applied sciences

MDPI

Article

Enhancing Semantic-Consistent Features and Transforming Discriminative Features for Generalized Zero-Shot Classifications

Guan Yang [1,2], Ayou Han [1,2], Xiaoming Liu [1,2,*], Yang Liu [3], Tao Wei [4] and Zhiyuan Zhang [1,2]

[1] School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China
[2] Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou 450007, China
[3] School of Telecommunications Engineering, Xidian University, Xi'an 710071, China
[4] School of Software, Henan University of Engineering, Zhengzhou 450007, China
* Correspondence: ming616@zut.edu.cn

Abstract: Generalized zero-shot learning (GZSL) aims to classify classes that do not appear during training. Recent state-of-the-art approaches rely on generative models, which use correlating semantic embeddings to synthesize unseen classes visual features; however, these approaches ignore the semantic and visual relevance, and visual features synthesized by generative models do not represent their semantics well. Although existing GZSL methods based on generative model disentanglement consider consistency between visual and semantic models, these methods consider semantic consistency only in the training phase and ignore semantic consistency in the feature synthesis and classification phases. The absence of such constraints may lead to an unrepresentative synthesized visual model with respect to semantics, and the visual and semantic features are not modally well aligned, thus causing the bias between visual and semantic features. Therefore, an approach for GZSL is proposed to enhance semantic-consistent features and discriminative features transformation (ESTD-GZSL). The proposed method can enhance semantic-consistent features at all stages of GZSL. A semantic decoder module is first added to the VAE to map synthetic and real features to the corresponding semantic embeddings. This regularization method allows synthesizing unseen classes for a more representative visual representation, and synthetic features can better represent their semantics. Then, the semantic-consistent features decomposed by the disentanglement module and the features output by the semantic decoder are transformed into enhanced semantic-consistent discriminative features and used in classification to reduce the ambiguity between categories. The experimental results show that our proposed method achieves more competitive results on four benchmark datasets (AWA2, CUB, FLO, and APY) of GZSL.

Keywords: generalized zero-shot learning; disentangled representation; semantic consistency; enhanced features

Citation: Yang, G.; Han, A.; Liu, X.; Liu, Y.; Wei, T.; Zhang, Z. Enhancing Semantic-Consistent Features and Transforming Discriminative Features for Generalized Zero-Shot Classifications. *Appl. Sci.* **2022**, *12*, 12642. https://doi.org/10.3390/app122412642

Academic Editor: Rubén Usamentiaga

Received: 27 October 2022
Accepted: 7 December 2022
Published: 9 December 2022

1. Introduction

The high-speed development of deep learning is dependent on a large amount of labeled data. However, in the real world, the collection of large-scale labeled samples is a very difficult problem, and some specific categories do not have a large number of labeled samples, such as species with endangered statuses, and sample information with respect to these species is extremely difficult to obtain. In addition, canonical deep learning models can only recognize categories that have already been observed during the training phase (seen classes) and cannot recognize classes that have not been seen by the model (unseen classes). It is a challenge to train the model by using only samples from the seen classes to recognize samples from the unseen classes.

Humans can use previous experiences to quickly learn new concepts. For example, suppose a child has never seen a zebra (zebra is an unseen class to child), but he knows that a zebra is a horse-like animal with black and white stripes (horse is a seen class to child); then, when he sees a zebra for the first time, he usually recognizes it as a zebra [1].

This is the core idea of the zero-shot learning approach, which is to use common sense or prior knowledge for deductive reasoning. The zero-shot learning (ZSL) [2] method provides a good solution to the above challenges. The traditional ZSL approach is based on the assumption that the test set contains only unseen classes, which is very easy to violate in reality. Therefore, a generalized zero-shot learning (GZSL) [3] approach emerged that is broader and more challenging than ZSL; that is, the test set contains seen and unseen classes.

There is a challenge in combining the knowledge learned by humans with the model's rules. Ranaldi et al. [4] proposed a new vision of knowledge in AI models based on a combination of rules, learning, and human knowledge. This study attempts to provide "form" to the origins and development of artificial intelligence and thoroughly considers the ability of neural networks to learn from so-called hooks. GZSL is a method that combines human knowledge with model inference. Recent GZSL classification methods [5–10] are usually based on generative models, such as generative adversarial networks (GAN) [11] and variational autoencoder (VAE) [12], etc. The goal of the generative model approach is to optimize the difference between real and generated data. Based on the sliding mode control, Gohari et al. [13] proposed a new strategy for a self-adjusting boundary layer using a robust controller to prevent the chattering phenomenon; this method can effectively suppress the transmitted noise and maintain the consistency of the system. Some other works [14,15] use GAN to generate unseen features and also design auxiliary modules, such as decoders, classifier, etc. The reconstruction of semantic embeddings is performed during training to establish cycle-consistent constraints, and this auxiliary module allows the a generative model to synthesize semantic-consistent features. However, these constraints are only used in the training phase and are discarded in the classification phase; they are rarely used in VAE-based methods. These modules can synthesize discriminative features in the feature synthesis phase and can reduce ambiguities between categories in the classification phase. SDGZSL [10] learns a VAE by instances of the seen classes and the corresponding semantic embedding and then synthesizes the corresponding instances for the unseen classes by the trained VAE and trains a disentanglement module together with the seen classes instances to separate semantic-unrelated and semantic-consistent features. However, the focus on semantic-consistent in this study is only in the training phase, and semantic consistency is neglected in the feature generation and classification phases, resulting in synthesized unseen classes instances that do not represent their semantics well and visual features and corresponding semantic embeddings that are not well aligned in the modalities. Using only one module to extract semantic-consistent features is often not enough; the supervision of the corresponding semantic information is also missing in the disentangling module. Therefore, it will lead to the problematic domain shifts.

