mathematics

Article

Variational Disentangle Zero-Shot Learning

Jie Su [1], Jinhao Wan [2,*], Taotao Li [2], Xiong Li [2] and Yuheng Ye [2]

[1] School of Computing, Newcastle University, Newcastle upon Tyne NE4 5TG, UK; j.su4@newcastle.ac.uk
[2] ISPNU Lab, Zhejiang University of Technology, Hangzhou 310023, China; 2111903074@zjut.edu.cn (T.L.); 2112003075@zjut.edu.cn (X.L.); 211122120010@zjut.edu.cn (Y.Y.)
* Correspondence: 111122030015@zjut.edu.cn

Abstract: Existing zero-shot learning (ZSL) methods typically focus on mapping from the feature space (e.g., visual space) to class-level attributes, often leading to a non-injective projection. Such a mapping may cause a significant loss of instance-level information. While an ideal projection to instance-level attributes would be desirable, it can also be prohibitively expensive and thus impractical in many scenarios. In this work, we propose a variational disentangle zero-shot learning (VDZSL) framework that addresses this problem by constructing variational instance-specific attributes from a class-specific semantic latent distribution. Specifically, our approach disentangles each instance into class-specific attributes and the corresponding variant features. Unlike transductive ZSL, which assumes that unseen classes' attributions are known beforehand, our VDZSL method does not rely on this strong assumption, making it more applicable in real-world scenarios. Extensive experiments conducted on three popular ZSL benchmark datasets (i.e., AwA2, CUB, and FLO) validate the effectiveness of our approach. In the conventional ZSL setting, our method demonstrates an improvement of 12∼15% relative to the advanced approaches and achieves a classification accuracy of 70% on the AwA2 dataset. Furthermore, under the more challenging generalized ZSL setting, our approach can gain an improvement of 5∼15% compared with the advanced methods.

Keywords: zero-shot learning; computer science; pattern recognition; deep learning

MSC: 68T07; 68T10; 68T45

check for updates

Citation: Su, J.; Wan, J.; Li, T., Li, X.; Ye, Y. Variational Disentangle Zero-Shot Learning. *Mathematics* **2023**, *11*, 3578. https://doi.org/10.3390/math11163578

Academic Editors: Xiangtao Zheng, Jinchang Ren and Ling Wang

Received: 11 July 2023
Revised: 13 August 2023
Accepted: 16 August 2023
Published: 18 August 2023

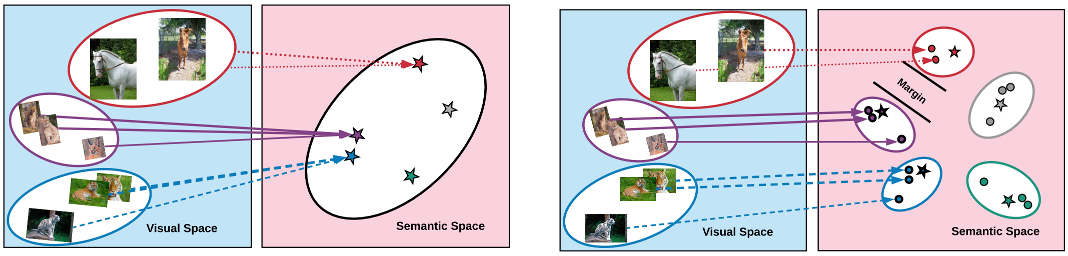
1. Introduction

Remarkable success in object classification has been achieved in recent years with the advances in deep convolutional neural networks. However, a common limitation in most state-of-the-art models is that they are generally trained on data with known labels and do not generalize to unseen classes To address this issue, zero-shot learning (ZSL) [1] was proposed, which assumes unseen classes share some auxiliary modalities (with seen classes), through which a generalized model can be trained to recognize both seen and unseen classes. Extensive studies have been conducted in various computer vision tasks such as super-resolution, object detection, and style transfer.

In ZSL, one of the most popular forms of auxiliary modalities is class-level attributes, usually provided by domain experts, crowd-sourcing annotations, or word embedding. With the shared attributes, the knowledge learned from seen classes can be transferred to unseen classes. For example, the model learns the visual features corresponding to the attribute 'stripes' from the seen class zebra, and these features should also be able to be used to predict the 'strips' from the unseen class zebra crossing in the city.

Conventional zero-shot learning (ZSL) models can be categorized into three primary groups: linear mapping models, non-linear mapping models, and semantic and probabilistic approaches. In the realm of linear mapping models, the deep visual-semantic embedding (DeViSE) model [2] leverages a pretrained neural language model coupled with a deep neural network to learn the parameters for the linear transformation layer. Subsequently, the attribute label embedding (ALE) model [3] employs a WSABIE ranking objective [4] to learn a linear compatibility function, thereby prioritizing the top of the ranking list. Later, the structured joint embedding (SJE) framework [5] integrated an SVM [6] to assess its importance, improving upon previous work. The embarrassingly simple zero-shot learning (ESZSL) [7] model unifies a linear mapping with a straightforward empirical loss, along with regularization terms that penalize both the projection of feature vectors from a Euclidean space into the attribute space and vice versa. Nonlinear methods such as the latent embeddings (LatEm) model [8] extend the DeViSE model by incorporating a piecewise linear compatibility function, thus amalgamating multiple mappings learned through a pairwise ranking loss. Later, the semantic autoencoder (SAE) [9] introduced an encoder-decoder network to prevent information loss and improve feature learning. Semantic and probabilistic approaches to zero-shot learning include direct and indirect attribute prediction (DAP and IAP, respectively) models [10], which learn a probabilistic attribute classifier and predict the label by combining classifier scores. Those approaches have yielded encouraging results during the past few decades [11].

