2021 IEEE/CVF International Conference on Computer Vision (ICCV)

FREE: Feature Refinement for Generalized Zero-Shot Learning

Shiming Chen[1], Wenjie Wang[1], Beihao Xia[1], Qinmu Peng[1], Xinge You[1]*, Feng Zheng[2], Ling Shao[3]
[1]Huazhong University of Science and Technology (HUST), China
[2]Southern University of Science and Technology (SUSTech), China
[3]Inception Institute of Artificial Intelligence (IIAI), UAE
{shimingchen,pengqinmu,youxg}@hust.edu.cn zfeng02@gmail.com ling.shao@ieee.org

Abstract

Generalized zero-shot learning (GZSL) has achieved significant progress, with many efforts dedicated to overcoming the problems of visual-semantic domain gap and seen-unseen bias. However, most existing methods directly use feature extraction models trained on ImageNet alone, ignoring the cross-dataset bias between ImageNet and GZSL benchmarks. Such a bias inevitably results in poor-quality visual features for GZSL tasks, which potentially limits the recognition performance on both seen and unseen classes. In this paper, we propose a simple yet effective GZSL method, termed feature refinement for generalized zero-shot learning (FREE), to tackle the above problem. FREE employs a feature refinement (FR) module that incorporates semantic→visual mapping into a unified generative model to refine the visual features of seen and unseen class samples. Furthermore, we propose a self-adaptive margin center loss (SAMC-loss) that cooperates with a semantic cycle-consistency loss to guide FR to learn class- and semantically-relevant representations, and concatenate the features in FR to extract the fully refined features. Extensive experiments on five benchmark datasets demonstrate the significant performance gain of FREE over its baseline and current state-of-the-art methods. The code is available at https://github.com/shiming-chen/FREE .

1. Introduction

A key challenge of artificial intelligence is to generalize machine learning models from seen data to unseen scenarios. Zero-shot learning (ZSL) is a typical research topic targeting this goal [25, 27, 41]. ZSL aims to classify the images of unseen classes by constructing a mapping relationship between the semantic and visual domains. It is usually based on the assumption that both seen and unseen classes can be described through a set of semantic vectors, e.g., sentence embeddings [45], and attribute vectors [26], in the same semantic space. According to their classification range, ZSL methods can be categorized into conventional ZSL (CZSL) and generalized ZSL (GZSL) [58]. CZSL aims to predict unseen classes, while GZSL can predict both seen and unseen classes. Recently, GZSL has attracted more attention as it is more realistic and challenging. We are thus also interested in the GZSL setting in this paper.

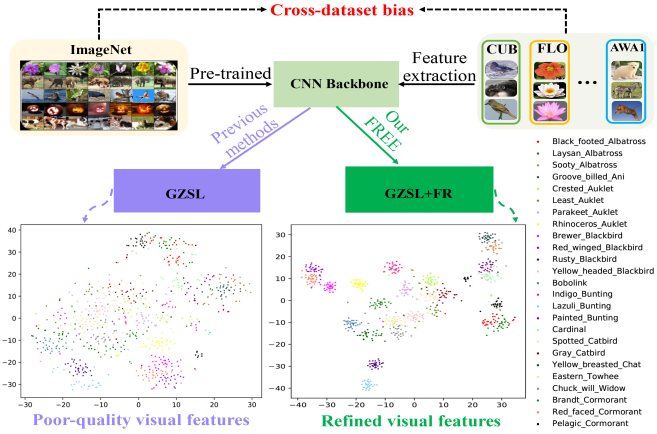


Figure 1. The core idea of our FREE. The cross-dataset bias between ImageNet and GZSL benchmarks (e.g., CUB) is harmful for feature extraction from GZSL benchmarks, which results in poor-quality visual features for unsatisfying performance in GZSL. Our FREE refines the visual features and improves the *semantic→visual* mapping using feature refinement (FR) in a unified network for GZSL classification.

