes and word embeddings. Unlike CZSL with the strong assumption that the query objects are only from unseen classes, generalized zero-shot learning (GZSL) [Xian *et al.*, 2018a] aims to recognize both seen and unseen classes, which is more challenging.

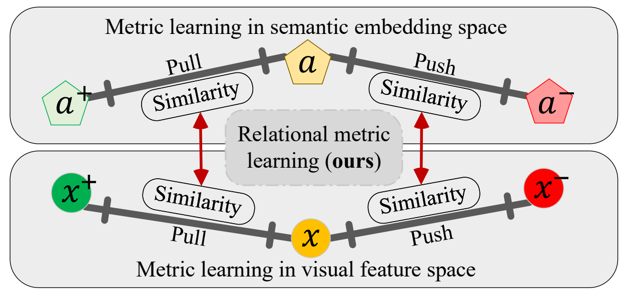


Figure 1: Traditional metric learning is performed in a single space and simply pushes positive instances closer and pulls negative instances farther. The proposed relational metric learning measures the similarity between instances in two spaces and aligns the similarity relationship between the two spaces.

To better transfer knowledge from seen classes to unseen classes, for each category, zero-shot learning introduces a corresponding semantic embedding as prior knowledge *e.g.* manual annotated attributes [Lampert *et al.*, 2013], word embeddings extracted by language models [Reed *et al.*, 2016], etc. The current mainstream solutions for GZSL are semantic embedding methods [Huynh and Elhamifar, 2020] [Xie *et al.*, 2019] and feature generation methods [Xian *et al.*, 2018b] [Li *et al.*, 2019a] [Xian *et al.*, 2019]. The semantic embedding methods project features into the semantic space and perform metric learning in the semantic space to learn a visual-to-semantic inference, and finally perform classification in semantic space using nearest neighbors. Due to the absence of unseen classes, embedding methods are usually biased towards seen classes and performance is inferior to generation methods. The feature generation methods first train a generator to synthesize unseen features conditional on unseen semantic embeddings and Gaussian noises, then the synthetic features and real seen features are used to train a GZSL classifier in a supervised way. Recently, to enhance feature discriminability, some methods [Han *et al.*, 2020] [Han *et al.*, 2021] [Chen *et al.*, 2021a] introduce metric learning into feature generation methods, which use triplet loss [Wen *et al.*, 2016] or contrastive loss [Hadsell *et al.*, 2006] and their vari-

*Corresponding authors.

ants to increase inter-class distance and decrease intra-class distance. However, as illustrated in Figure 1, these feature-refined methods and embedding methods only perform metric learning in feature space alone or in semantic space alone, ignoring the association between the feature space and the semantic space. As is known to all, GZSL uses semantics as prior knowledge to transfer visual knowledge from seen classes to unseen classes and there is a gap between visual and semantic information, thus the association between vision and semantics becomes a crucial problem. To boost visual-semantic association for GZSL, we propose a new feature generation-based method termed VS-Boost and introduce a novel relational metric learning which can bridge the metric learning between two different spaces.

VS-Boost first uses a semantic embedding network to constrain the visual features, where the features extracted by ResNet101 [He *et al.*, 2016] (parameter freezing) are first fine-tuned by the encoder and then projected into the semantic space, and metric learning is performed in the semantic space. The fine-tuned features will be more relevant to the semantics and more discriminative through the constraint of the semantic embedding network. After obtaining semantic-relevant features, the proposed relational metric learning is used to further enhance the consistency between the visual and semantic spaces. Concretely, relational metric learning measures the similarity between instances in the feature space and the semantic space and aligns the similarity of the same categories between the two spaces. We use binary cross-entropy loss to align the similarity between the two spaces and give proof of the validity of the loss function. It is well known that the APY [Farhadi *et al.*, 2009] dataset is the least generalized dataset due to the huge difference between seen and unseen classes, and VS-Boost greatly improves the SOTA level on APY, which indicates that boosting the association between vision and semantics is an effective way to solve the GZSL problem.

In this paper, our contributions are as follows:

- We propose a novel relational metric learning, which can relate the metric learning of two different spaces and enhance the consistency of the distribution of the two spaces.
- Based on the feature generation method and relational metric learning, we propose a novel framework for GZSL, termed VS-Boost, which effectively enhances the association between visual space and semantic space, thus greatly improving the generalization of the model to unseen classes.
- We evaluated our method on five GZSL benchmark datasets and experimentally find that ours achieves competitive results with significant gains.

2 Related Work

2.1 Conventional Zero-Shot Learning

Early zero-shot learning methods focused on the conventional zero-shot learning (CZSL) problem, where the testing set only contains unseen classes. Semantic embedding models [Frome *et al.*, 2013] [Akata *et al.*, 2015a] [Akata *et al.*, 2015b] [Romera-Paredes and Torr, 2015] [Kodirov *et al.*, 2017] [Xian *et al.*, 2016] learn a mapping from an image feature space to a semantic space. The classic semantic embedding methods DAP and IAP [Lampert *et al.*, 2013] make use of the semantic embeddings within a two-stage approach to infer the label of an image that belongs to one of the unseen classes. In addition, other hybrid models [Zhang and Saligrama, 2015] [Norouzi *et al.*, 2013] [Changpinyo *et al.*, 2016] embed both images and semantic embeddings into another intermediate space to perform classification. These embedding methods have achieved good results on CZSL task.