To solve the above problems, we improve the SDGZSL method and propose an disentanglement generalized zero-shot classification method involving enhanced semantic-consistent and discriminative feature transformation (ESTD-GZSL). A semantic decoder is first added to decode the seen class instances and the synthesized instances of unseen classes back into the semantic space, by which we can make the synthesized instances better represent their semantics, enhance the alignment of the two modalities, and reduce the offset between visual and semantic features. Then, we feed the output of this decoder as a supervised signal into the disentanglement auto-encoder to reconstruct visual features. Finally, since the semantic decoder is trained with cycle-consistent loss, we consider that the output obtained by the visual features via the semantic decoder is also semantic-consistent. Therefore, we further improved the classification module of the method. That is, generalized zero-shot classification is performed by enhanced semantic-consistent discriminant features. Our source code is available at https://github.com/HanAccount/ESTD-GZSL (accessed on 6 December 2022). The proposed approach is based on SDGZSL [10] with the following contributions:

- In order to alleviate the domain shift problem. An additional semantic decoder structure is added to enable the generator synthesized instances to better represent

their semantics, strengthen the semantic and visual alignment, and reduce the bais between visual and semantic.
- To enhance the ability of the disentanglement module to break down semantic consistency, we input the semantic decoder output as a supervised signal into the feature reconstruction of the disentanglement auto-encoder.
- We improve the generalized zero-shot classifier by introducing transformation discriminative features in the classification stage and splicing semantic consistency features output from the semantic decoder and semantic consistency features from the disentanglement module decomposition to enhance semantic consistency and reduce the ambiguity between different categories.
- The proposed method achieves more competitive results on four datasets in GZSL than baseline.

2. Related Work

2.1. Generalized Zero-Shot Learning

The recently advanced GZSL approach is implemented by generative models, which can synthesize their visual features by semantically embedding unseen classes. With these synthetic visual features of unseen classes, the ZSL problem becomes a relatively simple supervised classification problem. The classifier of seen classes in [14] is replaced by the integration of decoder and cycle-consistent loss [16], and CADA-VAE [17] introduces cross-reconstruction and distribution alignment loss to align the potential representations of the two modes using two VAEs. f-VAEGAN-D2 [18] combines the advantages of VAE and GAN by sharing the encoder of VAE and the generator of GAN to synthesize features. FREE [7] proposes a feature reinforcement approach that uses self-adaptive margin center loss to reduce the bias between the dataset used for training backbone and the GZSL task dataset. OT-GZSL [19] establishes optimal transmissions between synthetic and real feature distributions, while CE-GZSL [20] proposes a hybrid GZSL framework that mixes generative and embedding-based approaches using contrast learning, and TF-VAEGAN [6] proposes a feedback module that feeds the decoder output back to the generator. RPGN [21] proposes a residual-prototype-generating network to extract the residual visual features from the original visual features and to synthesize the prototype's visual features associated with semantic attributes by a disentangle regressor.

2.2. Disentangled Representation Learning

Traditional disentangled representation learning is performed by decomposing the original features into multiple mutually independent factors via an encoder–decoder structure [22]; generally speaking, the better the feature representation ability of the disentanglement, the better the learning ability of the model. β-VAE [23] leverages to balance the independence and reconstruction performance of the decoupling factor by adjusting the weights of the KL term. FactorVAE [24] suggests that the distribution of representation be factorial to decompose features. To achieve cross-dimensional independence, DLFZRL [25] proposes a hierarchical decomposition method to learn distinguished latent features. InfoGAN [26] achieves disentangling by maximizing the mutual information between latent and original feature variables. Li et al. [27] proposes a new classification method, disentangled-VAE, which aims to decompose classification extraction factors and classification dispersion factors from visual and semantic features. SDGZSL [10] leverages a total correlation penalty-based disentangled auto-encoder module to decompose semantically consistent and semantically irrelevant latent representations in visual features. This paper is a research improvement on the model based on SDGZSL.

3. Method

3.1. Problem Definition

For zero-shot learning (ZSL), the dataset is divided into seen dataset $\mathcal{D}_s$ and unseen class dataset $\mathcal{D}_u$. The categories are $\mathcal{Y}_s$ and $\mathcal{Y}_u$, $\mathcal{Y}_s \cap \mathcal{Y}_u = \oslash$. We have a training set, which consists

only of samples marked in the seen classes: $\mathcal{D}_s^{tr} = \{x_s, a_s, y_s | x_s \in \mathcal{X}_s, a_s \in \mathcal{A}_s, y_s \in \mathcal{Y}_s\}$, where $x_s \in \mathcal{X}_s$ denotes the visual features of seen classes, $a_s \in \mathcal{A}_s$ is semantic descriptor of category (e.g., semantic attribute), and $y_s \in \mathcal{Y}_s$ is class labels of seen classes. There is also an unseen class test set $\mathcal{D}_u^{te} = \{x_u, a_u, y_u | x_u \in \mathcal{X}_u, a_u \in \mathcal{A}_u, y_u \in \mathcal{Y}_u\}$, where the visual features of the unseen classes are not available during training. Traditional zero-shot learning aims to learn test set $\mathcal{D}_{te} = \{\mathcal{D}_u^{te}\}$ on the evaluated classifier $f^{ZSL} : \mathcal{X}_u \rightarrow \mathcal{Y}_u$. However, in generalized zero-shot learning, test set $\mathcal{X}_{te}$ consists of both seen and unseen classes, it is the classifier that learns the evaluation on all features $f^{GZSL} : \mathcal{X} \rightarrow \mathcal{Y}_s \cup \mathcal{Y}_u$. In this paper, we focus on the classification of GZSL tasks.

3.2. Model Overview

The model architecture of the proposed method is shown in Figure 1. The proposed approach is divided into two phases. The first phase is introduced in Figure 1 and corresponds to Sections 3.2.1–3.2.5, and the second phase is for the final classification, corresponding to Figure 2 and Section 3.2.6.