However, there are some practical issues that have been neglected for many years; in most embedding models, the mapping from a visual to a semantic space is generally non-injective. The definition of a non-injective problem in the ZSL task is as follows: Multiple visual instances or images are mapped to a single vector of class-level attributes, creating a situation where the inverse function does not exist. Under such conditions, the mapped space lacks the ability to uniquely trace back from the class-level attribute vector to a specific visual instance. As depicted in Figure 1a, two horse images (in white and brown) are mapped to the same class attribute (indicated by the red star) in conventional ZSL models. This implies that the intra-class variability, such as distinguishing details between the two horses, is neglected during the mapping process, resulting in an instance-level loss of information. To this end, we propose a variational disentangled zero-shot learning (VDZSL) framework which aims to learn an instance-level mapping (i.e., that shown in Figure 1b) across a visual-semantic space and preserve large inter-class margins (for better discrimination) simultaneously.



(**a**) Conventional embedding model for ZSL

(**b**) Proposed VDZSL model

Figure 1. A comparison between conventional ZSL and proposed VDZSL. (**a**) Conventional embedding model for ZSL. The mapping from a visual to a semantic space is non-injective (e.g., two horse images mapping to a single attribute in a semantic space). (**b**) Proposed VDZSL model. Images are projected to individual points, which can be modeled as a Gaussian distribution.

Our contribution can be summarized as follows:

- We identify the non-injective problem that results from a lack of instance-level attributes for ZSL classification tasks.
- We introduce a novel VDZSL method that leverages variation inference to disentangle instance-specific attributes from shared class-specific information.
- Extensive experiments are conducted on three benchmark datasets, and our model generally outperforms other state-of-the-art methods.

2. Related Work

The main difference between ZSL and most other machine learning methods is the prerequisite of human-understandable prior knowledge of the seen and unseen classes. The existing ZSL literature covers broad research topics, such as multimodal human priors [12,13], visual-semantic embedding [14], and seen-unseen domain adaptation [15]. The ZSL problem has many distinctive variations as well. Papers with different research purposes cannot simply be put together and compared. This section aims to scope out these distinctions so as to highlight our focus and contributions.

2.1. Zero-Shot Learning Variations

Traditional supervised learning models fail to generalize to new classes due to a lack of training examples for unseen categories. Zero-shot learning (ZSL) [16] utilizes a semantic modality to connect the visual and label spaces. The associations between semantic modality and labels come from human-understandable prior knowledge, and they require no training images from unseen categories. The challenge is how to achieve a consistent mapping from an image to semantic modality. Note that this conventional ZSL method was considered as an ill-posed problem in [17] because no unseen category information should be assumed available before the test. Moreover, images from arbitrary new classes may not follow the learned mapping from the seen category. If the unseen category's representation is not correlated with the seen category, then the learned visual-semantic mapping cannot generalize. Transductive zero-shot learning [18,19] is a compromise to the 'unseen' condition. Its setting assumes all the information of unseen classes is available except their labels. The test is then simplified into a dynamic domain alignment task between heterogeneous modalities. Such a task is sometimes referred to as unsupervised ZSL [20]. Because the entire visual distribution has been exposed during training, the transductive setting is significantly favorable in performance compared with conventional ZSL. However, collecting all unlabeled test images during the training stage can hardly be achievable in many realistic applications. A recent setting [14,21–25] was widely adopted by powerful generative models applied on a predefined semantic modality to synthesize training visual examples for both the seen and unseen classes. Different from the above settings, generalized zero-shot learning (GZSL) [17,26] considers a larger output space. More precisely, during inference, the output label space includes both the seen and unseen classes with increasing classification challenges. Our work will be evaluated for both the ZSL and GZSL settings.

2.2. Variation Autoencoder

The variational autoencoder (VAE) is a deep generative model which aims to learn complex density models from data via latent variables. Given a nonlinear generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ with input $\mathbf{x} \in \mathbb{R}^D$ associated with a latent variable $\mathbf{z} \in \mathbb{R}^L$ coming from some prior distribution $p(\mathbf{z})$, the VAE aims to use an encoding model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to approximate the posterior distribution of the latent variable (i.e., $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$). The learning process is achieved by maximizing the following variational lower bound:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(z|x)}[logp_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - \mathbf{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{1}$$

Typically, the posterior distribution of the latent variable $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ is defined as an isotropic normal distribution with its mean and standard deviation of the output of the deep neural network. After the learning process, a probabilistic encoded latent code $\mathbf{z}$ for the input can be generated by the encoding model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. We leverage the flexibility of the VAE to design a structured 'Siamese' encoding network that allows us to explore the class-specific information and class variation information for ZSL tasks.

Generative Model for ZSL

The early work on ZSL methods can mainly be divided into compatibility functions [2,3,7,27,28] and embedding approaches [29–33]. Compared with generative models, their training is purely based on conventional ZSL rather than semantic transductive settings. The learned classifier or embedding is directly applied to unseen images without further fine-tuning. In contrast, despite a massive performance gain, recent proposed generative approaches [14,21–25] rely on semantic transductive information so as to synthesize the visual features of unseen classes. However, knowing all the semantic information of the test classes in the semantic transductive setting can be difficult to achieve in realistic settings. Our method sticks to ZSL with a new usage that aims to model the noisy intra-class variability as a generative procedure so that we can distill more discriminative class representations.

3. Method

Our model, shown via the pictorial illustration in Figure 2, is based on a parallel autoencoder (AE) and variational autoencoder (VAE) architecture. The framework consists of two learning processes: variational disentangle learning and zero-shot learning. The variational disentangle learning process (i.e., the full network in Figure 2) can take advantage of a disentangle net and margin constraint to reduce the effect brought by redundant features and build a sparse latent space. The latter can leverage the extracted class-specific representation for ZSL classification tasks.

