Prototypical Model with Information-theoretic Loss Functions for Generalized Zero-Shot Learning

Chunlin Ji✉ CHUNLIN.JI@KUANG-CHI.ORG
Zhan Xiong ZHAN.XIONG@KUANG-CHI.COM
Kuang-Chi Institute of Advanced Technology, Shenzhen, China

Meiying Zhang ZHANGMY@SUSTECH.EDU.CN
Research Institute of Trustworthy Autonomous Systems
Southern University of Science and Technology, Shenzhen, China

Huiwen Yang HW.YANG@BERKELEY.EDU
University of California, Berkeley, CA, USA

Feng Chen FENG.CHEN@KUANG-CHI.COM
Hanchun Shen HANCHUN.SHEN@KUANG-CHI.COM
Kuang-Chi Institute of Advanced Technology, Shenzhen, China

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Generalized zero-shot learning (GZSL) is still a technical challenge of deep learning. To preserve the semantic relation between source and target classes when only trained with data from source classes, we address the quantification of the knowledge transfer from an information-theoretic viewpoint. We use the prototypical model and format the variables of concern as a probability vector. Taking advantage of the probability vector representation, information measurements can be effectively evaluated with simple closed forms. We propose two information-theoretic loss functions: a mutual information loss to bridge seen data and target classes; an uncertainty-aware entropy constraint loss to prevent overfitting when using seen data to learn the embedding of target classes. Simulation shows that, as a deterministic model, our proposed method obtains state-of-the-art results on GZSL benchmark datasets. We achieve $21\% - 64\%$ improvements over the baseline model – deep calibration network (DCN) and demonstrate that a deterministic model can perform as well as generative ones. Furthermore, the proposed method is compatible with generative models and can noticeably improve their performance.

Keywords: Generalized zero-shot learning; probability vector; prototypical model; mutual information; uncertainty-aware.

1. Introduction

Deep neural networks have made remarkable progress in object recognition in recent years. However, most successful deep neural networks are trained under supervised learning frameworks, which always require a large amount of annotated data for each class (Deng et al., 2009). Inspired by human's ability to recognize objects without having seen visual samples, zero-shot learning (ZSL) has recently gained a surge of interest and has been used in broad applications (Zhang and Saligrama, 2015; Zhang et al., 2017; Xian et al., 2018a).

ZSL offers an elegant way to extend classifiers from source categories, of which labeled images are available during training, to target categories, of which labeled images are not accessible. The goal of ZSL is to recognize objects of target classes by transferring knowledge from source classes through the relation in the semantic space, while generalized zero-shot learning (GZSL), a more general and challenging scenario of ZSL, tries to recognize objects from the joint set of both source and target classes. Generally, methods for ZSL/GZSL can be categorized into two major categories - deterministic and generative: Deterministic methods focus on carefully designed models and semantic relations preserving the knowledge from source classes to target classes, using only the seen data from source classes; Generative methods leverage generative models/networks to transfer the knowledge of the paired relation between the semantic representation and visual feature of source classes, in order to generate the data for target classes. With the generated data, although less reliable, generative methods always achieve better performance than deterministic methods. Broad studies show that filling the performance gap between these two methods is a challenge.

Two technical problems as envisioned in deterministic ZSL/GZSL methods (Changpinyo et al., 2016; Liu et al., 2018): (i) how to bridge source classes to target classes for knowledge transfer and (ii) how to make predictions on target classes without labeled training data. Toward the first problem, deterministic ZSL/GZSL methods typically embed the image features and the semantic representations into a predefined common embedding space (with properly defined distances) using a regression model. The choice of embedding space and the regression model/neural network design are essential to inheriting the semantic relation while maintaining the discriminative ability. As for the second problem, we need to effectively bridge target classes to source classes and prevent overfitting when using the seen data of source classes, as they are blind to the semantic representations of target classes. The seminal work, deep calibration network (DCN) (Liu et al., 2018), introduces an entropy loss for target classes which brings their semantic representations close to the seen data of source classes. However, the entropy loss with a calibration parameter is inadequate to accurately control how much the target classes should learn from the seen data, preventing the DCN from obtaining superior performance.

In this work, we address the two problems discussed above from an information-theoretic point of view and make three major contributions: 1) we choose the visual space as the common embedding space and propose a probability vector (PV) representation, illustrated in Fig. 1. By considering the semantic representation of either source or target classes as clustering centroids, the position of a visual feature can be formatted into a soft assignment PV under the prototypical model (Snell et al., 2017): we characterize the position of the visual feature indirectly by measuring its assignment probability to the reference points. Given the PV representation, we can evaluate information-theoretic measurements in closed forms. Thus, the PV representation significantly benefits us by quantifying several intuitive relations in the GSZL setting into information-theoretic loss functions; 2) we propose a mutual information loss to link the semantic representations of target classes to seen data of source classes. The mutual information consists of two parts: the conditional entropy encourages the seen data to attach to a certain prototype/centroid of the target classes, while the entropy term preserving the semantic representations collapses to trivial solutions when projecting them to the visual space; 3) we propose an uncertainty-aware loss which prevents overfitting when using seen data of source classes to train the embedding model of