In Figure 1, the structure of the proposed method consists of three modules: (i) the visual feature generation module, consisting of a variational encoder Q and a variational decoder P; (ii) a disentanglement module, consisting of encoder E, decoder D, relation module R, and discriminator Dis (where R and Dis are the structures proposed by SDGZSL [10], which are not modified in this paper and therefore are not drawn in Figure 1); and (iii) a semantic decoder module, consisting of a semantic decoder Dec. First, the feature generation module synthesizes the corresponding visual features using a conditional variational autoencoder (CVAE) [28] from the unseen classes semantic embeddings; this module uses visual features from seen classes for training. Then, a semantic decoder is used to decode the synthesized visual features from unseen classes and real features from seen classes and reconstructs them relative to the corresponding semantic space in order to ensure a better alignment between the synthesized visual features and corresponding semantics. Finally, the synthesized unseen classes' visual features and seen classes' real features are disentangled into semantic-consistent and semantic-unrelated visual features using disentanglement module encoder E; here, the output of the semantic decoder Dec is fed into disentanglement decoder D to reconstruct the original visual features.

3.2.1. Visual Feature Generation Module

Since the visual features of the unseen classes cannot be used in training, the conditional variational autoencoder (CVAE) [28] is used here to synthesize the corresponding visual features for the unseen classes. CVAE is trained by visual features of seen classes x_s and semantic embeddings a_s. The objective function of CAVE can be written as follows:

$$\begin{aligned} \mathcal{L}_{\mathrm{CVAE}} = & -D_{KL}\left[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_s, \boldsymbol{a}_s) \| p_\theta(\boldsymbol{z} \mid \boldsymbol{a}_s)\right] \\ & + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}_s, \boldsymbol{a}_s)}[\log p_\theta(\boldsymbol{x}_s \mid \boldsymbol{z}, \boldsymbol{a}_s)] \end{aligned} \tag{1}$$

where the first term is the Kullback–Keibler (KL) divergence between $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_s, \boldsymbol{a}_s)$ and $p_\theta(\boldsymbol{z} \mid \boldsymbol{a}_s)$, and the second term is the reconstruction loss. x_s and a_s are the seen classes' visual features and the seen classes' semantic embeddings.The encoder Q of the CVAE module uses the visual features of the seen classes x_s and semantic embeddings a_s to generate latent variables $\boldsymbol{z}$. CVAE decoder P uses latent variables $\boldsymbol{z}$ and seen class semantic embeddings a_s to synthesize visual features.

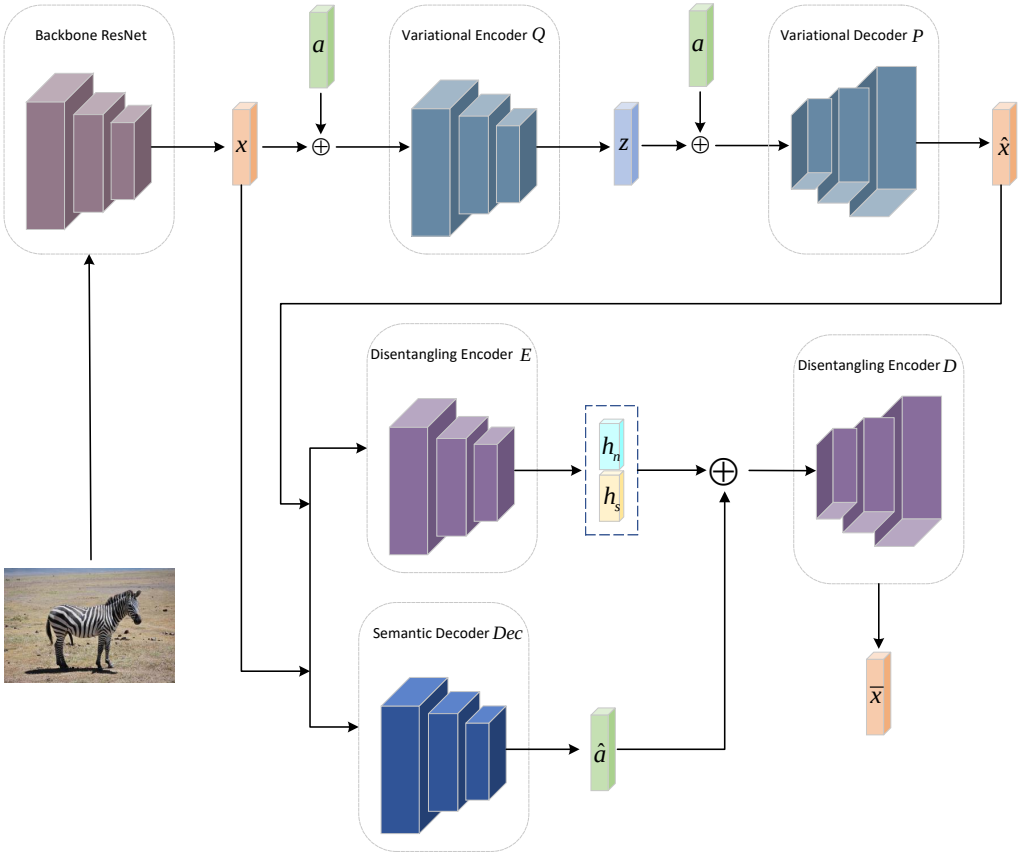


Figure 1. The proposed architecture for ESTD-GZSL.