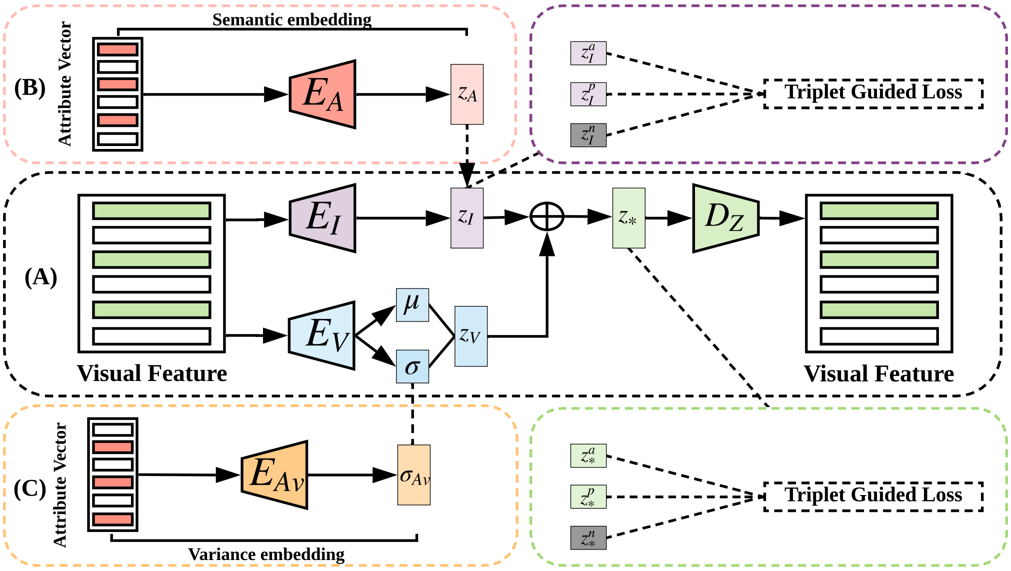


Figure 2. The modules of our VDZSL framework consist of three parts: (**A**) Variational disentangle network; (**B**) Semantic encoding module; (**C**) Variance encoding module. The Variational disentangle network consists of a class-specific feature encoder E_A and a class-variant feature encoder E_V. They take the visual features as input and extract class-specific and variant features, respectively. These features are supervised by the attribute and attribute-variance features encoded by the attribute feature encoder E_A and attribute variance feature encoder E_{Av}. Finally, the feature reconstructor decoder combines the class-specific feature $\mathbf{z}_I^i$ and class-variant feature $\mathbf{z}_V^i$ to reconstruct the original input features.

3.1. Problem Definition

Let $\boldsymbol{\mathcal{S}} = \{s_1, \ldots, s_p\}$ denote the set of seen classes and $\boldsymbol{\mathcal{U}} = \{u_1, \ldots, u_q\}$ denote that of unseen classes, where p and q are the total numbers of seen and unseen classes, respectively. $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{U}}$ are disjoint sets (i.e., $\boldsymbol{\mathcal{S}} \cap \boldsymbol{\mathcal{U}} = \varnothing$). Similarly, we define $\boldsymbol{\mathcal{A}}_s = \{\mathbf{a}_1^s, \ldots, \mathbf{a}_p^s\} \in \mathbb{R}^{p \times k}$ and $\boldsymbol{\mathcal{A}}_u = \{\mathbf{a}_1^u, \ldots, \mathbf{a}_q^u\} \in \mathbb{R}^{q \times k}$, which are the seen and unseen class semantic representations, respectively, where k is the dimension of the attribute vector. Given a set of labeled training data $\boldsymbol{\mathcal{D}}_s = \{\mathbf{x}_i^s, \mathbf{a}_i^s, y_i^s\}$, where $i \in \{1, \ldots, n\}$ and $\mathbf{x}_i^s \in \mathbb{R}^{1\times d}$ represents the d-dimensional visual feature vector of the seen classes, $y_i^s \in \{1, \ldots, p\}$ is the class label of $\mathbf{x}_i^s$, $\mathbf{a}_i^s$ is the semantic representation vector of $\mathbf{x}_i^s$, and n denotes the total number of instances in the seen classes. Similarly, the data of the unseen classes to be tested are defined as $\boldsymbol{\mathcal{D}}_u = \{\mathbf{x}_i^u, \mathbf{a}_i^u, y_i^u\}$, where $i \in \{1, \ldots, m\}$ and m denotes the total number of unseen instances. The goal of zero-shot learning is to leverage the seen training data $\mathcal{D}_s$ to learn a classifier $f : \boldsymbol{\mathcal{X}}_u \to \mathcal{U}$, where $\boldsymbol{\mathcal{X}}_u = \{\mathbf{x}_1^u, \ldots, \mathbf{x}_m^u\}$.

3.2. Variational Disentangle Network

Due to the diversity of images, different images from the same classes may suffer from noise factors (e.g., the background), which might lead to misclassification. To reduce the effect caused by those noise factors, a potential solution is to perform feature extraction such that the discriminant and noise features can be separated. Our VDZSL model presents a disentangle projection scheme to extract the class-specific and class-variant latent codes from the visual feature space. As shown in Figure 2, E_I and E_V denote the class-specific and class-variant feature encoding network, respectively. To be specific, giventhe feature data $\boldsymbol{\mathcal{X}}_s = \{\mathbf{x}_1^s, \cdots, \mathbf{x}_n^s\}$, for each $\mathbf{x}_i^s$, we have the class-specific representation $\mathbf{z}_I^i = E_I(\mathbf{x}_i^s)$ and variant representation $\mathbf{z}_V^i = E_V(\mathbf{x}_i^s)$. The downstream network E_V (i.e., block A in Figure 2) uses a basic variation autoencoder network, and the upstream network E_I uses a basic autoencoder network.