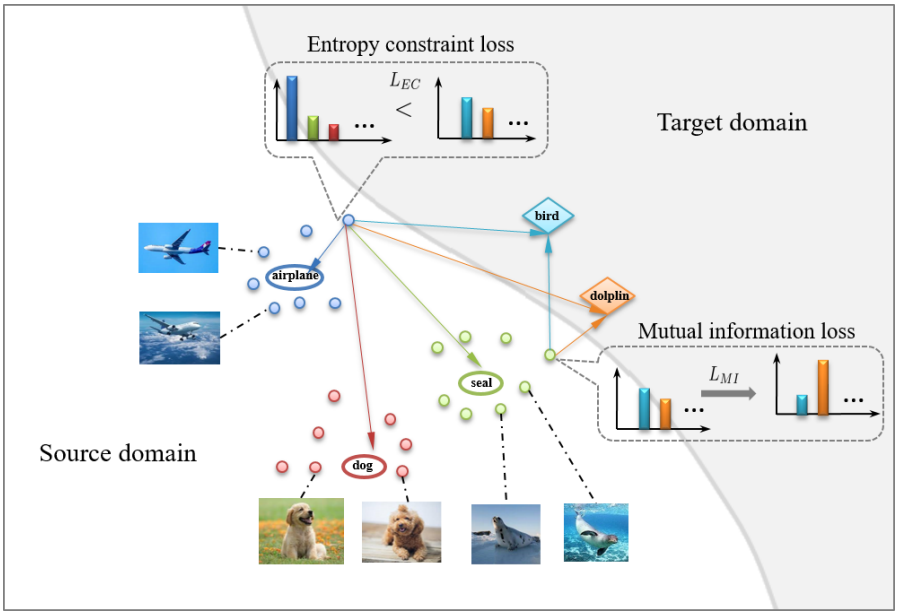

Figure 1: Illustration of the probability vector representation and information-theoretic loss functions. Circles and diamonds with text names denote the semantic representations of all classes (source and target classes respectively), small dots denote seen data of source classes, and unseen data of target classes are unavailable. The probability vector, illustrated by the series of color bars, represents the probability that the data is assigned to different clustering centroids, where the assignment is shown by the arrow. The proposed loss is intuitively illustrated by the change of bars in the PV representation.

target classes. We define a regularized entropy which allows us to control the uncertainty of the seen data when assigning them to source and target classes. With the proposed information-theoretic loss, the change in PV is illustrated by the bars shown in Fig. 1.

We evaluate the performance of our proposed methods on broadly studied benchmark datasets. Simulation shows that, as a deterministic GZSL model, our proposed method obtains SOTA results, significantly outperforms the recent deterministic models on all benchmark datasets, and even performs as well as generative ones. Moreover, our proposed model is compatible with generative models as well. We present additional loss functions to learn with generated data by considering their higher uncertainty than seen data. The experiments show that, by incorporating with the generated data from GANs (Xian et al., 2018b; Vyas et al., 2020), we also gain noticeable improvements over these GAN models.

2. Related works

Deterministic models for GZSL. Deterministic models try to utilize the knowledge of semantic embedding of both source and target classes sufficiently to make inferences on visual data. To this end, previous works typically embed visual samples and semantic embeddings into a common embedding space (Frome et al., 2013; Zhang et al., 2017), such as

the visual space, the semantic embedding space, or an intermediate space between semantic and visual domains. The choice of embedding space is critical for model performance. Previous works (Shigeto et al., 2015; Zhang et al., 2017) show that using visual space instead of semantic space or any other intermediate space as the common embedding space alleviates the negative effect of the hubness problem (Lazaridou et al., 2015). The choice of distance function in the common embedding space also plays an important role. In previous studies (Ravi and Larochelle, 2017; Liu et al., 2018), Euclidean distance, dot product similarity, and cosine similarity are broadly applied.

Most ZSL/GZSL methods tend to compensate the lack of visual representation of the unseen classes with the learning of a semantic preserving mapping. For instance, a fairly successful approach is based on a bilinear compatibility function that associates visual representation and semantic features, such as ALE (Akata et al., 2013), DEVISE (Frome et al., 2013), SJE (Akata et al., 2015) and ESZSL (Paredes and Torr, 2015). A straightforward extension is the exploration of a nonlinear compatibility function between visual and semantic spaces, such as a ridge regression (Shigeto et al., 2015). Furthermore, in Annadani and Biswas (2018), they introduce explicit regularization for semantic preserving but require an additional threshold for similarity. In DCN (Liu et al., 2018), they introduce an entropy loss to allow the embedding network of target classes trained by seen data, and a calibration parameter is required to balance the training of source classes and target classes. In our work, we introduce a series of information-theoretic loss functions that enable the use of nonlinear compatibility functions. Meanwhile, these functions allow us to translate several intuitive assumptions on the semantic relation to easy computing formulas. Moreover, the conditional entropy in our mutual information loss is consistent with the entropy loss in DCN, while the new marginal entropy term in our loss makes additional effects to encourage cluster balancing. The proposed MI loss is a more effective term to bridge the seen data and unseen class semantic embedding than the conditional entropy used in DCN. Furthermore, although MI loss has been explored in Han et al. (2020), they use a variational upper bound of MI as a surrogate instead of a direct evaluation of MI. In our work, leveraging the prototypical model based probability vector representation, we can directly evaluate both the MI and entropy terms with closed-form expressions. Furthermore, our method significantly outperforms the work (Han et al., 2020) for all common datasets.

Generative models for GZSL. Generative models have the advantage of utilizing generated image features to remove blindness caused by the inaccessibility of target classes' data during training. Variational Autoencoders (VAE) (Kingma and Welling, 2014) and conditional VAE (Sohn et al., 2015) based generative models are proposed with the aim of aligning the visual embedding with the semantic embedding (Schonfeld et al., 2019). A VAE based algorithm can train stably but fails to capture the complex distribution (Bao et al., 2017), leading to unsatisfied results. The generative adversarial network (GAN) (Goodfellow et al., 2014) has the advantage of generating more diverse data. The f-CLSWGAN (Xian et al., 2018b) is the model including a variant of an improved WGAN (Arjovsky et al., 2017) and a softmax classifier. f-CLSWGAN synthesizes visual features conditioned on semantic representations, offering a shortcut directly from a semantic descriptor to a class-conditional feature distribution. Recently, advanced GAN models (Li et al., 2019a; Narayan et al., 2020; Vyas et al., 2020) have been proposed that have outstanding performance in GZSL. Despite the favorable performance, GAN always suffers from mode collapse problems and has an

unstable training phase. Fortunately, an improved deterministic model incorporated with a generative model has shown advanced performance (Tong et al., 2019). In this work, we notice that the synthetic data from the generative model are generally less reliable than the seen data, so the proposed uncertainty-aware entropy constraint loss is also applicable here. Thus, instead of constructing a complex generative model, we train the proposed model additionally with generated data from known GANs, and obtain competitive results compared to recent advanced generative models.