3.2.2. Semantic Decoder Module

To create visual features (synthesized by CVAE) that better express their semantics, a semantic decoder *Dec* module, $\mathcal{X} \longrightarrow \mathcal{A}$, is introduced here to reconstruct the synthesized visual features relative to their corresponding semantic embeddings, and a cycle-consistent loss is used for the reconstructed semantic embeddings to ensure that the synthesized visual features can be reconstructed relative to their corresponding semantic embeddings. This module can ensure that the visual features and their semantics can be modally aligned, reducing semantic bias and enhancing semantic-consistent features. The cycle-consistent loss of the semantic embedding is achieved via the $\mathcal{L}_1$ reconstruction loss as follows.

$$\mathcal{L}_R = \mathbb{E}[\|Dec(x) - a\|_1] + \mathbb{E}[\|Dec(\hat{x}) - a\|_1] \tag{2}$$

3.2.3. Disentanglement Module

In order to enhance the ability of the disentanglement module to decompose the semantic consistency features, the proposed approach improves the visual reconstruction function of this module in SDGZSL. First, the visual features are encoded as potential features h by disentanglement encoder E. Then, h is decomposed into semantic-consistent features, h_s, and semantic-unrelated features, h_n, as follows.

$$E(x) = h = [h_s, h_n] \tag{3}$$

In the disentanglement decoder D, we use the output of semantic decoder *Dec* as a supervisory signal to reconstruct visual features jointly with the latent features h of the output of disentanglement encoder D. The ability of the disentanglement module

to decompose the semantic-consistent features can be enhanced by this approach. The reconstruction loss is as follows:

$$\mathcal{L}_{rec} = \sum_{x \in X^s} \|x - D[h, Dec(x)]\|^2 \tag{4}$$

3.2.4. Enhance and Transformation Semantic-Consistent Discriminative Features

In SDGZSL's work, only a single disentanglement module is used to extract semantic-consistent features; we believe that the semantic-consistent features extracted using only one module are not sufficient. Since our proposed semantic decoder *Dec* is trained with the semantic embedding via cycle-consistent loss, therefore, the output of *Dec* is also highly consistent with the semantics, and the output of *Dec* can also be considered as a semantic-consistent feature. Thus, we propose enhancing and transforming semantic-consistent features by transforming the semantic-consistent features decomposed by the disentanglement module and the semantic-consistent features output by *Dec* into an enhanced semantic-consistent discriminative feature for the final classification.

CVAE is a mapping that learns "single semantic embedding to multiple instances", and the semantic decoder is an inverse mapping that learns "multiple instances to one semantic embedding". We believe that *Dec* and CVAE can encode the complementary information of categories, so using the output information of *Dec* in the classification stage can reduce the ambiguity between feature instances of different categories.

3.2.5. Total Loss

In order to learn the enhanced semantic-consistent features, the proposed overall loss of the method is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{SDGZSL} + \beta * \mathcal{L}_R \tag{5}$$

where $\mathcal{L}_{SDGZSL}$ is the training loss proposed by baseline, but the reconstruction loss of the disentanglement module in this loss is replaced by Equation (4) proposed in this paper. β is a hyperparameter that measures the loss of weight relative to semantic decoder *Dec*.

3.2.6. Generalized Zero-Shot Classification

The classification process of the proposed method in this paper is shown in Figure 2. First, the decoder of CVAE generates unseen visual features with unseen classes of semantic embeddings, a_u, and Gaussian noise, z. Then, the generated visual features of the unseen classes and the real features of the seen classes are further fed into semantic decoder *Dec* and the disentanglement module to extract the corresponding semantic-consistent features. Finally, the semantic-consistent features output by *Dec* are concatenated with the semantic-consistent features decomposed by the disentanglement module to obtain the transformed and enhanced semantic-consistent discriminative features $h_s^s \oplus h_r^s$ and $h_s^u \oplus h_r^u$; they are then used to train the generalized zero-shot classifier, f^{GZSL}.

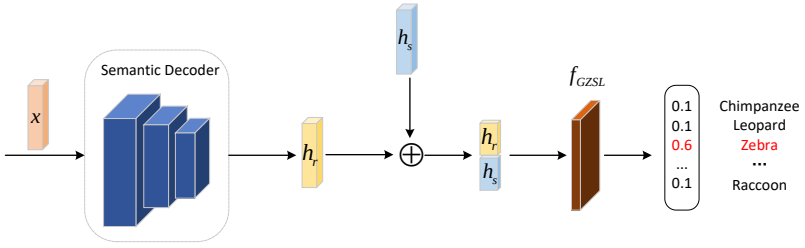


Figure 2. Classification architecture. A feature transformation is performed by concatenating ($\oplus$) the semantic-consistent features, h_s, of the output of the disentanglement module with the semantic-consistent features, h_r, of the output of semantic decoder module, *Dec*. The transformed discriminative features are then used for GZSL classification.

3.3. Training and Inference

Algorithm 1 shows the pseudocode of the model's training. We iteratively train the model with the overall framework for N_{iter} steps. In Algorithm 1, the weight for L_R is denoted as β. When the training of the model converges, the generative network P, the disentangling encoder E, and the semantic decoder Dec can generate two semantic-consistent unseen classes features from Gaussian noise z and the unseen classes' semantic embeddings a_u. We can extract semantic-consistent features h_s^s and h_r^s from training seen features using modules E and Dec, together with the generated unseen semantic-consistent features h_s^u and h_r^u. We can concatenate two semantic-consistent features (h_s and h_r) to train a supervised classifier. In this paper, a softmax classifier is adopted for evaluation.

Algorithm 1 ESTD-GZSL training

Input: training data $\{X^s, Y^s\}$, semantic embeddings A^s, learning rate λ
while model not converged **do**
  Randomly select a batch data $\{x^s_{(t)}, y^s_{(t)}\}_{t=1}^{B}$, $\{a^s_{(c)}\}_{c=1}^{N}$
  for step = 0,...,N_{iter} **do**
    Compute CVAE loss $\mathcal{L}_{CVAE}$ by Equation (1)
    Computer visual reconstruction loss $\mathcal{L}_{rec}$ by Equation (4)
    Computer semantic cycle-consistent loss $\mathcal{L}_R$ by Equation (2)
    Computer $\mathcal{L}_{all1} = \mathcal{L}_{SDGZSL} + \mathcal{L}_{rec} + \beta * \mathcal{L}_R$ by Equations (1)–(5)
    Update $\nabla \mathcal{L}_{all1}$
  end for
  Randomly select a batch data $\{x^s_{(t)}, y^s_{(t)}\}_{t=1}^{B}$,$\{a^s_{(c)}\}_{c=1}^{N}$
  Computer $\mathcal{L}_{all2} = \mathcal{L}_{SDGZSL} + \mathcal{L}_{rec} + \beta * \mathcal{L}_R$ by Equations (1), (2), (4) and (5)
  Update $\nabla \mathcal{L}_{all2}$
end while
Output: trained generative network P, disentangling encoder E and semantic decoder Dec