2.2 Generalized Zero-Shot Learning

The concept of generalized zero-shot learning (GZSL) [Xian *et al.*, 2018a] has received significant attention since its proposal. In GZSL, the testing set contains both seen and unseen classes, due to the overfitting of seen classes, the existing CZSL methods decline dramatically in performance and suffer from a very serious strong-bias problem. In order to solve the problem of shortage of unseen-class samples, the generative adversarial networks (GAN) [Goodfellow *et al.*, 2014] [Mirza and Osindero, 2014] [Arjovsky *et al.*, 2017] and variational auto-encoding (VAE) [Kingma and Welling, 2013] were introduced for GZSL, where a generator was trained to synthesize unseen-class visual features conditional on corresponding semantic embeddings. Most of the current feature generation methods [Xian *et al.*, 2018b] [Felix *et al.*, 2018] [Li *et al.*, 2019a] [Sariyildiz and Cinbis, 2019] [Xian *et al.*, 2019] [Narayan *et al.*, 2020] attempt to learn an inference from semantic embeddings to visual features and some methods [Verma *et al.*, 2020] [Liu *et al.*, 2021] introduce the meta-learning strategy into the feature generation method to improve the generalization of the model. The common space methods [Ma and Hu, 2020] [Schonfeld *et al.*, 2019] [Chen *et al.*, 2021b] propose to learn a common space into which both visual features and semantic embeddings are projected for effective knowledge transfer. In addition to the feature generation methods, the prototype generation methods [Li *et al.*, 2019b] [Yu *et al.*, 2020] [Liu *et al.*, 2020] also achieved good results in GZSL, where the semantics-to-prototype mapping is trained and the synthetic prototypes are used as a classifier for different classes. The attention-based methods [Xie *et al.*, 2019] [Huynh and Elhamifar, 2020] [Min *et al.*, 2020] [Chen *et al.*, 2022] usually use attention mechanisms to extract visual features which fit better with the semantics and design the new loss functions to balance the predictions between seen and unseen classes in GZSL task. There are also some open-set classification methods [Yue *et al.*, 2021] [Chou *et al.*, 2020] applied in zero-shot learning, which first separate the unseen classes from the seen classes, and then classify them separately.

Recently, in order to enhance the discriminability of features, some methods [Han *et al.*, 2020] [Han *et al.*, 2021] [Chen *et al.*, 2021a] introduce metric learning *e.g.* triplet loss [Wen *et al.*, 2016] [Schroff *et al.*, 2015] and contrastive loss [Hadsell *et al.*, 2006] to the generation method, but these methods only perform metric learning in the feature space without considering linking features to the semantic space, which is not conducive to model generalization.

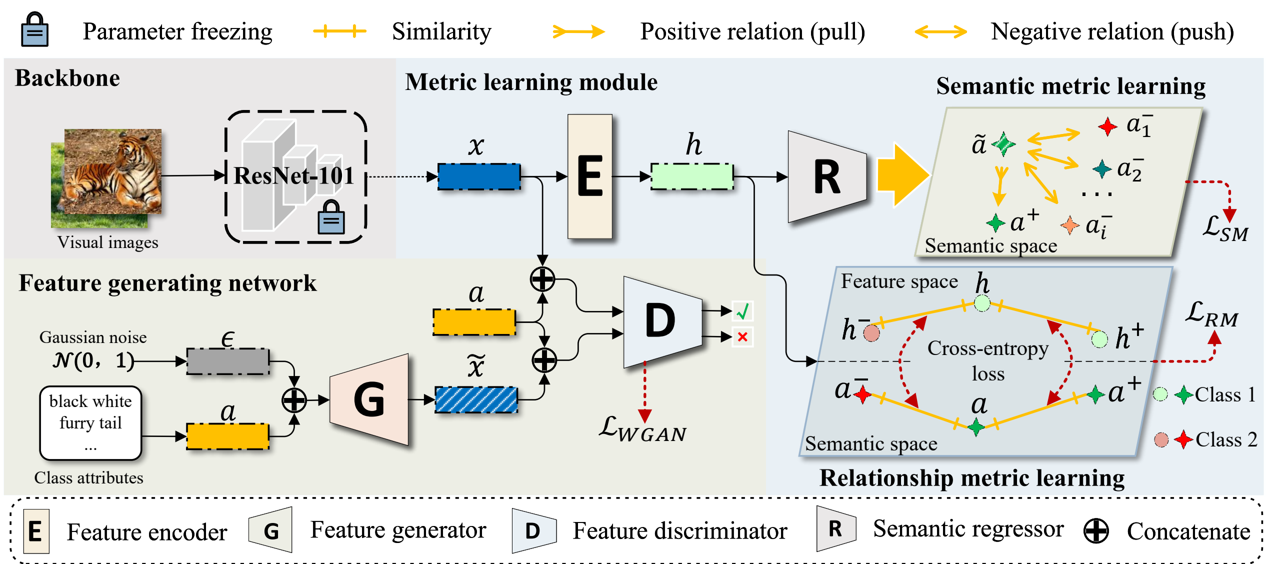


Figure 2: The framework of our proposed VS-Boost. Visual feature x is extracted from ResNet-101 and the black dashed line indicates there is no gradient back propagation. The feature generating network and metric learning module are trained on the fly. metric learning module contains the semantic metric learning module and the relational metric learning module.

3 Method

3.1 Problem Definition

The zero-shot learning problem is defined as follows: a training (seen classes) dataset $\mathcal{D}^{tr} = \{(x_i, y_i)|x_i \in \mathcal{X}, y_i \in \mathcal{S}\}$, where x_i is the visual feature and y_i is its corresponding label, $\mathcal{S}$ is the label set of seen classes. In CZSL task, the testing set is denoted as $\mathcal{D}^{te} = \{(x_j, y_j)|x_j \in \mathcal{X}, y_j \in \mathcal{U}\}$, where $\mathcal{U}$ is the label set of unseen classes and $\mathcal{S} \cap \mathcal{U} = \varnothing$. While in GZSL task, the testing set is denoted as $\mathcal{D}^{te} = \{(x_j, y_j)|x_j \in \mathcal{X}, y_j \in \mathcal{S} \cup \mathcal{U}\}$. In zero-shot learning, each seen and unseen class has its own corresponding semantic embedding $a_k \in \mathcal{A}, \forall k \in \mathcal{S} \cup \mathcal{U}$. Given $\mathcal{D}^{tr}$ and $\mathcal{A}$, the task of CZSL is to learn the classifier $f_{czsl} : \mathcal{X} \to \mathcal{U}$, and the task of GZSL is to learn the classifier $f_{gzsl} : \mathcal{X} \to \mathcal{S} \cup \mathcal{U}$. Due to the strong-bias to seen classes, GZSL task is more challenging than CZSL.