3. Generalized Zero-shot learning

Following the notation in Liu et al. (2018), we first present the definition of generalized zero-shot learning as follows: suppose that we have the seen data $\mathcal{D} = \left\{(x^{(n)}, y^{(n)})\right\}_{n=1}^{N}$, where $x^{(n)} \in \mathbb{R}^P$ is the feature of the n-th image in the visual space $\mathbb{R}^P$ and $y^{(n)} \in \mathcal{S}$ is the label from the source classes $\mathcal{S} = \{1, ..., S\}$. In this study, we assume that the image feature x (also named visual embedding) has already been extracted by a pre-trained deep convolutional network, such as ResNet (He et al., 2016). Let $\mathcal{T} = \{S+1, ..., S+T\}$ denote the target classes, where no seen data is available in the training phase. For each class $c \in \mathcal{S} \cup \mathcal{T}$, let $v_c \in \mathbb{R}^Q$ denote the semantic representation in the semantic space $\mathbb{R}^Q$, such as word embedding generated by Word2Vec (Mikolov et al., 2013) or visual attributes annotated by humans to describe visual patterns (Lampert et al., 2014), and let $\mathcal{V} = \{v_c\}_{c=1}^{S+T}$ denote the set of semantic representations. In the test phase, we predict unseen data $\mathcal{D}' = \left\{x^{(m)}\right\}_{m=N+1}^{N+M}$ of M points from either source or target classes. The task of ZSL is that, given $\mathcal{D}$ and $\{v_c\}_{c=1}^{S}$, learn a model $\phi : x \to y$ to classify $\mathcal{D}'$ over target classes $\mathcal{T}$. The task of GZSL is that, given $\mathcal{D}$ and $\{v_c\}_{c=1}^{S+T}$ of both source and target classes, learn a model $f : x \to y$ to classify $\mathcal{D}'$ over both source and target classes $\mathcal{S} \cup \mathcal{T}$.

4. Proposed methods

4.1. Prototype model

To link the visual embedding of the seen data to the class semantic representations, an intuitive way is to view the semantic representations (or their projection in another space) as the centroids of their corresponding classes and learn to push the visual embedding to surround the centroid of its belonging class. In this work, we utilize the prototypical model/networks (Snell et al., 2017) to realize this goal. Prototypical networks learn a metric space in which classification can be performed by computing distances between samples and the prototype representation (or centroid) of each class. Under the GZSL settings, we assume that the semantic representation v_c or its projection by a network or a liner model $\psi(v_c)$ in a common embedding space, $\mathbb{R}^K$, is the prototype of each class. For the image feature x, we assume a network $\phi(x)$ to transform the image feature to the same space of the prototype $\psi(v_c)$. Given a distance function $d : \mathbb{R}^K \times \mathbb{R}^K \to [0, +\infty)$ to measure the distances between the samples and the prototypes, the prototype model produces a soft assignment probability vector (PV), $\boldsymbol{p} = [p_1(y = 1|x), ..., p_C(y = C|x)]^T$,

over the prototypes of each class for the data sample x,

$$p_c(y = c|x) = \frac{\exp[-d(\phi(x), \psi(v_c))]}{\sum_{c'} \exp[-d(\phi(x), \psi(v_{c'}))]}. \tag{1}$$

In this formulation of soft assignment PV $\boldsymbol{p}$, two concerns are presented, namely the selection of the specified embedding metric space and the specification of the distance function in that space.

4.1.1. Choice of common embedding space

Motivated by previous works (Shigeto et al., 2015; Zhang et al., 2017), we map the semantic representations into the visual space such that the semantic relation between the mapped semantic representations and the visual features reflect the relation between their corresponding classes. We propose a *multilayer perceptrons (MLP)* (Rumelhart et al., 1986) as the compatibility function to map semantic representations to the visual space $\psi : v_c \to z$, where $z \in \mathbb{R}^P$. Therefore, the soft assignment PV $\boldsymbol{p}$ expression becomes

$$p_c(y = c|x) = \frac{\exp[-d(x, \psi(v_c))]}{\sum_{c'} \exp[-d(x, \psi(v_{c'}))]}. \tag{2}$$

In previous works (Akata et al., 2013; Frome et al., 2013; Tong et al., 2019), they use a linear mode to project semantic representations into another space, since a linear model makes it easy to maintain semantic relationships. However, MLP is more flexible and can learn the nonlinear relation between the original semantic representations and the mapping in the visual space. To prevent unreasonable nonlinear transform, we will introduce the information-theoretic loss functions in Section 4.2.

4.1.2. Choice of distance function

The distance function $d(\cdot, \cdot)$ plays another important role in the prototypical model, while the Euclidean distance, cosine similarity, and dot product similarity based distances have been used in previous works (Snell et al., 2017; Ravi and Larochelle, 2017; Liu et al., 2018). Generally, when we map the semantic representations to the visual space, it is a too strong assumption that the mapped semantic representation can be well aligned to the visual feature under Euclidean distance. Thus, in this work, we utilize the dot product similarity based distance, besides, dot product similarity has more degrees of freedom than the cosine similarity. In the simulation study (Section 5.3), we show the advanced performance of the dot product based distance function.

Furthermore, as will be addressed in Section 4.2, when we bridge the semantic embedding of target classes to the seen data, the uncertainty of the learned model should be higher than that of learning the semantic embedding of source classes by the seen data. To reflect this viewpoint, we propose an asymmetrical dot product based distance

$$d(x, \psi(v_c)) = -\max\{m < x, \psi(v_c) >, 0\} \tag{3}$$

where $m = \mathrm{m}_1$ when $c \in \mathcal{S}$ and $m = \mathrm{m}_2$ when $c \in \mathcal{T}$. The setting of m is similar to the calibration parameter ρ in the DCN model (Liu et al., 2018), which was introduced to balance the confidence of source classes and the uncertainty of target classes.