4. Experimental Results

4.1. Datasets

In our experiments, we use four popular benchmark datasets to evaluate the performance of our models: Animals with Attributes 2 (AwA2) [29], Caltech-UCSD Birds-200-2011 (CUB) [30], Oxford Flowers (FLO) [2], and Attribute Pascal and Yahoo (APY) [31]. The APY dataset is a coarse-grained dataset consisting of 20 seen classes and 12 unseen classes, each with 64 annotated attributes. The AwA2 dataset is commonly used for animal classification and consists of 40 seen classes and 10 unseen classes, each annotated with 85 attributes. The FLO dataset contains 102 flower classes, 82 seen classes, and 20 unseen classes. The CUB dataset contains 200 species of birds, of which 150 species are seen classes and 50 species are unseen classes. The semantic embeddings of FLO and CUB are 1024-dimensional character-based CNN-RNN features [32] extracted from the fine-grained visual descriptions (10 sentences per image). The details of each dataset are shown in Table 1.

Table 1. Statistics of the datasets used in our experiments, including the dimensions of visual features (visual dimension), the dimensions of semantic vectors per class (attribute), seen class size (seen), and unseen class size (unseen).

Dataset	Visual Dimension	Attribute	Seen	Unseen
CUB	2048	1024	150	50
FLO	2048	1024	82	20
AWA2	2048	85	40	10
APY	2048	64	20	12

4.2. Evaluation Protocols

The metric evaluated on the GZSL task uses a harmonic mean, which calculates the joint accuracy of the seen and unseen classes, and it can be written as $H = (2 \times U \times S)/(U + S)$, where U and S denote the average per-class top-1 accuracy of unseen and seen classes, respectively. A high harmonic mean indicates the good performance of both seen and unseen classes.

4.3. Implementation Details

Following the options of most methods, we first use the pre-trained ResNet101 [33] to extract image features of dimension 2048. We use three fully connected layers with 2048 hidden units for the VAE encoder and decoder. LeakyReLU [34] is used as an activation function. We use the same hyperparameters of the visual feature generation module and disentanglement module as SDGZSL for fair comparisons. We use the 64 mini-batch sizes in all datasets, and the learning rate is set to 0.0001. For semantic decoder *Dec*, we use two fully connected layers with twice the semantic embedding dimension. Three fully connected layers for the disentanglement module are used with 2048 hidden units. We optimize all networks with the Adam [35] optimizer for each module. We use a single fully connected layer for the classifier for GZSL. All models are implemented with PyTorch 1.10.0, CUDA 11.2.0, and cuDNN 8.2.0. We use a single RTX 3060 6 GB GPU for each training operation.

4.4. Comparison with State-of-the-Art Methods

To compare ESTD-GZSL with the recent GZSL methods, we select the recent state-of-the-art generative-based methods such as f-CLSWGAN [8], CANZSL [36], cycle-CLSWGAN [14], FREE [7], LisGAN [37], CADA-VAE [17], f-VAEGAN-D2 [18], TF-VAEGAN [6], CE-GZSL [20], RPGN [21], and SDGZSL [10]. Table 2 shows the GZSL performance of the compared methods and ours with and without finetuning the backbone on the datasets.

Our method achieves better performances than the baseline on CUB, FLO, and APY, and it obtains the second-best H results of AwA2, where SDGZSL decomposes visual features into those that are semantic-consistent and semantic-unrelated. However, we believe that using only a single encoder to extract semantic-consistent features is not sufficient, and the visual features synthesized by VAE do not express their semantics well, which can lead to semantic and visual biases. Therefore, to improve disentanglements, we introduced a semantic decoder module that maps the visual space to the semantic space by using cycle-consistent loss to reduce bias. We combine the semantic decoder's output with the disentanglement encoder output for the reconstruction of the original image to improve generalization. In the classification phase, we transform semantic-consistent features decomposed by the semantic decoder output and the disentanglement module into enhanced semantic-consistent discriminative features for classification. Compared with the performance of SDGZSL, as shown in Table 2, for all the datasets without finetuning, our method outperforms SDGZSL in H results, which improved by 1.2% on the AwA2 dataset, improved by 1.9% on the CUB dataset, improved by 1.8% on the FLO dataset, and improved by 0.5% on the APY dataset. For all the finetuned datasets, our method outperforms SDGZSL in H results on CUB, FLO, and APY, and comparable H results were also obtained on AwA2, which improved by 2.3% on the CUB, improved by 0.7% on the FLO, and improved by 0.3% on the APY. The proposed method achieves better results because it is able to generate semantically aligned visual features for unseen classes and extract more sufficient semantic-consistent features. The experimental results show that the proposed method can improve the effect of the baseline, which proves the effectiveness of the method.

Table 2. Performance comparison in accuracy(%) on four datasets. We show the accuracies of seen and unseen classes and their harmonic mean for GZSL, which are denoted as U, S, and H. * means a finetuned backbone is used. Red font and blue font denote the highest and the second highest results, respectively.