3.2 Method Overview

The architecture of VS-Boost is illustrated in Fig. 2, and it contains two streamlines: the feature generating network and the metric learning module. In the feature generating network, we train a generator G to synthesize the visual features from the semantic embeddings. And in the metric learning module, a feature encoder E is trained for refining original features. Different from the existing metric learning used in GZSL which only focuses on metric learning in the semantic space or feature space, our VS-Boost introduces a novel relational metric learning to relate the measures of semantic spaces and feature spaces. The metric learning module contains the classical semantic embedding network and the proposed relational metric learning. The visual features extracted by ResNet101 [He *et al.*, 2016] are refined by encoder E as $h = E(x)$, and the h is mapped to the semantic space, and semantic metric learning is completed by InfoNCE loss [Van den Oord *et al.*, 2018]. Moreover, we enforce the relational metric learning to constrain encoder E. As illustrated in Figure 3, relational metric learning first calculates the similarity between refined features by a learnable function F and then measures the similarity of semantic embeddings. Cross-entropy loss is employed to bridge the similarity between features and their corresponding semantic embeddings. Through relational metric learning, the distribution of feature space and semantic space becomes more consistent, which is greatly conducive to the inference of visual tasks with semantic embedding as a cue.

In classification, the trained feature generator G will be used to synthesize features of unseen classes, then the synthetic unseen features and real seen class features are refined by the encoder E as the input to a classifier.

3.3 Feature Generating Network

The feature generating network [Xian *et al.*, 2018b] introduces GAN into GZSL for the first time and achieves outstanding results than previous methods. GAN learns a feature generator G to synthesize the visual features $\widetilde{x} = G(a, \epsilon)$ conditioned on a class-level semantic embedding a and Gaussian noise $\epsilon \in \mathcal{N}(0, 1)$. At the same time, the discriminator of generator D is cross-iteratively trained with the generator to discriminate between a real pair (x, a) and a synthetic pair $(\widetilde{x}, a)$. The generator tries to generate a more realistic synthetic feature $\widetilde{x}$ with its corresponding semantic embedding a. The generative model adopts the Wasserstein Generative Adversarial Networks (WGAN) [Arjovsky *et al.*, 2017] and introduces the gradient penalty term [Gulrajani *et al.*, 2017] to train G_F and D, the adversarial training loss of WGAN can be formulated as:

$$\mathcal{L}_{WGAN} = \mathbb{E}\left[D(x, a)\right] - \mathbb{E}\left[D(\widetilde{x}, a)\right] - \gamma \mathbb{E}\left[(||\nabla_{\hat{x}} D(\hat{x}, a)||_2 - 1)^2\right], \quad (1)$$

where $\mathcal{E}$ indicates expectation, $\widetilde{x} = G(a, \epsilon), \hat{x} = \alpha x + (1 - \alpha)\widetilde{x}$ with $\alpha \sim U(0, 1)$ and γ is the penalty coefficient. As suggested in [Gulrajani *et al.*, 2017], we fix $\gamma = 10$.

3.4 Semantic Metric Learning

The semantic embedding network [Frome *et al.*, 2013] [Akata *et al.*, 2015a] [Akata *et al.*, 2015b] [Romera-Paredes and Torr, 2015] [Kodirov *et al.*, 2017] [Xian *et al.*, 2016] was originally used in CZSL to learn a mapping function R that maps a visual feature x into the semantic space denoted as $R(x)$. The commonly-used semantic embedding methods rely on a structured loss function [Akata *et al.*, 2015b][Frome *et al.*, 2013] formulated as below:

$$\mathcal{L}_{STR} = \mathbb{E}_x \left[\max \left(0, \Delta - (a^+)^\top R(x) + \left(a^-\right)^\top R(x) \right) \right], \tag{2}$$

where a^+ is the semantic embedding corresponding to class of x, $a^- \neq a$ is a randomly-selected semantic embedding of other classes, and $\delta > 0$ is a margin. The structured loss is of the same form as triplet loss [Wen *et al.*, 2016] [Schroff *et al.*, 2015; Zhang *et al.*, 2022], allowing the model to perform metric learning in semantic space. Recently, semantic embedding networks is used by many generation methods [Felix *et al.*, 2018] [Narayan *et al.*, 2020] [Chen *et al.*, 2021a] as a reconstructor to give synthetic features a consistency constraint guaranteed that the features synthesized from semantic embeddings can be reconstructed back to semantic embeddings. In this paper, to boost the associations between features and semantics, we introduce the semantic embedding network to impose a semantic measure constraint on original features through training an encoder E to refine features. Unlike some methods[Han *et al.*, 2020] [Han *et al.*, 2021] [Chen *et al.*, 2021a; Hu *et al.*, 2021] that perform metric learning directly in visual space, mapping the visual features to the semantic space and performing metric learning makes the visual features more relevant to their semantics. Furthermore, since semantics have excellent discriminability, using semantic metric learning also makes the model learn to represent more discriminative visual features. Concretely, as illustrated in Fig. 2, the original features x are refined by encoder E, and the refined features h are mapped to the semantic space to obtain the mapped semantic embeddings $\widetilde{a} = R(h)$. In order to ensure the discriminability of $\widetilde{a}$ in the semantic domain, we employ the current popular infoNCE loss [Van den Oord *et al.*, 2018] instead of structure loss as the objective function for semantic metric learning, which is formulated as:

$$\mathcal{L}_{SM} = -log \frac{exp(\widetilde{a}^\top \cdot a^+ / \tau)}{exp(\widetilde{a}^\top \cdot a^+ / \tau) + \sum_{i=1}^{N-1} exp(\widetilde{a}^\top \cdot a_i^- / \tau)}, \tag{3}$$

where $\tau > 0$ is the temperature parameter for infoNCE loss, N is the total number of semantic embeddings. Although semantic metric learning can make features and semantics more correlated, it is still not enough, because it only performs metric learning in a separate space without really bridging the semantic space and the feature space together.