4.1.3. Cross entropy loss for seen data

Given the expression of PV, $\boldsymbol{p} = [p_1(y = 1|x), ..., p_S(y = S|x)]^T$ with $p_c(y = c|x) = \frac{\exp[-d(x,\psi(v_c))]}{\sum_{c'=1}^{S}\exp[-d(x,\psi(v_{c'}))]}$ for each $c \in \mathcal{S}$, and the label y of the seen data from source classes, we can define the loss function to train the prototypical model. To be specific, given the seen data $x^{(n)} \in \mathcal{D}$ from source classes $\mathcal{S}$, we can learn the proposed mapping network $\psi(\cdot)$ by minimizing the cross entropy loss,

$$L_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{S} y_c^{(n)} \log p_c(x^{(n)}). \tag{4}$$

However, only the cross entropy loss is insufficient to train a prototypical model for GSZL. Therefore, we propose information-theoretic loss functions to boost the performance of a deterministic prototypical model.

4.2. Information-theoretic loss functions

The proposed information-theoretic loss functions are specified as follows: 1) to bridge the source and target classes through the seen data of source classes, we propose the mutual information loss; 2) to reflect the factor that the seen data should be closer to prototypes of sources classes rather than target classes, we propose an entropy constraint loss.

4.2.1. Mutual information loss to link seen data and target classes

To link the semantic embedding of target classes to visual images of source classes, we leverage an intuitive factor that each seen image can be classified as a target class that is most similar to the image's label in the source classes, rather than being classified to each target class with an equal assignment probability (or say equal uncertainty) (Liu et al., 2018). Here, we translate this intuitive factor into a formal information-theoretic measurement and let the mutual information $\mathrm{MI}(x, c)$, quantify the relationship (or say closeness) between the seen data x and the target class c. With the prototypical model discussed in Section 4.1, we can also obtain the probability vector that the seed data x belong to the prototypes of target classes, $p_c(x) = \frac{\exp[-d(x,\psi(v_c))]}{\sum_{c'=S+1}^{S+T}\exp[-d(x,\psi(v_{c'}))]}$. To bridge the seen data and the prototypes of target classes, we minimize the MI loss as follows,

$$\begin{aligned} L_{\mathrm{MI}} &= -\mathrm{MI}(x,c) = -\mathrm{H}(c) + \mathrm{H}(c|x) = \sum_c P_c \log P_c + \mathbb{E}_x\left[-\sum_c p_c(x)\log p_c(x)\right] \\ &\approx \sum_{c=S+1}^{S+T}\left(\frac{1}{N}\sum_{n=1}^{N} p_c(x^{(n)})\right)\log\left(\frac{1}{N}\sum_{n=1}^{N} p_c(x^{(n)})\right) - \frac{1}{N}\sum_{n=1}^{N}\sum_{c=S+1}^{S+T} p_c(x^{(n)})\log p_c(x^{(n)}) \end{aligned} \tag{5}$$

where $\mathbb{E}_x[\cdot]$ denotes the expectation with respect to x, which is always approximated by the Monte Carlo approach, as sample $\{x^{(n)}\}_{n=1}^{N}$ are available here. $P_c = \mathbb{E}_x[p_c(x^{(n)})] \approx \frac{1}{N}\sum_{n=1}^{N} p_c(x^{(n)})$. P_c can be viewed as a marginal assignment probability that a sample data belongs to a number of T target classes. Furthermore, increasing the marginal entropy $\mathrm{H}(c)$ encourages cluster balancing, which prevents trivial solutions that map the semantic embedding of all target classes to a single prototype in the visual space.

The second term $\mathrm{H}(c|x)$ in eqn (5), commonly called conditional entropy, measures the uncertainty that a seen visual feature belongs to the target classes. Conditional entropy, here indicated by $L_{\mathrm{Ent}}^{org} \triangleq -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=S+1}^{S+T} p_c(x^{(n)})\log p_c(x^{(n)})$, can significantly improve prediction over target classes while having little harm on classifying seen data (Liu et al., 2018). Here, we further introduce a margin for this conditional entropy, that

$$L_{\mathrm{Ent}} = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{-1}{\log_2(T)}\sum_{c=S+1}^{S+T} p_c(x^{(n)})\log p_c(x^{(n)}) - \mathrm{margin}_1\right]_+ \tag{6}$$

where the term $\log_2(T)$ denotes the information capacity of T bits and $[x]_+ := \max\{0, x\}$. Here, we propose to regularize the entropy by dividing the information capacity term, as a consequence, the resulting *regularized* entropy varies only in a small fixed interval $(0, C_0]$, where $C_0 = \log(n)/\log_2(n)$, $\forall n > 1.0$ and $n \in R$. Therefore, the selection of margin_1 becomes easy and consistent, even though the number of target classes varies among different datasets. We also apply regularization for the marginal entropy $\mathrm{H}(c)$, by dividing the term $\log_2(T)$. Finally, the improved MI loss becomes,

$$L_{\mathrm{MI}} = \frac{1}{\log_2(T)}\sum_{c=S+1}^{S+T} P_c\log P_c + L_{\mathrm{Ent}} \tag{7}$$

4.2.2. Entropy constraint loss for uncertainty-aware training

When training the embedding network of target classes using the seen data from source classes, a notable factor is that the data seen, to some extent, are out of distribution data for target classes. Therefore, the uncertainty of classifying the seen data to target classes should be larger than that of classifying them to source classes. By counting this factor, we propose an information constraint loss to control the entropy of the seen image with respect to the prototypes of source classes to be less than that with respect to the prototypes of target classes. Let us define the *regularized* entropy terms as follows:

$$\mathrm{E_u} = \frac{-1}{\log_2(T)}\sum_{c=S+1}^{S+T} p_c(x^{(n)})\log p_c(x^{(n)})$$

where $p_c(x) = \frac{\exp[-d(x,\psi(v_c))]}{\sum_{c'=S+1}^{S+T}\exp[-d(x,\psi(v_{c'}))]}$ is the PV that assigns the seen data to each prototype of the target classes and

$$\mathrm{E_s} = \frac{-1}{\log_2(S)}\sum_{c=1}^{S} p_c(x^{(n)})\log p_c(x^{(n)})$$

where $p_c(x) = \frac{\exp[-d(x,\psi(v_c))]}{\sum_{c'=1}^{S}\exp[-d(x,\psi(v_{c'}))]}$ is the PV that assigns the seen data to each prototype of the source classes. Then, the uncertainty-aware entropy constraint loss is defined as

$$L_{\mathrm{EC}} = \frac{1}{N}\sum_{n=1}^{N}[\mathrm{E_u} - (\mathrm{E_s} + \mathrm{margin}_2)]_+ \tag{8}$$

This loss reflects the expectation that the entropy $\mathrm{E_u}$ should be greater than the sum of $\mathrm{E_s}$ plus a margin margin_2. As discussed in the previous section, the regularized entropy $\mathrm{E_u}$ and $\mathrm{E_s}$ vary in the interval $(0, C_0]$, so it is not difficult to set a proper value for margin_2.