Methods	AwA2			CUB			FLO			APY		
	U	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*
f-CLSWGAN [8]	56.1	65.5	60.4	43.7	57.7	49.7	59	73.8	65.6	32.9	61.7	42.9
CANZSL [36]	49.7	70.2	58.2	47.9	58.1	52.5	58.2	77.6	66.5	-	-	-
LisGAN [37]	52.6	76.3	62.3	46.5	57.9	51.6	57.7	83.8	68.3	34.3	68.2	45.7
CADA-VAE [17]	55.8	75.0	63.9	51.6	53.5	52.4	51.6	75.6	61.3	31.7	55.1	40.3
f-VAEGAN-D2 [18]	57.6	70.6	63.5	48.4	60.1	53.6	56.8	74.9	64.6	-	-	-
TF-VAEGAN [6]	59.8	75.1	66.6	52.8	64.7	58.1	62.5	84.1	71.7	-	-	-
TF-VAEGAN * [6]	55.5	83.6	66.7	63.8	79.3	70.7	69.5	92.5	79.4	-	-	-
FREE [7]	60.4	75.4	67.1	55.7	59.9	57.7	67.4	84.5	75.0	-	-	-
cycle-CLSWGAN [14]	-	-	-	59.3	47.9	53.0	59.2	72.5	65.1	-	-	-
CE-GZSL [20]	63.1	78.6	70.0	63.9	66.8	65.3	69.0	78.7	73.5	-	-	-
RPGN [21]	68.3	78.8	73.2	61.0	62.2	61.6	68.2	88.9	77.0	-	-	-
SDGZSL [10]	64.6	73.6	68.8	59.9	66.4	63.0	83.3	90.2	86.6	38.0	57.4	45.7
SDGZSL * [10]	69.6	78.2	73.7	73.0	77.5	75.1	86.1	89.1	87.8	39.1	60.7	47.5
ESTD-GZSL(Ours)	65.1	73.4	70.0	65.3	64.5	64.9	85.2	91.9	88.4	36.5	63.2	46.2
ESTD-GZSL *(Ours)	66.6	81.2	73.2	74.5	80.2	77.4	83.8	93.6	88.5	35.2	74.8	47.8

4.5. Zero-Shot Retrieval Results

For the sake of fairness, we follow the zero-shot retrieval protocol proposed in SDGZSL [10]. First, the backbone network (ResNet-101) extracts the visual features from all unseen images. Then, the disentanglement encoder extracts semantic-consistent features in the unseen visual features. Third, the semantic decoder further extracts the semantic-consistent features in the unseen visual features. Finally, these two semantic-consistent features are transformed into enhanced semantic-consistent features and the centroid point is computed as a retrieval query, which is further used to retrieve the nearest samples. To evaluate the performance on the retrieved samples, the mean average precision (mAP) score is adopted. In Figure 3, we compare our propose ESTD method with CVAE and SDGZSL when retrieving 100%, 50%, and 25% of the images from all unseen classes on APY, AwA2, CUB, and FLO. It can be seen that the proposed method can significantly boost the retrieval performance among all settings, which can also demonstrate the effectiveness of the proposes method from the retrieval perspective.

4.6. Model Analysis

In this subsection, we perform an experimental analysis of the proposed method with the CUB dataset.

4.6.1. Ablation Study

In this ablation study, we evaluated the usefulness of each of the proposed methods. In Table 3, we show the results of the ablation experiments. Experiment (a) only adds a semantic decoder to the baseline for training, and the GZSL's performance, H, is enhanced by 2.1%. Experiment (b) adds the semantic decoder's output to the visual reconstruction of the disentanglement module, and the GZSL's performance, H, is enhanced by 2.1%. Experiment (c) involves training a classifier by joining the semantic-consistent features from the semantic decoder's output and the disentanglement encoder output; the GZSL performance, H, is enhanced by 2.3%.

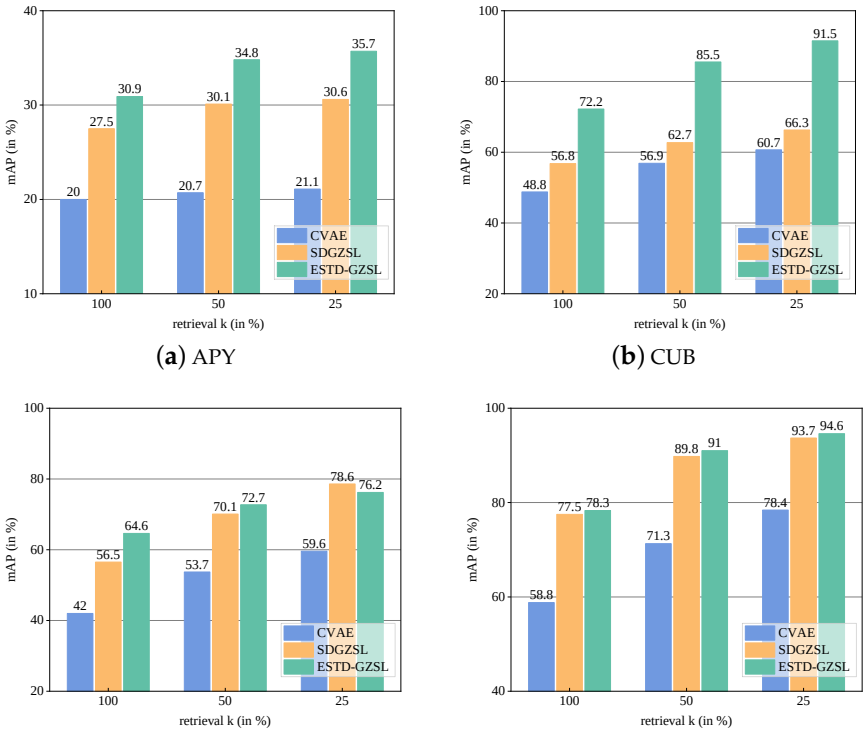


Figure 3. Zero-shot image retrieval result comparison between CVAE, SDGZSL, and ESTD-GZSL.

Table 3. Ablation study for different components of ESTD-GZSL on the CUB dataset. The best results are marked in **boldface**.