3.5 Relational Metric Learning

Recently, some methods [Han *et al.*, 2020] [Han *et al.*, 2021] [Chen *et al.*, 2021a] have proposed to refine visual features by triplet loss [Wen *et al.*, 2016] or its variants, which effectively improve the performance of GZSL through feature augmentation. However, these methods only perform metric learning in the feature space, which can effectively improve the discriminability of features but cannot enhance the association between features and semantics. In this subsection, we introduce a novel relational metric learning for instance-level constraint. Unlike conventional metric learning simply pulls intra-class instances closer and inter-class instances farther, relational metric learning effectively relates the metrics in the semantic space with the metrics in the feature space, thus making the distribution in the feature space more consistent with the distribution in the semantic space. Specifically, relational metric learning is based on the learnable similarity function F (see Fig. 3), which learns to predict similarity probability between two features. F is achieved through a learnable inner product similarity and activated by a sigmoid activation function σ, which is formulated as:

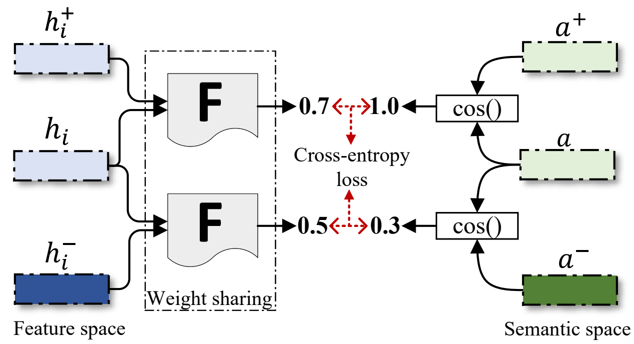


Figure 3: Illustration of our proposed relational metric learning. F indicates the learnable similarity function and $cos()$ indicates cosine similarity.

$$F(h_i, h_j) = \sigma(w^F(h_i \circ h_j)), \tag{4}$$

where w^F is a $[2048, 1]$ fully connected layer, 2048 is the dimension of h and $\circ$ indicates element-wise multiplication. By scoring the similarity of two instances, F can be modeled as a probability prediction problem. We take the cosine similarity between semantic embeddings as the ground truth and the cross-entropy loss is as follows:

$$\begin{aligned}\mathcal{L}_{bce}(h_i, h_j) = -[&cos(a_i, a_j) log F(h_i, h_j) + \\ &(1 - cos(a_i, a_j)) log(1 - F(h_i, h_j))],\end{aligned} \tag{5}$$

where h_i, h_j are the refined features of pair-wise instances and a_i, a_j are their corresponding semantic embeddings. $cos(a_i, a_j) = \frac{a_i \cdot a_j}{\|a_i\| \|a_j\|}$ indicates the cosine similarity.

Theorem. *Since the semantic embedding space does not make any changes, supposed that $\zeta = cos(a_i, a_j) \in [0, 1]$ is a constant after the calculation and $\xi = F(h_i, h_j) \in (0, 1)$ is an independent variable, Equation (5) is expressed as $\mathcal{L}_{bce}(\xi) = -\zeta log \xi - (1 - \zeta) log(1 - \xi)$, and the partial derivation is formulated as follows:*

$$\frac{\partial \mathcal{L}_{bce}(\xi)}{\partial \xi} = -\frac{\zeta}{\xi} + \frac{1 - \zeta}{1 - \xi} = \frac{\xi - \zeta}{\xi(1 - \xi)}, \tag{6}$$

where $\mathcal{L}_{bce}(\xi)$ achieves a minimum value when and only when $\xi = \zeta$ i.e. the distribution relation of feature space is consistent with the distribution relation of semantic space.

Furthermore, imitating triplet loss [Schroff *et al.*, 2015], each instance h is compared with its positive sample h^+ and negative sample h^- through F, where h^+ and h^- are sampled at random. Based on Eq.(5), the relational metric loss is as follows:

$$\mathcal{L}_{RM} = \mathbb{E}_h \left[\mathcal{L}_{bce}(h, h^+) + \mathcal{L}_{bce}(h, h^-) \right]. \quad (7)$$

Through relational metric learning, on the one hand, the discriminability of visual features is improved by the inter-class and intra-class metric learning, on the other hand, it enables a more consistent distribution between semantics and features. The distribution of the visual space is consistent with the feature space in favor of the classification of unseen classes because unseen class classification is guided by semantic cues.

3.6 Optimization

The full model of our VS-Boost optimizes G, D, E, R, and F simultaneously with the following objective function:

$$\min_{G,R,E,F} \max_{D} \mathcal{L}_{WGAN} + \mathcal{L}_{SM} + \mathcal{L}_{RM}, \quad (8)$$

where no hyper-parameters are needed to balance the different losses to achieve the desired results.