4.3. Learning and inference

We combine all loss functions with different weights $\lambda_{1:2}$. Therefore, we optimize the parameters of our prototypical model by jointly learning the following loss functions:

$$L_D = L_{\text{CE}} + \lambda_1 L_{\text{MI}} + \lambda_2 L_{\text{EC}} \tag{9}$$

The network parameters in $\psi(\cdot)$ can be efficiently optimized using the SGD or Adam algorithm with the auto-differentiation technique supported in PyTorch (Paszke et al., 2017).

In the test stage, the predicted class $y(x^{(n)})$ of the image feature $x^{(n)}$ is given by $y(x^{(n)}) = \text{argmax} p_c(x^{(n)})$, where $p_c(x^{(n)}) = \frac{\exp[-d(x,\psi(v_c))]}{\sum_{c'}\exp[-d(x,\psi(v_{c'}))]}$ and $\psi(\cdot)$ is the trained network that maps semantic embedding to the visual feature space. Therefore, the prediction is made over both source and target classes, as $c \in \mathcal{S} \cup \mathcal{T}$ in generalized zero-shot learning. In conventional zero-shot learning, we only need the prediction over target classes $c \in \mathcal{T}$.

4.4. Cooperate with generative models

Most generative methods emphasize the development of a sophisticated model to generate more 'realistic' data for target classes. However, effective utilization of generated data is still largely ignored. We noticed that the synthetic data from the generative model are generally less reliable than the seen data, so we propose an uncertainty-aware entropy constraint to select the generated data before they are applied in training the discriminative model. Specifically, we put the generated data $\{\widetilde{x}^{(m)}\}_{m=1}^{M}$ in the prototypical model where the prototypes are from target classes and obtain the PV, $p_c(\widetilde{x}^{(m)}) = \frac{\exp[-d(\widetilde{x}^{(m)},\psi(v_c))]}{\sum_{c'=S+1}^{S+T}\exp[-d(\widetilde{x}^{(m)},\psi(v_{c'}))]}$. Then, we define the uncertainty of the generated data by the regularized entropy, $\widetilde{\text{E}}_{\text{u}}(\widetilde{x}^{(m)}) = \frac{-1}{\log_2(T)}\sum_{c=S+1}^{S+T} p_c(\widetilde{x}^{(m)})\log p_c(\widetilde{x}^{(m)})$. After that, we select the generated data using the criterion that $\widetilde{\text{E}}_{\text{u}}(\widetilde{x}^{(m)}) < \text{margin}_3$ with a predefined threshold margin_3. This uncertainty based selection can prevent improperly generated data from negatively affecting the prediction of target classes. Let $\widetilde{x}_{sel} = \{\widetilde{x}_{sel}^{(m)}\}_1^{M_s}$ denote the selected generated data, then we can train the prototypical model using these data via a cross entropy loss,

$$\widetilde{L}_{\text{CE}}(\widetilde{x}) = -\frac{1}{M_s}\sum_{m=1}^{M_s}\sum_{c=S+1}^{S+T}\widetilde{y}_c^{(m)}\log p_c(\widetilde{x}_{sel}^{(m)}), \tag{10}$$

where the label $\widetilde{y}_c$ ($c \in \mathcal{T}$) is known in the generation of the data.

Moreover, we can also link the generated data to prototypes of source classes. To this end, we define another mutual information loss as eqn (5), the difference is that the prototypes change from target classes to source classes, $p_c(\widetilde{x}) = \frac{\exp[-d(\widetilde{x},\psi(v_c))]}{\sum_{c'=1}^{S}\exp[-d(\widetilde{x},\psi(v_{c'}))]}$, that

$$\widetilde{L}_{\text{MI}}(\widetilde{x}) = \sum_{c=1}^{S}\left(\frac{1}{M_s}\sum_{m=1}^{M_s}p_c(\widetilde{x}_{sel}^{(m)})\right)\log\left(\frac{1}{M_s}\sum_{m=1}^{M_s}p_c(\widetilde{x}_{sel}^{(m)})\right) - \frac{1}{M_s}\sum_{m=1}^{M_s}\sum_{c=1}^{S}p_c(\widetilde{x}_{sel}^{(m)})\log p_c(\widetilde{x}_{sel}^{(m)}) \tag{11}$$

Finally, combine all loss functions, eqs (9), (10) and (11) weighted by γ_1 and γ_2, and then we obtain the loss to train the proposed model with both seen data x and generated data $\widetilde{x}$,

$$L_G = L_D + \gamma_1\widetilde{L}_{\text{CE}}(\widetilde{x}) + \gamma_2\widetilde{L}_{\text{MI}}(\widetilde{x}). \tag{12}$$

5. Experiments

5.1. Experimental settings

Datasets. The benchmark datasets are briefly described as follows: Animals with Attributes (AwA1) (Lampert et al., 2014) is a widely used dataset for ZSL/GZSL, which contains 30,475 images from 50 different animal classes. A standard split into 40 source classes and 10 target classes is provided in Lampert et al. (2014). A variant of this dataset is Animal with Attributes2 (AwA2) (Xian et al., 2017) which has the same 50 classes as AwA1, but AwA2 has 37,322 images in all, which do not overlap with images in AwA1. Caltech-UCSD-Birds-200-2011 (CUB) (Wah et al., 2011) is a fine-grained dataset with a large number of classes and attributes, containing 11,788 images from 200 different types of birds annotated with 312 attributes. The split of CUB with 150 source classes and 50 target classes is provided in Akata et al. (2016). SUN Attribute (SUN) (Patterson and Hays, 2012) is another fine-grained dataset, containing 14,340 images from 717 types of scenes annotated with 102 attributes. The split of SUN with 645 source classes and 72 target classes is provided in Lampert et al. (2014). Attribute Pascal and Yahoo (aPY) (Farhadi et al., 2009) is a small-scale dataset with 64 attributes and 32 classes(20 Pascal classes as source classes and 12 Yahoo classes as target classes).