Methods	*U*	*S*	*H*
Baseline	73.0	77.5	75.1
(a)	**75.0**	79.6	77.2
(b)	74.6	80.0	77.2
(c)	73.8	**81.4**	**77.4**

In Figure 4, we show the performance comparison between SDGZSL and h_s, h_r, and $h_s \oplus h_r$. It can be seen that a classifier trained using only the feature's output from the semantic decoder can achieve comparable results to SDGZSL. This shows that using semantic-consistent features of the semantic decoder (*Dec*)'s output h_r can further improve performance.

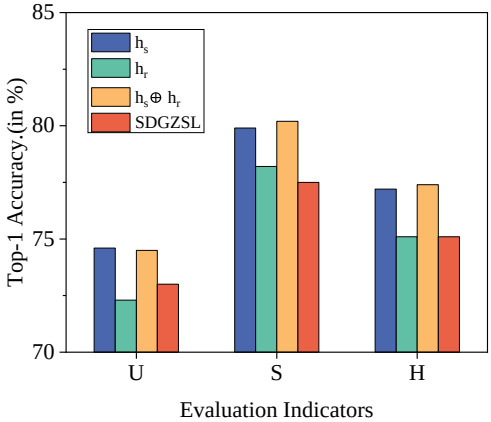
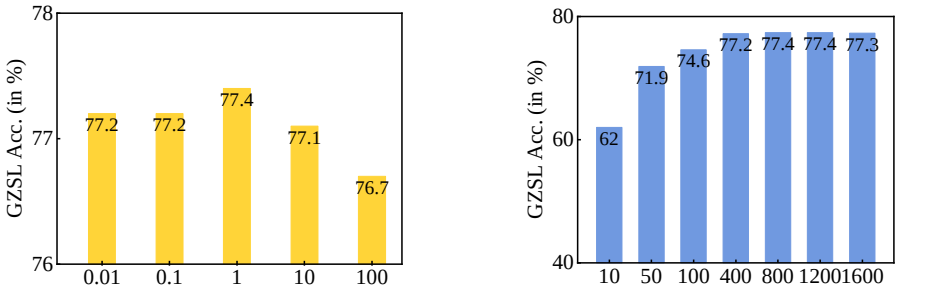


Figure 4. U, S, and H accuracy (%) comparison between SDGZSL, h_s, h_r, and $h_s \oplus h_r$.

4.6.2. Hyper-Parameter Study

In this experiment, we evaluate the impact of the number of synthesized unseen visual features and the loss of weight, β, on the semantic decoder, Dec. Figure 5 shows the harmonic mean (H) when changing the L_R weight β from 0.01 to 100 and the number of synthesized unseen visual features from 10 to 1600. For the loss weight β of semantic decoder Dec, when β equals 1, our model achieves the best H result (77.4%), and when β becomes larger, the H result becomes lower. For the number of synthesized unseen visual features, it produces the best H result (77.4%) in both 800 and 1200.

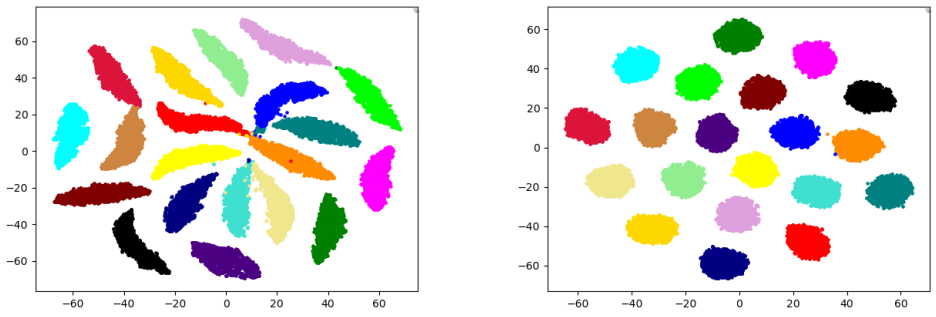


(a) semantic decoder loss weight β (b) numbers of synthesized unseen visual features

Figure 5. The effect of hyper-parameters L_R weight β and the number of synthesized unseen visual features.

4.6.3. T-SNE Visualization

To further validate the ability of the proposed method to synthesize unseen visual features, we visualize the distribution of unseen visual features synthesized by SDGZSL and ESTD-GZSL (ours) in Figure 6. We choose 20 unseen classes from the CUB dataset that have enough classes to show the class-wise comparison. Clearly, our proposed method generates more discriminative and robust visual representations. This will have a great positive impact on the disentanglement module.



(**a**) SDGZSL (**b**) ESTD-GZSL

Figure 6. t-SNE visualization of the synthesized unseen visual representations of 20 unseen classes on CUB: (**a**) baseline: SDGZSL; (**b**) ours: ESTD-GZSL .

5. Conclusions

In this paper, we propose a generalized zero-shot learning classification method with enhanced semantic-consistent discriminative features, which can effectively alleviate the domain shift problem in GZSL. The proposed method can enhance semantic-consistent features at all stages of GZSL. It can use real seen classes' features to assist in synthesizing unseen classes features of semantic alignments to further extract semantic-consistent features. Then, the two semantic-consistent features are transformed into enhanced semantic consistency features for training the classifier. Experiments on multiple datasets, ablation experiments, hyper-parameter analysis experiments, and t-SNE visualization experiments show that our method is comparable or superior to the existing method. Furthermore, we

further apply the proposed method to a zero-shot retrieval task for a comparison with the baseline and achieved better results than the baseline.

Compared with the same type of methods, the proposed method achieves more competitive results. However, its performance is limited by the quality of the visual features extracted by the pre-trained neural network, which is because the ability and quality of visual features extracted by different deep models are different. In addition, the performance of existing methods is closely related to semantics, and this kind of strong prior external knowledge is difficult to obtain in real scenarios. How to break through the restriction of such external semantic knowledge is a research difficulty. The proposed method focuses on image type data and may not work well for video type data.