Motivated by the VAE, we assume the data are generated by an unobserved continuous random variable $\mathbf{z}_*$ in the embedding space. The generation process consists of three parts: (1) a variant feature $\mathbf{z}_V^i$ is generated from some prior distribution $p'_{\boldsymbol{\theta}}(\mathbf{z}_V)$; (2) $\mathbf{z}_I^i$ is a class-specific latent code, and $\mathbf{z}_*^i$ is defined by the aggregation of $\mathbf{z}_I^i$ and $\mathbf{z}_V^i$ such that $\mathbf{z}_*^i = \mathbf{z}_I^i \otimes \mathbf{z}_V^i$; and (3) finally, the decoder net D_Z aims to reconstruct the region features from some conditional distribution $p'_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_*)$, i.e., $\mathbf{x}_i^{s'} = D_Z(\mathbf{z}_*^i)$. Following the lower boundary on the VAE, the loss for modeling the class variant code could be written as

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) \approx -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}_V|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}_V)) + \frac{1}{N}\sum_{i=1}^{N} logp_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_V^i). \tag{2}$$

Moreover, different classes may suffer from different variances, and thus we model the prior distribution of $\mathbf{z}_V$ to be a center isotropic multivariate Gaussian which is conditioned on the class-specific attribute (i.e., $p_{\boldsymbol{\theta}}(\mathbf{z}_V) \sim \mathcal{N}(0, \sigma_{Av})$). Here, the class-conditioned variance vector is encoded by the variance embedding net (i.e., block C in Figure 2), such that $\boldsymbol{\sigma}_{Av}^j = D_{Av}(\mathbf{a}_j^s)$ and $\mathbf{a}_j^s \in \mathcal{A}_s$. Then, Equation (2) can be rewritten as

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) \approx -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}_V|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}_V|\sigma_{Av})) + \frac{1}{N}\sum_{i=1}^{N} logp_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_V^i). \tag{3}$$

For the approximation posterior in Equation (3), we model a multivariate Gaussian with diagonal covariance (i.e., $q_{\boldsymbol{\phi}}(\mathbf{z}_V|\mathbf{x}_i) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2\mathbf{I})$). The KL divergence between $q_{\boldsymbol{\phi}}(\mathbf{z}_V|\mathbf{x})$ and $p_{\boldsymbol{\theta}}(\mathbf{z}_V|\sigma_{Av})$ could be written as

$$\begin{aligned} D_{KL} &= \int q_{\boldsymbol{\phi}}(\mathbf{z}_V|\mathbf{x}) \log q_{\boldsymbol{\phi}}(\mathbf{z}_V|\mathbf{x})dx - \int q_{\boldsymbol{\phi}}(\mathbf{z}_V|\mathbf{x}) \log p_{\boldsymbol{\theta}}(\mathbf{z}_V|\sigma_{Av})dx \\ &= \left\{-\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\left(\log\sigma_i^2 + 1\right)\right\} \\ &\quad - \left\{-\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log\sigma_{Av}^2 - \frac{1}{2}\sum_{i=1}^{N}\left[\frac{\sigma_i^2}{\sigma_{Av}^2} + \frac{\boldsymbol{\mu}_i^2}{\sigma_{Av}^2}\right]\right\} \\ &= -\frac{1}{2}\sum_{i=1}^{N}\left[\log\frac{\sigma_i^2}{\sigma_{Av}^2} - \frac{\sigma_i^2}{\sigma_{Av}^2} - \frac{\boldsymbol{\mu}_i^2}{\sigma_{Av}^2} + 1\right] \end{aligned} \tag{4}$$

Benefiting from the elegant reparameterization trick from the VAE, the variational disentangle network's loss function could be written as

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}) \approx -\frac{1}{2N}\sum_{i=1}^{N}(1+log(\frac{\sigma_i^2}{\sigma_{Av}^2})-(\frac{\sigma_i^2+\mu_i^2}{\sigma_{Av}^2})))\\ +\frac{1}{N}\sum_{i=1}^{N}logp_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_*^i). \tag{5}$$

More specifically, $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x})$ (i.e., the objective function of the disentangle network) can be rewritten as

$$\mathcal{L}_{disentangle} = -\frac{1}{2N}\sum_{i=1}^{N}(1+log(\frac{\sigma_i^2}{\sigma_{Av}^2})-(\frac{\sigma_i^2+\mu_i^2}{\sigma_{Av}^2})))\\ +\frac{1}{N}\sum_{i=1}^{N}||\mathbf{x}_i - D_z(E_I(\mathbf{x}_i)+E_V(\mathbf{x}_i))||. \tag{6}$$

3.3. Margin Regularizer

The objective of the variational disentangle network naturally encourages the network to extract the class-specific latent code and class variance latent code. In classification tasks, for large inter-class separation in the latent space, we further use a margin regularizer (i.e., triplet guided loss [34]):

$$\mathcal{L}_{margin} = \varphi\frac{1}{N}\sum_{i=1}^{N}max(0,\alpha+d(\mathbf{z}_I^a,\mathbf{z}_I^p)-d(\mathbf{z}_I^a,\mathbf{z}_I^n))\\ +\varphi'\frac{1}{N}\sum_{i=1}^{N}max(0,\alpha+d(\mathbf{z}_*^a,\mathbf{z}_*^p)-d(\mathbf{z}_*^a,\mathbf{z}_*^n)) \tag{7}$$

where $d(\cdot)$ represents the distance function.

The first term of Equation (7) indicates the margin regularization in the class-specific latent code which may push the negative latent code (i.e., $\mathbf{z}_I^n$, where $y^n \neq y^p$) far away from the anchor and positive latent code (i.e., $\mathbf{z}_I^a,\mathbf{z}_I^p$, where $y^a = y^p$) in a margin α. Similarly, the second term aims to push the embedded latent code (i.e., $\mathbf{z}_*$) far away from other classes' latent codes. It should be noted that not all of the triplet samples can contribute loss to the training. In this case, during the training process, only validate-hard samples are used.

Note that in Equation (7), φ and φ' are two hyperparameters that can control the constraint effect on the latent code $\mathbf{z}_I$ and $\mathbf{z}_*$. As the class-specific latent code should become more representative (used for classification), we set larger values for large margins. However, the combining latent code $\mathbf{z}_*$ should maintain the original information for the reconstruction process (with variant information added). Thus, we gave it a relatively small value to relax the constraint. In this work, we empirically set $\varphi = 0.9$ and $\varphi' = 0.1$.