3.7 Classification

After the proposed VS-Boost has been well trained, the seen features extracted by ResNet-101 [He *et al.*, 2016] and unseen features synthesized by the generator are refined through encoder E as $h = E(x)$. Suppose that the refined feature sets of seen classes and synthetic unseen classes are $\mathcal{H}_s$ and $\widetilde{\mathcal{H}}_u$, which can be used to train a standard classifier through minimizing the cross-entropy loss:

$$\min_{\theta} \mathbb{E}_h \left[-logP(y|h;\theta) \right], \quad (9)$$

where θ is the parameter of the classifier, and $P(y|h)$ is the softmax prediction. We denote in CZSL, $h \in \widetilde{\mathcal{H}}_u$,$y \in \mathcal{U}$ while in GZSL $h \in \widetilde{\mathcal{H}}_u \cup \mathcal{H}_s$, $y \in \mathcal{S} \cup \mathcal{U}$.

In the testing, the testing features x_t are also refined as $h_t = E(x_t)$. The classification function is:

$$f(x) = arg \max_{y} P(y|h_t;\theta), \quad (10)$$

where in CZSL, $y \in \mathcal{U}$ and in GZSL,$y \in \mathcal{S} \cup \mathcal{U}$.

4 Experiments

Dataset. We evaluate our method on the five benchmark datasets for zero-shot learning: Attribute Pascal and Yahoo (**APY** [Farhadi *et al.*, 2009]), Animals with Attributes (**AWA** [Xian *et al.*, 2018a]), Caltech-UCSD Birds-200-2011(**CUB**) [Welinder *et al.*, 2010], Oxford Flowers (**FLO**) [Nilsback and Zisserman, 2008] and SUN Attribute (**SUN**) [Patterson and Hays, 2012]. Among them, AWA, APY, and SUN use class attributes as semantic embeddings, and CUB and FLO use word embeddings extracted by CNN-RNN [Reed *et al.*, 2016] as semantic embeddings. APY is annotated with 64-dimensional attributes and combines datasets a-Pascal and a-Yahoo, which has 30 and 12 classes respectively. AWA is a coarse-grained animal dataset with manually annotated 85-dimensional attributes. While CUB and FLO are two fine-grained datasets with 1,024-dimensional word embeddings. And SUN is a scenario dataset with annotated 102-dimensional attributes. Table 1 shows the detailed statistics of the five datasets. Similar to the state-of-the-art generation methods, we extract the 2048-dimensional visual features for five datasets with the backbone ResNet-101 [He *et al.*, 2016] pre-trained on ImageNet [Krizhevsky *et al.*, 2012] without finetuning. In addition, we adopt the Proposed Split(PS) [Xian *et al.*, 2018a] to divide all classes on each dataset into seen and unseen classes.

Dataset	$*\mathcal{A}$	$\#\mathcal{D}^{tr}$	$\#\mathcal{D}^{te}_s$ / $\#\mathcal{D}^{te}_u$	$\#\mathcal{S}$ / $\#\mathcal{U}$
APY	64	5,932	7,924 / 1,483	20 / 12
AWA	85	23,527	5,882 / 7,913	40 / 10
CUB	1,024	7,057	1,764 / 2,967	150 / 50
FLO	1,024	5,631	1,403 / 1,155	82 / 20
SUN	102	10,320	2,850 / 1,440	645 / 72

Table 1: The statistics of five benchmark datasets. * denotes dimension size, # denotes the number. $\mathcal{A}$ is the set of semantic embeddings, $\mathcal{D}^{tr}$, $\mathcal{D}^{te}_s$, and $\mathcal{D}^{te}_u$ are training set, testing seen classes set, and testing unseen classes set, respectively. $\mathcal{S}$ and $\mathcal{U}$ are categories of seen classes and unseen classes.

Evaluation Protocols. Following the evaluation strategy in [Xian *et al.*, 2018a], we compute the average per-class Top-1 recognition accuracy (Acc) as the criteria. We evaluate Acc of unseen classes (noted as U) and seen classes (noted as S). And the performance of GZSL is measured by their harmonic mean: $H = 2 \times S \times U/(S + U)$.

Implementation Details. We implement our model by using PyTorch based on Python 3.7 platform. The the proposed model is trained and evaluated on one GeForce RTX 3090 GPU. As a pre-processing step, we normalize the visual features like [Li *et al.*, 2019a]. Feature generator G, discriminator D, and semantic regressor R are multilayer perceptrons that contain a 4,096-unit hidden layer with LeakyReLU activation. The feature encoder E is a $[2048, 2048]$ Linear layer with LeakyReLU activation. Finally, we use the task with N way, K shot (N-K) random sampling for training, and use a random mini-batch size of 8-64 for APY and AWA, 4-16 for CUB, 1-32 for FLO and 64-2 for SUN in our method.

4.1 Compared with State-of-the-arts

To evaluate the performance of our method, we compare VS-Boost with fourteen SOTA methods: GXE[Li *et al.*, 2019b], DVBE[Min *et al.*, 2020], DAZLE[Huynh and Elhamifar, 2020], AREN[Xie *et al.*, 2019], MSDN[Chen *et al.*, 2022], f-CLSWGAN[Xian *et al.*, 2018b], LisGAN[Li *et al.*, 2019a], RFF-GZSL[Han *et al.*, 2020], TF-VAEGAN[Narayan *et al.*, 2020], TGMZ[Liu *et al.*, 2021], FREE[Chen *et al.*, 2021a], SDGZSL[Chen *et al.*, 2021c], CE-GZSL[Han *et al.*, 2021], and ICCE[Kong *et al.*, 2022]. From Table 2, it is ob-