Image features. Due to the variations in image features used by different zero-shot learning methods, for a fair comparison, we use the widely used features: 2048-dimensional ResNet-101 features provided by Xian et al. (2018a).

Semantic representations. We use the per-class continuous attributes provided with the datasets of aPY, AwA, CUB and SUN. Note that we can also use Word2Vec representations as class embeddings (Mikolov et al., 2013).

5.2. Implementation Details

The compatibility function in the prototypical model is implemented as MLP. The input dimension of the attribute embedding depends on the problem. The MLP has 2 fully connected layers with 2048 hidden units. We use LeakyReLU as the nonlinear activation function, Dropout function, for the first layer, and Tanh for the output layer to squash the predicted values within $[-1, 1]$. For the distance eqn (3) in the prototypical model, experimentally we set $m_1 = 0.5$ and $m_2 = 1.0$ for the asymmetric dot product distance. The batch size of the visual feature is set to 512. For optimization, we use Adam optimizer (Kingma and Ba, 2015) with a constant learning rate 0.001.

We design a cross-validation for the selection of hyperparameters in loss functions: first, we divide the source classes $\mathcal{S}$ into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ where we keep $\frac{\#\mathcal{S}_1}{\#\mathcal{S}_2} \approx \frac{\#\mathcal{S}}{\#\mathcal{T}}$, where $\#\mathcal{S}$ represents the number of elements in the set $\mathcal{S}$; then we discard the seen data (image features) of $\mathcal{S}_2$ and treat them as target classes denoted by $\mathcal{T}_1$; let the seen data of $\mathcal{S}_1$, the semantic representation of both $\mathcal{S}_1$ and $\mathcal{T}_1$ to format another GZSL problem and choose hyperparameters that achieve the best performance in this new GZSL setting; finally we use the chosen hyperparameters for the original GZSL problem. By this cross-validation process, we choose the value of λ_2 as 0.5 for all datasets; while the value of λ_1 depends loosely on the data sets, we choose 0.05 for dataset AwA1/2, aPY, CUB, and a larger value 0.5 for dataset SUN. The margin value is chosen as follows: $\text{margin}_1 = 0.15$ and $\text{margin}_2 = 0.05$ for dataset

Table 1: Comparison of the contribution of different approaches

Methods	AWA1	AWA2	CUB	SUN	aPY
L_{CE}+L_{Ent}^{org} (DCN (Liu et al., 2018))	-	39.1	38.7	30.2	23.9
L_{CE} (*space C*)	20.2	23.7	35.9	29.6	17.4
+L_{Ent} (*space A*)	46.3	51.7	41.8	31.6	32.4
+L_{Ent} (*space B*)	48.3	54.2	42.8	34.7	32.4
+L_{Ent} (*space C*)	50.8	56.2	43.3	36.3	31.5
+L_{MI} (*space C*)	51.1	58.2	45.9	38.6	33.7
+L_{MI}+L_{EC}(*space C*)	**55.7**	**61.6**	**46.6**	**39.5**	**37.8**

aPY, CUB and SUN, while $\text{margin}_1 = 0.15$ and $\text{margin}_2 = 0.0$ for AwA1 and AwA2. For hyperparameters in the loss function eqn (12) when incorporated with generated data, we use a similar cross-validation process and obtain the hyperparameter setting: $\text{margin}_3 = 0.5$, $\gamma_1 = 0.2$ and $\gamma_2 = 0.01$.

Following the Proposed Split (PS) in the Rigorous Protocol (Xian et al., 2017), we compare three accuracies: ACC_{ts}, the accuracy of all unseen images in target classes; ACC_{tr}, the accuracy of some seen images from source classes that are not used for training;the harmonic mean of the two accuracies as $ACC_H = 2(ACC_{ts} * ACC_{tr})/(ACC_{ts} + ACC_{tr})$, which is used as the final criterion to favor high accuracies on both source and target classes.

5.3. Ablation Study

We investigate how each strategy in the proposed approach contributes to the model performance for GZSL. We include the result of the DCN model and the prototypical model trained with only cross entropy loss as baseline methods. We compare the choice of common embedding spaces, attribute space, and feature space, and the choice of different distance functions. We represent the combinations as follows: *space A* uses the attribute space as the embedding space and uses dot product similarity based distance; *space B* uses the visual space as the embedding space and uses cosine similarity based distance; *space C* uses the visual space as the embedding space and uses the distance based on dot product similarity. Furthermore, we show the contribution of different losses: L_{Ent}, L_{MI} and L_{EC}.

All simulation results are shown in Table 1, where we quote the result of the DCN directly from. The DCN model introduces an entropy regularization L_{Ent}^{org} for bridging seen data and target classes, which is similar to the L_{Ent}. The DCN uses the dot product distance and projects the feature and attribute onto a common space. The third to fifth rows show that the entropy loss L_{Ent} significantly improves the performance of GZSL, compared to the result obtained with only the cross entropy loss L_{CE}. It should be noted that the results of the third to fifth rows significantly outperform the DCN, which could be due to the factors that it is easier to train the model using the original attribute/feature space rather than looking for a common space, and the proposed entropy loss L_{Ent} seems more effective than the entropy regularization L_{Ent}^{org} in the DCN model. Furthermore, the third to the fifth rows demonstrate the importance of choosing the common embedding space and distance function: using the visual feature space rather than the semantic space as the embedding space gains strong improvement; using dot product similarity based distance yields better performances than using cosine similarity based distance. Comparing the sixth row with the fifth row, we show that the proposed marginal entropy $\text{H}(c)$ in L_{MI} brings additional