3.4. Zero-Shot Learning

For ZSL classification, we link the visual and semantic space by mapping the class-specific attribute to the center of a class-specific latent space. It should be noted that this is a 'one-to-one' mapping, which is different from the other ZSL methods. This mapping can be performed by minimizing the distance between the encoded attribute code and the corresponding class center (i.e., the mean vector of the class-specific latent code):

$$\mathcal{L}_{cls} = \frac{1}{p}\sum_{j=1}^{p}||\mathbf{z}_{As}^j - \frac{1}{M_j}\sum_{i=1}^{M_j}\mathbf{z}_I^{i,j}||_F, \tag{8}$$

where $||\cdot||_F$ indicates the Frobenius norm and M_j is the sample size in class j, $\mathbf{z}^j_{As} = E_A(\mathbf{a}^s_j)$ where $\mathbf{a}^s_j \in \boldsymbol{\mathcal{A}}_s$.

3.5. Final Training Objective

The variational disentangle learning and zero-shot classification are trained simultaneously. The final objective function is a combination of Equations (6)–(8) which can be written as

$$\mathcal{L}_{total} = \lambda\mathcal{L}_{disentangle} + \beta\mathcal{L}_{margin} + \sigma\mathcal{L}_{cls}, \tag{9}$$

where λ, β, and σ are the regularization coefficients of different loss functions.

Once the parameters of the encoding network E_A are learned, the unseen class's semantic feature can be obtained by encoding their corresponding class attribute vector (i.e., $\mathbf{z}^u_a = E_A(\mathbf{a}^u)$). Finally, the ZSL classification task can be solved with a nearest neighbor search between $\mathbf{z}^u_a$ and the unseen class-specific latent code (i.e., $\mathbf{z}^u_I = E_I(\mathbf{x}^u)$):

$$\hat{y} = arg\min_{y\in\mathcal{Y}_u} d(\mathbf{z}^u_I, \mathbf{z}^u_{aj}), \mathbf{z}^u_{aj} \in \{\mathbf{z}^u_1, \cdots, \mathbf{z}^u_q\} \tag{10}$$

where $d(\cdot)$ represents the distance function, and the cosine distance is used here.

4. Experiment Set-Up

4.1. Datasets and Settings

We evaluated our method on three public datasets. AWA2 [10] is a coarse-grained dataset with 37,322 images and 50 classes. CUB-200-2011 Birds (CUB) [35] is a fine-grained and medium-scale dataset with respect to both the number of images and number of classes (i.e., 11,788 images from 200 different types of birds annotated with 312 attributes). Oxford-Flowers (FLO) [36] is a small-scale dataset with respect to both the number of images and number of classes (i.e., FLO contains 6786 images coming from 102 types of flower annotated with 1024 attributes). For the seen and unseen class split, we followed the protocol used in [17].

4.1.1. Input Space

The input space of ZSL includes two parts: a semantic space and a visual space. For the semantic space, attributes and word2vec embeddings are two popular input sources, where the former is used to form the semantic space for the visual content while the latter is used as the semantic representation for a large-scale dataset (e.g., ImageNet). For the visual space, the visual features extracted by the CNN models, which are pretrained on ImageNet in ILSVRC 2012 [37], are widely used. In our experiments, we used pretrained ResNet-101 [38] features and attributes as our input sources.

4.1.2. Evaluation Metrics

For the ZSL setting, only accuracy was used. For the generalized ZSL setting, three metrics were used: (1) acc_{tr},which is the average per-class classification accuracy on the seen class test data using a classifier trained for all classes, (2) acc_{ts}, which is the average per-class classification accuracy on the unseen class test data using a classifier trained for all classes, and (3) H, which is the harmonic mean of acc_{tr} and acc_{ts} (i.e., $H = \frac{2\times acc_{tr}\times acc_{ts}}{acc_{tr}+acc_{ts}}$).

There were two phases in the training procedure. In the first phase, we cut off the backpropagation of the gradient from the reconstruction term, which aims to ensure the class-specific latent space's construction. For the latent space dimension, we followed [39] and fixed the embedding size to 512 in all the experiments. For the regularization coefficients, we first set $\lambda = 1, \beta = 1$, and $\sigma = 1$ and tuned them for optimal results. For the network in our experiment, we used linear layers to construct the encoder and decoder, where tanh was used as the activation function.

5. Discussion

5.1. Quantitative Results Discussion

In this section, we present the results for both the ZSL and GZSL settings. The performance of our method, along with other baseline comparisons on the CUB, AWA2, and FLO datasets, is summarized in Table 1. In the ZSL setting, we can see the superior performance improvements in the AwA2 and FLO datasets brought about by the proposed VDZSL framework. Benefiting from the designed class-specific feature extraction and the incorporation of triplet margin loss, the clustering of different classes became more discriminative. This led to a reduction in the intra-class distances, thereby yielding improved classification results. However, we observed that the results from the CUB dataset were on par with previous approaches. One major reason for this similarity in performance can be attributed to the large class number of the CUB dataset (i.e., 200 classes), making fine-grained classification more challenging. The differences across various bird classes may be subtle, rendering the extraction of discriminative features more difficult and thus posing challenges. In the GZSL setting, our method generally achieved higher values for *ts* and *H* on the three datasets, with the exception of *H* in the CUB dataset. We observed that most algorithms, including ours, tended to have higher *tr* values compared with *ts*, which is attributed to overfitting the data with seen classes. Nevertheless, our method still maintained reasonable performance with the GZSL setting when compared with other approaches.

Table 1. Results of our VDZSL and other baselines in ZSL and GZSL settings.