		Method	APY			AWA			CUB			FLO			SUN		
			U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
†	2019	GXE	26.5	74.0	39.0	56.4	81.4	66.7	47.4	47.6	47.5	-	-	-	36.3	42.8	39.3
†	2020	DVBE	32.6	58.3	41.8	63.6	70.8	67.0	53.2	60.2	56.5	-	-	-	45.0	37.2	40.7
†	2020	DAZLE	-	-	-	60.3	75.7	67.1	56.7	59.6	58.1	-	-	-	24.3	52.3	33.2
†	2020	AREN	30.0	47.9	36.9	54.7	79.1	64.7	63.2	69.0	66.0	-	-	-	40.3	32.3	35.9
†	2022	MSDN	-	-	-	62.0	74.5	67.7	68.7	67.5	68.1	-	-	-	52.2	34.2	41.3
‡	2018	f-CLSWGAN	-	-	-	56.1	65.5	60.4	43.7	57.7	49.7	59.0	73.8	65.6	42.6	36.6	39.4
‡	2019	LisGAN	34.3	68.2	45.7	-	-	-	46.5	57.9	51.6	57.7	83.8	68.3	42.9	37.8	40.2
‡	2020	RFF-GZSL	-	-	-	59.8	75.1	66.5	52.6	56.6	54.6	65.2	78.2	71.1	45.7	38.6	41.9
‡	2020	TF-VAEGAN	-	-	-	59.8	75.1	66.6	52.8	64.7	58.1	62.5	84.1	71.7	45.6	40.7	43.0
‡	2021	TGMZ	34.8	77.1	48.0	64.1	77.3	70.1	60.3	56.8	58.5	-	-	-	-	-	-
‡	2021	FREE	-	-	-	60.4	75.4	67.1	55.7	59.9	57.7	67.4	84.5	75.0	47.4	37.2	41.7
‡	2021	SDGZSL	38.0	57.4	45.7	64.6	73.6	68.8	59.9	66.4	63.0	62.2	79.3	69.8	-	-	-
‡	2021	CE-GZSL	-	-	-	63.1	78.6	70.0	63.9	66.8	65.3	69.0	78.7	73.5	48.8	38.6	43.1
‡	2022	ICCE	45.2	46.3	45.7	65.3	82.3	72.8	67.3	65.5	66.4	66.1	86.5	74.9	-	-	-
		VS-Boost	49.8	69.6	58.1	67.9	81.6	74.1	68.0	68.7	68.4	69.1	84.0	75.8	49.2	37.4	42.5

Table 2: Comparisons with the SOTA GZSL methods. U and S are the Top-1 recognition accuracy of unseen and seen classes, respectively. H is the harmonic mean of U and S. ‡ denotes feature generation methods and † denotes other methods. The best and second best results are respectively marked in red and blue.

served that our VS-Boost achieves competitive results. In the harmonic mean H, the main criteria of GZSL, VS-Boost achieves the best results on APY, AWA, CUB, and FLO, and the gains are 10.1%, 1.3%, 0.3%, and 0.8% against TGMZ [Liu *et al.*, 2021], ICCE [Kong *et al.*, 2022], MSDN [Chen *et al.*, 2022] and FREE[Chen *et al.*, 2021a] respectively. VS-Boost makes significant gains on GZSL, especially on APY which is recognized as a challenging dataset due to the huge difference between seen and unseen domains. On SUN, the result of the CE-GZSL[Han *et al.*, 2021] is 0.6% higher than ours and VS-Boost is inferior to CE-GZSL[Han *et al.*, 2021] and TF-VAEGAN[Narayan *et al.*, 2020] in terms of S and U. We speculate that the disadvantage is that SUN has 727 classes but only 102-dimensional semantic embeddings that provide very limited information than visual features, which degrades the performance of our VS-Boost. Furthermore, in terms of unseen classes, VS-Boost achieves the best results on APY, AWA, and FLO and achieves second place on other datasets. The gains for unseen classes are 4.6%, 2.7%, and 0.1% on APY, AWA, and FLO, respectively. The performance improvement achieved on GZSL fully validates the effectiveness of our proposed VS-Boost. And the best results on four datasets (especially on APY) indicate that the generalization of our method is more excellent than other methods.

4.2 Conventional Zero-Shot Learning

In addition to the GZSL task, we implement our method on the conventional zero-shot learning (CZSL) task, where the test set contains only unseen classes. We compare VS-Boost with nine CZSL methods: DAP&IAP [Lampert *et al.*, 2013], SSE [Zhang and Saligrama, 2015], LATEM [Xian *et al.*, 2016], DEVISE [Frome *et al.*, 2013], SJE [Akata *et al.*, 2015b], ALE [Akata *et al.*, 2015a], ESZSL [Romera-Paredes and Torr, 2015], SYNC[Changpinyo *et al.*, 2016], and four resent GZSL methods: LisGAN [Li *et al.*, 2019a], TF-VAEGAN [Narayan *et al.*, 2020], CE-GZSL [Han *et al.*, 2021], ICCE [Kong *et al.*, 2022]. As documented in Table 3, our VS-Boost achieves significant gains on five benchmark datasets, whether compared with the CZSL methods or the GZSL methods. In detail, VS-Boost made 16.7%, 1.5%, 1.4%, and 0.4% improvements on APY, AWA, CUB, and FLO, respectively. Although the results of VS-Boost are not as satisfactory as some GZSL methods on SUN, it is still 4.3% better than the best CZSL method. The competitive results achieved under CZSL verify the superior capabilities of our VS-Boost.