Table 2: Results of conventional zero-shot learning

Method	SUN	CUB	AwA2	aPY
DAP(Lampert et al., 2014)	39.9	40.0	46.1	33.8
CONSE(Mohammad et al., 2014)	38.8	34.3	44.5	26.9
ALE(Akata et al., 2013)	58.1	54.9	62.5	39.7
DEVISE(Frome et al., 2013)	56.5	52.0	59.7	39.8
SJE(Akata et al., 2015)	53.7	53.9	61.9	32.9
ESZSL(Paredes and Torr, 2015)	54.5	53.9	58.6	38.3
SYNC(Changpinyo et al., 2016)	40.3	55.6	46.6	23.9
PSR(Annadani and Biswas, 2018)	61.4	56.0	63.8	38.4
DLFZRL(Tong et al., 2019)	59.3	**57.8**	63.7	44.5
Proposed	**62.1**	57.6	**64.6**	**44.7**

improvements. Comparing the results in the seventh and sixth rows, we show that the entropy constraint loss L_{EC} further boosts the model performance.

5.4. Conventional zero-shot learning results

We investigate the proposed method for conventional ZSL that only recognizes target classes in the test stage. And we compare the result of our method with several state-of-the-art results from recent works (Akata et al., 2015; Paredes and Torr, 2015; Changpinyo et al., 2016; Annadani and Biswas, 2018; Tong et al., 2019). As shown in Table 2, the proposed approach compares favorably with existing approaches in the literature, obtaining state-of-the-art results on SUN, AwA2 and aPY datasets. On the CUB dataset, our result is 2% lower than DLFZRL (Tong et al., 2019).

5.5. Generalized zero-shot learning results

Comparison with deterministic models.

We compare the performance of our proposed model with several recent deterministic models for GZSL. Taking DCN (Liu et al., 2018) as the baseline model, as shown in Table 3, our method gains superior accuracy compared to other deterministic models on all datasets: it obtains $21\% - 64\%$ improvements over DCN and significantly outperforms a previous state-of-the-art deterministic model-DLFZRL (Tong et al., 2019). In addition, we observe that our deterministic model obtains results comparable to some generative models, such as f-CLSWGAN and DLFZRL+softmax.

Comparison with generative models.

We investigate the performance of our proposed method incorporating three generative models: a baseline model: f-CLSWGAN (Xian et al., 2018b), and two advanced models: LsrGAN (Vyas et al., 2020), TFVAEGAN (Narayan et al., 2020). Seen image features and class-level attributes are used to train GAN models, and image features of unseen classes can be generated by unseen class-level attributes. Including the generated features for the target classes into the whole training, we train the model by the loss function defined in eqn (12). As shown in Table 3, our proposed model significantly outperforms the baseline model f-CLSWGAN with $7\% - 13\%$ improvements. Compared to two similar

Table 3: Results of Generalized Zero-Shot Learning on four datasets under PS.

Methods	AwA2			CUB			SUN			aPY		
	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H
Non-Generative Models												
ALE(Akata et al., 2013)	16.8	76.1	27.5	23.7	62.8	34.4	21.8	33.1	26.3	4.6	73.7	8.7
DeViSE(Frome et al., 2013)	13.4	68.7	22.4	23.8	53.0	32.8	16.9	27.4	20.9	4.9	76.9	9.2
ZSKL(Zhang and Koniusz, 2018)	18.9	82.7	30.8	21.6	52.8	30.6	20.1	31.4	24.5	10.5	76.2	18.5
DCN(Liu et al., 2018)	25.5	84.2	39.1	28.4	60.7	38.7	25.5	37.0	30.2	14.2	75.0	23.9
DLFZRL(Tong et al., 2019)	-	-	45.1	-	-	37.1	-	-	24.6	-	-	31.0
PQZRL(Li et al., 2019b)	31.7	70.9	43.8	43.2	51.4	46.9	27.9	64.1	35.1	35.3	35.2	38.8
Proposed	52.7	74.1	**61.6**	41.6	54.3	**47.1**	41.7	37.4	**39.5**	30.1	50.5	**37.8**
Generative Models												
f-CLSWGAN(Xian et al., 2018b)	52.1	68.9	59.4	43.7	57.7	49.7	42.6	36.6	39.4	-	-	-
DLFZRL+softmax(Tong et al., 2019)	-	-	60.9	-	-	51.9	-	-	42.5	-	-	38.5
F-VAEGAN-D2(Xian et al., 2019)	57.6	70.6	63.5	48.4	60.1	53.6	45.1	38.0	41.3	-	-	-
CADA-VAE(Schonfeld et al., 2019)	55.8	75.0	63.9	51.6	53.5	52.4	47.2	35.7	40.6	-	-	-
GDAN(Huang et al., 2019)	32.1	67.5	43.5	39.3	66.7	49.5	38.1	89.9	***53.4***	30.4	75.0	43.4
LisGAN(Li et al., 2019a)	52.6	76.3	62.3	46.5	57.9	51.6	42.9	37.8	40.2	-	-	-
LsrGAN (Vyas et al., 2020)	54.6	74.6	63.0	48.1	59.1	53.0	44.8	37.7	40.9	-	-	-
RFF-GZSL(Han et al., 2020)	59.8	75.1	66.5	52.6	56.6	54.6	45.7	38.6	41.9	-	-	-
TF-VAEGAN(Narayan et al., 2020)	59.8	75.1	66.6	52.8	64.7	58.1	45.6	40.7	43.0	-	-	-
DR-VAE(Li et al., 2021)	56.9	80.2	66.6	51.1	58.2	54.4	36.6	47.6	41.4	-	-	-
DGN(Xu et al., 2021)	60.1	76.4	67.3	53.8	61.9	57.6	48.3	37.4	42.1	36.5	61.7	45.9
IZF(Shen et al., 2020)	60.6	77.5	68.0	52.7	68.0	*59.4*	52.7	57.0	*54.8*	42.3	60.5	**49.8**
IAS(Chou et al., 2021)	65.1	78.9	*71.3*	41.4	49.7	45.2	29.9	40.2	34.3	35.1	65.5	45.7
ZLAP-GZSL(Chen et al., 2022)	64.8	80.9	**72.0**	71.2	66.2	**68.6**	50.9	35.7	42.0	38.3	60.9	*47.0*
f-CLSWGAN+Proposed	56.4	83.2	67.2	52.1	55.8	53.9	53.3	35.0	42.3	37.1	57.7	45.2
LsrGAN+Proposed	60.4	79.1	***68.5***	51.7	57.4	54.3	49.0	37.2	42.3	35.1	59.7	44.2
TF-VAEGAN+Proposed	62.4	73.8	67.6	53.0	64.8	***58.3***	47.3	41.0	*43.9*	38.2	57.5	***45.9***