Methods	**ZSL**			**Generalized ZSL**								
	AwA2	**CUB**	**FLO**	**AwA2**			**CUB**			**FLO**		
				ts	**tr**	**H**	**ts**	**tr**	**H**	**ts**	**tr**	**H**
DAP	46.1	40.0	-	0.0	84.7	0.0	1.7	67.9	3.3	-	-	-
IAP	35.9	24.0	-	0.9	87.6	1.8	0.2	72.8	0.4	-	-	-
LATEM	55.8	49.3	40.4	11.5	77.3	20.0	15.2	57.3	24.0	6.6	47.6	21.5
ALE	62.5	54.9	48.5	14.0	81.8	23.9	23.7	62.8	34.4	13.3	61.6	21.9
DEVISE	59.7	52.0	45.9	17.1	74.7	27.8	23.8	53.0	32.8	9,9	44.2	16.2
SJE	61.9	53.9	53.4	8.0	73.9	14.4	23.5	59.2	33.6	13.9	47.6	21.5
ESZSL	58.6	53.9	51.0	5.9	77.8	11.0	12.6	63.8	21.0	11.4	56.8	19.0
SAE	58.1	42.0	45.6	1.1	82.8	2.2	17.4	50.7	25.9	-	-	-
VDZSL (Ours)	**70.0**	53.0	**60.0**	**20.0**	85.0	**32.3**	**24.8**	48.5	32.9	**23.5**	**79.7**	**36.4**

5.2. Latent Space Visualization

To better understand the learned features, T-SNE visualization was used on the original CNN feature $\mathbf{x}^u$, reconstructed feature $\mathbf{x}^{u\prime}$, class-specific latent code $\mathbf{z}_I$, and combined feature latent code $\mathbf{z}_*$ as shown in Figure 3. As can be seen in Figure 3, the reconstructed CNN feature $\mathbf{x}^{u\prime}$ still kept a similar shape with the original features $\mathbf{x}^u$, and the class-specific latent code $\mathbf{z}_I$ led to an elegant separation while still keeping the distribution shape of the original feature space. We also observed that the 'seal' and 'walrus' overlapped in both the reconstructed feature space and class-specific feature space. A possible explanation could be the negative pulling effect during the triplet training caused by the similar input feature from these two classes.

We also calculated the inter-class distances in different latent codes (i.e., $\mathbf{z}_I$, $\mathbf{z}_*$) as shown in Figure 4, where a darker color represents a larger distance. From Figure 4, we can see that the classes tended to be far away in the class-specific latent space (corresponding to $\mathbf{z}_I$) than the combined latent space (corresponding to z_*), indicating the high-level discriminant capability of $\mathbf{z}_I$. On the other hand, z_* aggregated both the discriminant class-specific feature $\mathbf{z}_I$ and class-variant features (i.e., noises), yielding a high level of class overlapping in its feature space (i.e., Figure 3d) with shorter inter-class distances (as shown in Figure 4).

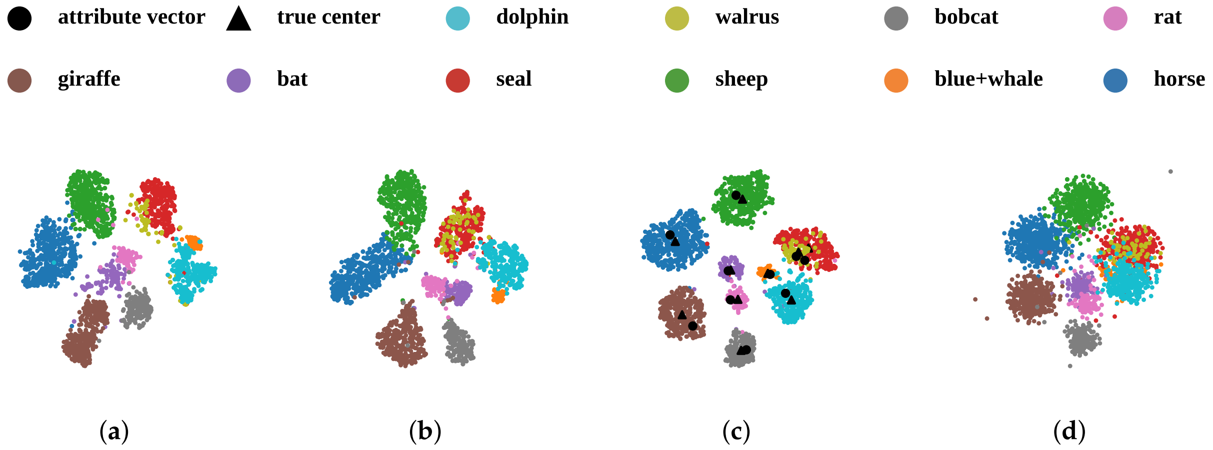


Figure 3. t-SNE visualization for AwA2 dataset. (**a**) Original CNN feature $\mathbf{x}^u$. (**b**) Reconstructed features $\mathbf{x}^{u\prime}$. (**c**) Class-specific latent code $\mathbf{z}_I$. (**d**) Combined latent code $\mathbf{z}_*$.

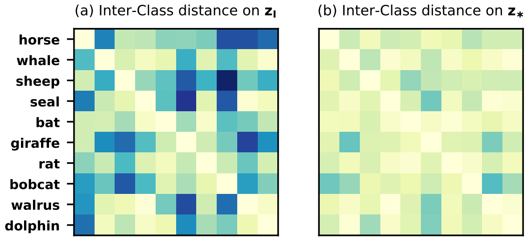


Figure 4. Inter-class distance comparison on AwA2 datasets under ZSL setting. (**a**) Inter-class distance for class-specific latent code $\mathbf{z}_I$. (**b**) Inter-class distance for combined latent code $\mathbf{z}_*$. Dark color indicates large distance, and vice versa.

5.3. Margin Analysis

We also evaluated the effectiveness of the triplet margin setting, and the accuracy distribution with respect to the margin size is shown in Figure 5. It is interesting to see that the performance deteriorated with large margins (e.g., >1), which might have been caused by a large class number (e.g., 50, 102, or 200 classes for AwA2, FLO, CUB, respectively). Although a large margin constraint was normally used for better class separation, with a large class number, the features may have been highly overlapped in the feature space, making the trained model less effective.