Method	APY	AWA	CUB	FLO	SUN
DAP	33.8	46.1	40.0	-	39.9
IAP	36.6	35.9	24.0	-	19.4
SSE	34.9	61.0	43.9	-	51.5
LATEM	35.2	55.8	49.3	40.4	55.3
DEVISE	39.8	59.7	52.0	45.9	56.5
SJE	32.9	61.9	53.9	53.4	53.7
ALE	39.7	62.5	54.9	48.5	58.1
ESZSL	38.3	58.6	53.9	51.0	54.5
SYNC	23.9	46.6	55.6	-	56.3
LisGAN	43.1	70.6	58.8	69.6	61.7
TF-VAEGAN	-	72.2	64.9	70.8	66.0
CE-GZSL	-	70.4	77.5	70.6	63.3
ICCE	49.5	72.7	78.4	71.6	-
VS-Boost	66.2	74.2	79.8	72.0	62.4

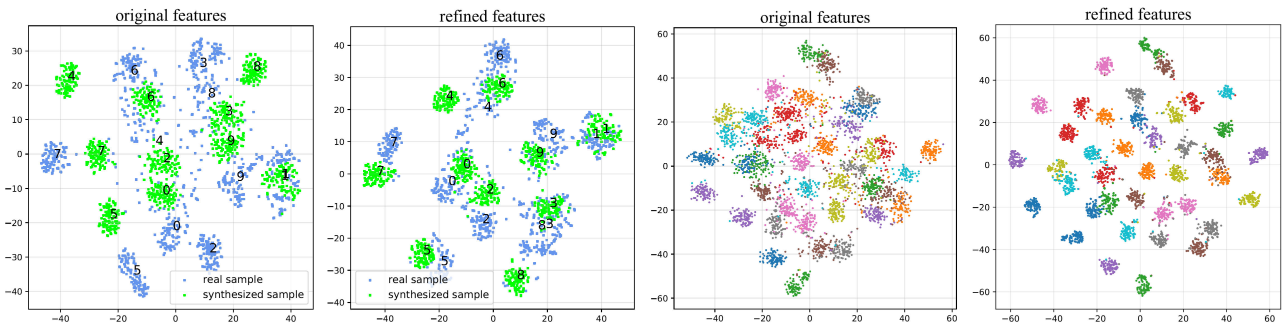
Table 3: Comparison results of CZSL. The first nine methods are early CZSL methods and the following four methods are recently proposed GZSL methods. The best and second best results are respectively marked in red and blue.

4.3 Ablation Study

To provide further insight into VS-Boost, we conduct ablation studies to evaluate the effects of semantic metric learning (SML) and relational metric learning (RML). Based on

			APY			AWA			CUB			FLO			SUN		
a	b	c	U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
✓	✗	✗	35.5	63.1	45.4	57.3	71.1	63.5	57.4	60.2	58.8	59.1	76.0	66.5	44.5	36.3	40.0
✓	✓	✗	39.9	**73.5**	51.7	64.6	80.9	71.8	64.2	67.7	65.9	64.9	**84.2**	73.3	45.6	**38.2**	41.6
✓	✗	✓	45.3	67.2	54.1	65.8	77.7	71.3	64.4	63.2	63.8	66.9	81.5	73.5	46.7	37.0	41.3
✓	✓	✓	**49.8**	69.6	**58.1**	**67.9**	**81.6**	**74.1**	**68.0**	**68.7**	**68.3**	**69.1**	84.0	**75.8**	**49.2**	37.4	**42.5**

Table 4: Ablation study results in the GZSL task on five datasets. a indicates a plain feature generation method. b and c indicate the use of semantic metric learning and relational metric learning, respectively. The best results are marked in **boldface**.



(a) (b) (c) (d)

Figure 4: Visualization of AWA dataset through t-distributed stochastic neighbor embedding (t-SNE), including original features and refined features.

the feature generating network, we introduce SML and RML independently and analyze the results. The results of the ablation study are shown in Table 4, after using SML individually, the results (in terms of H) on the five datasets are improved by 6.3%, 8.3%, 7.1%, 6.8%, and 1.6%, respectively. While using RML individually, the results on the five datasets are improved by 8.7%, 7.8%, 6.0%, 7.0%, and 1.3%, respectively. The great improvements on five datasets fully verify RML and RML both have a significant impact on the VS-Boost. Concretely, after using SML, the recognition accuracy of both seen and unseen classes is greatly improved, while, RML has more significant enhancements for unseen classes. We conjecture that it is because RML associates the semantic space with the feature space, which is very beneficial for the generalization of unseen classes. After using both SML and RML, the results (in terms of H) are greatly improved on the five datasets by 12.6%, 10.7%, 9.6%, 9.3%, and 2.5%, respectively.

4.4 Quantitative Analysis

Figure 4(a) and 4(b) show the visualization results of the unseen features synthesized from the semantic embeddings and the real unseen features, with different numbers representing different category centers. It can be seen that after refinement by VS-Boost, the gap between real features and synthetic features becomes significantly smaller. We speculate that this is due to the better connection between the feature space and the semantic space after fine-tuning by VS-Boost, which allows the generator to do the inference from semantic embeddings to visual features more efficiently. Furthermore, as shown in Figure 4(c) and 4(d), we visualize all seen and unseen features, with different colors representing different classes. After the VS-Boost refinement, it can be observed that the refined features are vastly improved in both inter-class discriminability and intra-class aggregation, which shows that our proposed VS-Boost can effectively enhance the discriminability of visual features while enhancing visual and semantic consistency.

5 Conclusion

In this paper, we propose a novel GZSL method termed VS-Boost, where a stronger association between visual features and semantic embeddings can be built. VS-Boost first uses a semantic embedding network to extract semantic-relevant visual features and then relates the visual feature space with the semantic embedding space by the proposed relational metric learning. The experimental results on five benchmark datasets demonstrate that VS-Boost improves SOTA performance on four datasets, in particular on APY with a 10% improvement in the harmonic mean. The huge performance improvement indicates that boosting the association between vision and semantics is a very effective solution to the GZSL problem.

Acknowledgements

This work is supported by the National Key Research and Development Program of China No.2020AAA0108301, the National Natural Science Foundation of China under Grant No.62176224, 62222602, and 62176092, Shanghai Science and Technology Commission No.21511100700, and CCF-Lenovo Blue Ocean Research Fund.