methods: DLFZRL+softmax (Tong et al., 2019) which also uses a deterministic model-DLFZRL combined with f-CLSWGAN, and RFF-GZSL (Han et al., 2020) which uses a mutual information-based approach to learn the redundancy free information to facilitate f-CLSWGAN, our proposed method gains superior performance. Furthermore, the last two rows show that our proposed method is also broadly applicable to improve the performance of an advanced generative model. Moreover, our results are also comparable with state-of-the-art generative models; unlike some generative models (Huang et al., 2019; Shen et al., 2020; Chou et al., 2021), which may gain superior results on one or two datasets but inferior results on other datasets, our proposed method obtains favorable results on all datasets.

6. Conclusion

This paper addresses information-theoretic loss functions to quantify knowledge transfer and semantic relations for GZSL/ZSL. Using the proposed probability vector representation based on the prototypical model, the proposed loss can be effectively evaluated with simple closed forms. Experiments show that our approach yields state-of-the-art performance for the deterministic approach for both the GZSL and the conventional ZSL tasks. Moreover, by incorporating the generated data from known GAN models, the proposed method also gains favorable performance. One limitation of this work is that we need pretty much extra cross-validation to select the hyperparameters; Another limitation is that the loss functions have a certain correlation, so further study is needed to simply the loss functions while keeping similar performance.

References

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, pages 819–826, 2013.

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

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. In *IEEE TPAMI*, pages 1425–1438, 2016.

Y. Annadani and S. Biswas. Preserving semantic relations for zero-shot learning. In *CVPR*, pages 7603–7612, 2018.

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

Jianmin Bao, D. Chen, Fang Wen, H. Li, and G. Hua. Cvae-gan: Fine-grained image generation through asymmetric training. *ICCV*, pages 2764–2773, 2017.

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

Dubing Chen, Yuming Shen, Haofeng Zhang, and Philip H.S. Torr. Zero-shot logit adjustment. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 813–819, 2022.

Yu-Ying Chou, Hsuan-Tien Lin, and Tyng-Luh Liu. Adaptive and generative zero-shot learning. In *ICLR*, 2021.

Jia Deng, W. Dong, R. Socher, L. Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

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

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. Devise:a deep visual-semantic embedding model. In *NeurIPS*, pages 2121–2129, 2013.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014.

Zongyan Han, Zhenyong Fu, and Jian Yang. Learning the redundancy-free features for generalized zero-shot object recognition. *CVPR*, pages 12862–12871, 2020.

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

He Huang, Changhu Wang, Philip S. Yu, and Chang-Dong Wang. Generative dual adversarial network for generalized zero-shot learning. *CVPR*, pages 801–810, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36:453–465, 2014.

A. Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *ACL*, 2015.

J. Li, M. Jin, K. Lu, Z. Ding, L. Zhu, and Z. Huang. Leveraging the invariant side of generative zero-shot learning. In *CVPR*, 2019a.

Jin Li, Xuguang Lan, Yang Liu, Le Wang, and Nanning Zheng. Compressing unknown images with product quantizer for efficient zero-shot classification. *CVPR*, pages 5458–5467, 2019b.

Xiangyu Li, Zhe Xu, Kun-Juan Wei, and Cheng Deng. Generalized zero-shot learning via disentangled representation. In *AAAI Conference on Artificial Intelligence*, 2021.

S. Liu, M. Long, J. Wang, and M. I. Jordan. Generalized zero-shot learning with deep calibration network. In *NeurIPS*, 2018.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

Norouzi Mohammad, Mikolov Tomas, Bengio Samy, SingerYoram, Shlens Jonathon, Frome Andrea, Corrado Greg, and Dean Jeffrey. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

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.

B. Romera Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015.

Adam Paszke, S. Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zach DeVito, Zeming Lin, Alban Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

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

S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Parallel distributed processing: Explorations in themicrostructure of cognition. *MIT Press*, 1, 1986.

E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata. Generalized zero and few-shot learning via aligned variational autoencoders. In *CVPR*, 2019.

Yuming Shen, Jie Qin, Lei Huang, Li Liu, Fan Zhu, and Ling Shao. Invertible zero-shot recognition flows. In *ECCV*, pages 614–631. Springer, 2020.

Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, MasashiShimbo, and Yuji Matsumoto. Ridge regression, hubness,and zero-shot learning. In *ECML-PKDD*, pages 135–151, 2015.

J. Snell, Kevin Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

Kihyuk Sohn, H. Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NeurIPS*, 2015.

Bin Tong, Chao Wang, Martin Klinkigt, Yoshiyuki Kobayashi, and Yuuichi Nonaka. Hierarchical disentanglement of discriminative latent features for zero-shot learning. *CVPR*, pages 11459–11468, 2019.

Maunil R Vyas, Hemanth Venkateswara, and Sethuraman Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *ECCV*, pages 70–86. Springer, 2020.

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

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

Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learninga comprehensive evaluation of the good, the bad and the ugly. In *IEEE TPAMI*, 2018a.

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

Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. F-vaegan-d2: A feature generating framework for any-shot learning. *CVPR*, pages 10267–10276, 2019.

Tingting Xu, Ye Zhao, and Xueliang Liu. Dual generative network with discriminative information for generalized zero-shot learning. *Complex.*, 2021:6656797:1–6656797:11, 2021.

H. Zhang and P. Koniusz. Zero-shot kernel learning. In *CVPR*, 2018.

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

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