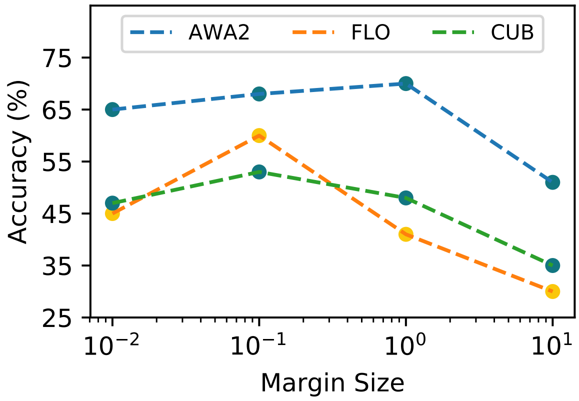


Figure 5. Performance with respect to margin size.

5.4. Limitations and Future Work

5.4.1. Limitation

Although the proposed VDZSL framework can significantly enhance the performance of zero-shot learning tasks by addressing the non-injective projection problem, it suffers from feature constraint. Specifically, the performance of zero-shot recognition is limited by the features that are either pretrained or extracted from the large-scale dataset. This means that even if the design of the learning framework is excellent, it may still fail to deliver decent performance for zero-shot object classification. This limitation stems from the coarse-grained features extracted during the initial dataset creation stage.

Another major challenge is the distribution distance measurement. Using KL divergence to measure the distribution distance could lead to challenges in the gradient calculation. Specifically, when two distributions are either too far away or too close, the KL divergence measurement may fail. This failure occurs because KL divergence is not symmetric, meaning that swapping the two distributions can result in a different value. Additionally, if one distribution has support while the other does not (i.e., it assigns a probability of zero to an event that has a positive probability under the other distribution), then the KL divergence becomes infinite. This behavior can cause issues in mathematical optimization tasks, such as gradient descent, where the negative or undefined gradients could disrupt the learning process. These aspects of KL divergence could be further investigated to understand their impact on specific applications.

5.4.2. Future Work

According to the aforementioned limitations, future work in the field of zero-shot learning may be directed toward the following aspects:

- Designing end-to-end training strategies for zero-shot learning (ZSL) recognition allows for the avoidance of pretrained features. The use of task-specific features can enhance recognition performance, leading to more accurate results.
- Investigate more advanced distance measurements (e.g., the Wasserstein distance in the earth mover's distance group) and their effects on the zero-shot learning task.
- By connecting the advanced generative model [40] with zero-shot learning, we can leverage its capabilities. Specifically, by generating images conditioned on attributes, we can produce a larger-scale dataset suitable for the zero-shot learning task.

5.4.3. Connecting to Real-World Applications

While ZSL has shown success in object classification tasks, its applications reach well beyond this realm, influencing various real-world domains. In healthcare, ZSL exhibits potential for diagnosing rare diseases and pinpointing anomalies within medical imaging. Within the context of environmental conservation and wildlife monitoring, ZSL can be harnessed to identify previously unobserved species or ecological phenomena. In the field of industrial automation, it can enable robots to recognize a wide variety of products, and in the financial sector, ZSL may be used to detect emerging fraud patterns. This provides a flexible and innovative approach to identifying suspicious activities without relying on previous examples. The extensive applications of ZSL underscore its adaptability and potential for enhancing technological capabilities and addressing complex challenges across an array of diverse domains.

6. Conclusions

We presented the variational disentangle zero-shot learning (VDZSL) framework for predicting unseen classes under both ZSL and GZSL settings. Unlike existing ZSL methods, our approach is designed to disentangle class-specific features from variance noise, rendering the classification process less sensitive to nuisance factors. By separating class-specific information from the original visual features, our framework substantially reduces the influence of variant noises. Moreover, we model the semantic space as a distribution to preserve variant information, aiding the reconstruction of visual features.

Experimental results on three public datasets have attested to the effectiveness of our VDZSL method in both the ZSL and GZSL settings, showing notable advantages over other algorithms. Finally, although we employed an isotropic Gaussian distribution to model the intra-class variant latent code, other distributions can be explored, and this will be a direction for future research.

Author Contributions: Conceptualization, J.S.; methodology, T.L. and Y.Y.; validation, J.W. and X.L.; formal analysis, T.L. and J.S.; investigation, T.L. and Y.Y.; resources, T.L., J.S. and J.W.; writing—original draft preparation, J.S., T.L. and Y.Y.; writing—review and editing, J.S., T.L. and X.L.; visualization, Y.Y.; supervision, J.S. and J.W.; project administration, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ZSL	Zero-shot learning