References

[Akata *et al.*, 2015a] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*, 38(7):1425–1438, 2015.

[Akata *et al.*, 2015b] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015.

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017.

[Changpinyo *et al.*, 2016] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016.

[Chen *et al.*, 2021a] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. In *ICCV*, pages 122–131, 2021.

[Chen *et al.*, 2021b] Shiming Chen, Guosen Xie, Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *NeurIPS*, 34, 2021.

[Chen *et al.*, 2021c] Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, and Zheng Zhang. Semantics disentangling for generalized zero-shot learning. In *ICCV*, pages 8712–8720, 2021.

[Chen *et al.*, 2022] Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao, and Xinge You. MSDN: Mutually semantic distillation network for zero-shot learning. In *CVPR*, pages 7612–7621, 2022.

[Chou *et al.*, 2020] Yu-Ying Chou, Hsuan-Tien Lin, and Tyng-Luh Liu. Adaptive and generative zero-shot learning. In *ICLR*, 2020.

[Farhadi *et al.*, 2009] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785. IEEE, 2009.

[Felix *et al.*, 2018] Rafael Felix, Ian Reid, Gustavo Carneiro, et al. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, pages 21–37, 2018.

[Frome *et al.*, 2013] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NeurIPS*, pages 2121–2129, 2013.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014.

[Gulrajani *et al.*, 2017] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

[Hadsell *et al.*, 2006] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742. IEEE, 2006.

[Han *et al.*, 2020] Zongyan Han, Zhenyong Fu, and Jian Yang. Learning the redundancy-free features for generalized zero-shot object recognition. In *CVPR*, pages 12865–12874, 2020.

[Han *et al.*, 2021] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *CVPR*, pages 2371–2381, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hu *et al.*, 2021] Runze Hu, Yutao Liu, Ke Gu, Xiongkuo Min, and Guangtao Zhai. Toward a no-reference quality metric for camera-captured images. *IEEE Transactions on Cybernetics*, 2021.

[Huynh and Elhamifar, 2020] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, pages 4483–4493, 2020.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kodirov *et al.*, 2017] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 3174–3183, 2017.

[Kong *et al.*, 2022] Xia Kong, Zuodong Gao, Xiaofan Li, Ming Hong, Jun Liu, Chengjie Wang, Yuan Xie, and Yanyun Qu. En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning. In *CVPR*, pages 9306–9315, 2022.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25:1097–1105, 2012.

[Lampert *et al.*, 2013] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 36(3):453–465, 2013.

[Li *et al.*, 2019a] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *CVPR*, pages 7402–7411, 2019.

[Li *et al.*, 2019b] Kai Li, Martin Renqiang Min, and Yun Fu. Rethinking zero-shot learning: A conditional visual classification perspective. In *CVPR*, pages 3583–3592, 2019.

[Liu *et al.*, 2020] Bo Liu, Qiulei Dong, and Zhanyi Hu. Zero-shot learning from adversarial feature residual to compact visual feature. In *AAAI*, volume 34, pages 11547–11554, 2020.

[Liu *et al.*, 2021] Zhe Liu, Yun Li, Lina Yao, Xianzhi Wang, and Guodong Long. Task aligned generative meta-learning for zero-shot learning. In *AAAI*, pages 8723–8731, 2021.

[Ma and Hu, 2020] Peirong Ma and Xiao Hu. A variational autoencoder with deep embedding model for generalized zero-shot learning. In *AAAI*, volume 34, pages 11733–11740, 2020.

[Min *et al.*, 2020] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Zheng-Jun Zha, and Yongdong Zhang. Domain-aware visual bias eliminating for generalized zero-shot learning. In *CVPR*, pages 12664–12673, 2020.

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[Narayan *et al.*, 2020] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *ECCV*, pages 479–495. Springer, 2020.

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[Norouzi *et al.*, 2013] M Norouzi, T Mikolov, S Bengio, Y Singer, J Shlens, A Frome, GS Corrado, and J Dean. Zero-shot learning by convex combination of semantic embeddings. arxiv 2013. *arXiv preprint arXiv:1312.5650*, 2013.

[Patterson and Hays, 2012] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758. IEEE, 2012.

[Reed *et al.*, 2016] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016.

[Romera-Paredes and Torr, 2015] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015.

[Sariyildiz and Cinbis, 2019] Mert Bulent Sariyildiz and Ramazan Gokberk Cinbis. Gradient matching generative networks for zero-shot learning. In *CVPR*, pages 2168–2178, 2019.

[Schonfeld *et al.*, 2019] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, pages 8247–8255, 2019.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[Van den Oord *et al.*, 2018] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.

[Verma *et al.*, 2020] Vinay Kumar Verma, Dhanajit Brahma, and Piyush Rai. Meta-learning for generalized zero-shot learning. In *AAAI*, volume 34, pages 6062–6069, 2020.

[Welinder *et al.*, 2010] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[Wen *et al.*, 2016] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016.

[Xian *et al.*, 2016] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016.

[Xian *et al.*, 2018a] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 41(9):2251–2265, 2018.

[Xian *et al.*, 2018b] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018.

[Xian *et al.*, 2019] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, pages 10275–10284, 2019.

[Xie *et al.*, 2019] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *CVPR*, pages 9384–9393, 2019.

[Yu *et al.*, 2020] Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang. Episode-based prototype generating network for zero-shot learning. In *CVPR*, pages 14035–14044, 2020.

[Yue *et al.*, 2021] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, pages 15404–15414, 2021.

[Zhang and Saligrama, 2015] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, pages 4166–4174, 2015.

[Zhang *et al.*, 2022] Yachao Zhang, Yuan Xie, Cuihua Li, Zongze Wu, and Yanyun Qu. Learning all-in collaborative multiview binary representation for clustering. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022.