Author Contributions: Conceptualization, G.Y., A.H. and X.L.; methodology, A.H., G.Y. and X.L.; validation, A.H., G.Y., X.L. and Z.Z.; formal analysis, A.H., G.Y. and T.W.; investigation, A.H.; resources, G.Y. and Y.L.; writing—original draft preparation, A.H.; writing—review and editing, A.H. and G.Y.; visualization, A.H. and Z.Z.; supervision, G.Y. and X.L.; funding acquisition, G.Y. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Outstanding Youth Science Fund Project of National Natural Science Foundation of China (grant no. 61906141), the Key Laboratory for Applied Statistics of MOE, Northeast Normal University (grant no. 135131007), and Key Scientific Research Projects of Colleges and Universities in Henan Province (grant no. 23A520022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the second author.

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

References

1. Pang, Y.; Wang, H.; Yu, Y.; Ji, Z. A decadal survey of zero-shot image classification. *Sci. Sin. Infor.* **2019**, *49*, 1299–1320. [CrossRef]
2. Larochelle, H.; Erhan, D.; Bengio, Y. Zero-data Learning of New Tasks. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI 2008), Chicago, IL, USA, 13–17 July 2008; pp. 646–651.
3. Chao, W.L.; Changpinyo, S.; Gong, B.; Sha, F. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 52–68.
4. Ranaldi, L.; Fallucchi, F.; Zanzotto, F.M. Dis-Cover AI Minds to Preserve Human Knowledge. *Future Internet* **2022**, *14*, 10. [CrossRef]
5. 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), Salt Lake City, UT, USA, 18–23 June 2018; pp. 1004–1013.
6. Narayan, S.; Gupta, A.; Khan, F.S.; Snoek, C.G.; Shao, L. Latent embedding feedback and discriminative features for zero-shot classification. In Proceedings of the ECCV 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 479–495.
7. Chen, S.; Wang, W.; Xia, B.; Peng, Q.; You, X.; Zheng, F.; Shao, L. FREE: Feature refinement for generalized zero-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 122–131.
8. Xian, Y.; Lorenz, T.; Schiele, B.; Akata, Z. Feature generating networks for zero-shot learning. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5542–5551.
9. Keshari, R.; Singh, R.; Vatsa, M. Generalized zero-shot learning via over-complete distribution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 13300–13308.
10. Chen, Z.; Luo, Y.; Qiu, R.; Huang, Z.; Li, J.; Zhang, Z. Semantics Disentangling for Generalized Zero-shot Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
11. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
12. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2014**, arXiv:1312.6114.

13. Gohari, H.; Zarastvand, M.; Talebitooti, R.; Loghmani, A.; Omidpanah, M. Radiated sound control from a smart cylinder subjected to piezoelectric uncertainties based on sliding mode technique using self-adjusting boundary layer. *Aerosp. Sci. Technol.* **2020**, *106*, 106141. [CrossRef]
14. Felix, R.; Reid, I.; Carneiro, G. Multi-modal cycle-consistent generalized zero-shot learning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 21–37.
15. Huang, H.; Wang, C.; Yu, P.S.; Wang, C.D. Generative Dual Adversarial Network for Generalized Zero-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
16. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
17. Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; Akata, Z. Generalized zero-and few-shot learning via aligned variational autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8247–8255.
18. Xian, Y.; Sharma, S.; Schiele, B.; Akata, Z. F-VAEGAN-D2: A feature generating framework for any-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10275–10284.
19. Wang, W.; Xu, H.; Wang, G.; Wang, W.; Carin, L. Zero-shot recognition via optimal transport. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 3471–3481.
20. Han, Z.; Fu, Z.; Chen, S.; Yang, J. Contrastive Embedding for Generalized Zero-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2371–2381.
21. Zhang, Z.; Li, X.; Ma, T.; Gao, Z.; Li, C.; Lin, W. Residual-Prototype Generating Network for Generalized Zero-Shot Learning. *Mathematics* **2022**, *10*, 3587. [CrossRef]
22. Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; Wang, J. CausalVAE: Disentangled representation learning via neural structural causal models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9593–9602.
23. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.P.; Glorot, X.; Botvinick, M.M.; Mohamed, S.; Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.
24. Kim, H.; Mnih, A. Disentangling by factorising. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2649–2658.
25. Tong, B.; Wang, C.; Klinkigt, M.; Kobayashi, Y.; Nonaka, Y. Hierarchical disentanglement of discriminative latent features for zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 11467–11476.
26. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016; pp. 2180–2188.
27. Li, X.; Xu, Z.; Wei, K.; Deng, C. Generalized Zero-Shot Learning via Disentangled Representation. *Proc. Aaai Conf. Artif. Intell.* **2021**, *35*, 1966–1974. [CrossRef]
28. Sohn, K.; Lee, H.; Yan, X. Learning structured output representation using deep conditional generative models. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015; pp. 3483–3491. Available online: https://dl.acm.org/doi/10.5555/2969442.2969628 (accessed on 6 December 2022).
29. 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]
30. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-Ucsd Birds-200-2011 Dataset*; California Institute of Technology: Pasadena, CA, USA, 2011.
31. Farhadi, A.; Endres, I.; Hoiem, D.; Forsyth, D. Describing objects by their attributes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 1778–1785.
32. Reed, S.; Akata, Z.; Lee, H.; Schiele, B. Learning Deep Representations of Fine-Grained Visual Descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 49–58.
33. 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), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical evaluation of rectified activations in convolutional network. *arXiv* **2015**, arXiv:1505.00853.
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference for Learning Representations, San Diego, CA, USA, 7–9 May 2015.

36. Chen, Z.; Li, J.; Luo, Y.; Huang, Z.; Yang, Y. Canzsl: Cycle-consistent adversarial networks for zero-shot learning from natural language. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 874–883.
37. Li, J.; Jing, M.; Lu, K.; Ding, Z.; Zhu, L.; Huang, Z. Leveraging the invariant side of generative zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7402–7411.