References

1. Palatucci, M.; Pomerleau, D.; Hinton, G.E.; Mitchell, T.M. Zero-shot learning with semantic output codes. In Proceedings of the Advances in Neural Information Processing Systems 22 (NIPS 2009), Vancouver, BC, Canada, 7–10 December 2009.
2. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; Mikolov, T. Devise: A deep visual-semantic embedding model. In Proceedings of the Advances in Neural Information Processing Systems 26 (NIPS 2013), Lake Tahoe, NV, USA, 5–10 December 2013.
3. Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C. Label-embedding for attribute-based classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013), Portland, OR, USA, 23–28 June 2013.
4. Weston, J.; Bengio, S.; Usunier, N. Wsabie: Scaling up to large vocabulary image annotation. In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence (IJCAI 2011), Barcelona, Spain, 16–22 July 2011; Volume 3, pp. 2764–2770.
5. Toutanova, K.; Chen, D.; Pantel, P.; Poon, H.; Choudhury, P.; Gamon, M. Representing text for joint embedding of text and knowledge bases. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), Lisbon, Portugal, 17–21 September 2015; pp. 1499–1509.
6. Suthaharan, S.; Suthaharan, S. Support vector machine. In *Machine Learning Models and Algorithms for Big DATA Classification: Thinking with Examples for Effective Learning*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 207–235.
7. Romera-Paredes, B.; Torr, P. An embarrassingly simple approach to zero-shot learning. In Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), Lille, France, 6–11 July 2015; pp. 2152–2161.
8. Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q.; Hein, M.; Schiele, B. Latent embeddings for zero-shot classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVP 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 69–77.
9. Kodirov, E.; Xiang, T.; Gong, S. Semantic autoencoder for zero-shot learning. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 3174–3183.
10. Lampert, C.H.; Nickisch, H.; Harmeling, S. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 453–465. [CrossRef] [PubMed]
11. Zhang, L.; Xiang, T.; Gong, S. Learning a deep embedding model for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 2021–2030.
12. Lu, C.; Krishna, R.; Bernstein, M.; Fei-Fei, L. Visual relationship detection with language priors. In Proceedings of the European Conference on Computer Vision (ECCV 2016), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 852–869.
13. Stafylakis, T.; Tzimiropoulos, G. Zero-shot keyword spotting for visual speech recognition in-the-wild. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018.

14. Long, Y.; Liu, L.; Shao, L.; Shen, F.; Ding, G.; Han, J. From Zero-shot Learning to Conventional Supervised Classification: Unseen Visual Data Synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017.
15. Kodirov, E.; Xiang, T.; Fu, Z.; Gong, S. Unsupervised Domain Adaptation for Zero-Shot Learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015), Santiago, Chile, 7–13 December 2015.
16. Lampert, C.H.; Nickisch, H.; Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR 2009), Miami, FL, USA, 20–25 June 2009.
17. Xian, Y.; Schiele, B.; Akata, Z. Zero-Shot Learning-The Good, the Bad and the Ugly. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017.
18. Rohrbach, M.; Ebert, S.; Schiele, B. Transfer learning in a transductive setting. In Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS 2013), Lake Tahoe, NV, USA, 5–8 December 2013.
19. Song, J.; Shen, C.; Yang, Y.; Liu, Y.; Song, M. Transductive Unbiased Embedding for Zero-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake, UT, USA, 18–22 June 2018.
20. Demirel, B.; Cinbis, R.G.I.C.N. Attributes2Classname: A discriminative model for attribute-based unsupervised zero-shot learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017.
21. Chen, L.; Zhang, H.; Xiao, J.; Liu, W.; Chang, S.F. Zero-Shot Visual Recognition Using Semantics-Preserving Adversarial Embedding Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake, UT, USA, 18–22 June 2018.
22. Xian, Y.; Lorenz, T.; Schiele, B.; Akata, Z. Feature Generating Networks for Zero-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake, UT, USA, 18–22 June 2018.
23. Kumar Verma, V.; Arora, G.; Mishra, A.; Rai, P. Generalized Zero-Shot Learning via Synthesized Examples. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake, UT, USA, 18–22 June 2018.
24. Zhu, Y.; Elhoseiny, M.; Liu, B.; Peng, X.; Elgammal, A. A Generative Adversarial Approach for Zero-Shot Learning From Noisy Texts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake, UT, USA, 18–22 June 2018.
25. Felix, R.; Kumar, V.B.G.; Reid, I.; Carneiro, G. Multi-modal Cycle-consistent Generalized Zero-Shot Learning. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018.
26. Zhang, H.; Long, Y.; Guan, Y.; Shao, L. Triple Verification Network for Generalized Zero-Shot Learning. *IEEE Trans. Image Process.* **2019**, *28*, 506–517. [CrossRef] [PubMed]
27. Mensink, T.; Verbeek, J.; Perronnin, F.; Csurka, G. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In Proceedings of the European Conference on Computer Vision (ECCV 2012), Florence, Italy, 7–13 October 2012.
28. Changpinyo, S.; Chao, W.L.; Gong, B.; Sha, F. Synthesized Classifiers for Zero-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016.
29. Farhadi, A.; Endres, I.; Hoiem, D.; Forsyth, D. Describing objects by their attributes. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009.
30. Zhang, Z.; Saligrama, V. Zero-shot learning via semantic similarity embedding. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015), Santiago, Chile, 7–13 December 2015.
31. Li, Y.; Zhang, J.; Zhang, J.; Huang, K. Discriminative Learning of Latent Features for Zero-Shot Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake, UT, USA, 18–22 June 2018.
32. Jiang, H.; Wang, R.; Shan, S.; Chen, X. Learning Class Prototypes via Structure Alignment for Zero-Shot Recognition. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018.
33. Long, Y.; Liu, L.; Shen, Y.; Shao, L. Towards Affordable Semantic Searching: Zero-shot Retrieval via Dominant Attributes. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018), New Orleans, LA, USA, 2–7 February 2018.
34. Hoffer, E.; Ailon, N. Deep metric learning using triplet network. In Proceedings of the International Workshop on Similarity-Based Pattern Recognition (SIMBAD 2015), Copenhagen, Denmark, 12–14 October 2015.
35. Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; Perona, P. Caltech-UCSD birds 200. In *Computation & Neural Systems Technical Report*; California Institute of Technology: Pasadena, CA, USA, 2010.
36. Nilsback, M.E.; Zisserman, A. Automated flower classification over a large number of classes. In Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP 2008), Bhubaneswar, India, 16–19 December 2008.
37. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

39. Wang, J.; Zhou, F.; Wen, S.; Liu, X.; Lin, Y. Deep metric learning with angular loss. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 2593–2601.
40. Amyar, A.; Ruan, S.; Vera, P.; Decazes, P.; Modzelewski, R. RADIOGAN: Deep convolutional conditional generative adversarial network to generate PET images. In Proceedings of the 7th International Conference on Bioinformatics Research and Applications (ICBRA 2020), Berlin, Germany, 13–15 September 2020; pp. 28–33.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.