GZSL has achieved significant progress, with many efforts focused on the problems of *visual-semantic domain gaps* [26, 1, 2, 52, 51, 61] and *seen-unseen bias* [57, 37, 66, 64, 49, 38, 36, 19]. Semantic embedding [33, 8, 31, 66, 34] or generative methods (e.g., variational autoencoders (VAEs) [3, 47], generative adversarial nets (GANs) [57, 31, 60, 65, 21, 51], and generative flows [49])

*Corresponding author

are typically applied to mitigate these challenges.

An important observation of ours is that the unsatisfying performance in GZSL that still nevertheless exists is closely related to the cross-dataset bias [50]. GZSL models usually extract visual features from coarse- and fine-grained benchmarks (e.g., AWA1 [26] and CUB [53]) using a convolutional neural network (CNN) backbone (e.g., ResNet-101 [16]) pre-trained on ImageNet [58]. However, cross-dataset bias, where the data collection procedure can be biased by human or systematic factors, can lead to a distribution mismatch between two datasets, e.g., Auklets are found in the CUB dataset but not in ImageNet. Thus, it is unwise to directly transfer knowledge from ImageNet to a new dataset for GZSL without any further sequential learning, because cross-dataset bias limits knowledge transfer and results in the extraction of poor-quality visual features from GZSL benchmarks, as shown in Fig. 1. Further, the larger the bias between ImageNet and the GZSL benchmark, the poorer the knowledge transfer and feature extraction. Since there is a more obvious bias for fine-grained datasets (e.g., CUB), these typically yield inferior performance to coarse-grained datasets (e.g., AWA) for all GZSL methods. The negative effect of cross-dataset bias on the performance of GZSL has been further validated experimentally. In [59], Xian fine-tuned a ResNet pre-trained on ImageNet using seen classes from GZSL benchmarks. Before fine-tuning, f-VAEGAN achieved a harmonic mean of 64.6% and 63.5% on FLO and AWA2, respectively, while these numbers increased to 75.1% and 65.2% afterward, as shown in Table 4. However, Xian did not analyze or discuss this phenomenon. Further, although fine-tuning may alleviate the cross-dataset bias to a certain degree, it inevitably results in other severer problems, e.g., overfitting [17, 28]. Thus, properly addressing the problem of cross-dataset bias in GZSL has become very necessary. To the best of our knowledge, we are the first to identify this as an open issue in GZSL, which will be tackled in this paper.

To address the above challenges, we propose a novel GZSL method, termed feature refinement for generalized zero-shot learning (FREE), to further boost the performance of GZSL. FREE refines visual features in a unified generative model, which simultaneously benefits *semantic→visual* learning, feature synthesis, and classification. Specifically, we take f-VAEGAN [59] as a baseline to learn a *semantic→visual* mapping. To improve the visual features of seen and unseen class samples, we employ a feature refinement (FR) module, which can be jointly optimized with f-VAEGAN to effectively avoid the drawbacks of fine-tuning. Since class label information is available, we introduce a self-adaptive margin center loss (SAMC-loss) to explicitly encourage intra-class compactness and inter-class separability that can adapt to different datasets, i.e., coarse-grained and fine-grained, and guide FR to learn discriminative class-relevant features. Thus, the distributions of different classes can be easily separated, as shown in Fig. 1. To better learn semantically-relevant and more discriminative visual features, a semantic cycle-consistency loss is also added after the restitution of features. From the residual information [16], we further concatenate the discriminative features of various layers in FR to extract the fully refined features.

To summarize, this paper provides the following important contributions: **(1)** We propose a novel GZSL method, termed feature refinement for generalized zero-shot learning (FREE), to address the problem of cross-dataset bias, which can further boost the performance of GZSL. To achieve this goal, a feature refinement (FR) module that cooperates with *semantic→visual* mapping in a unified framework is explored. Importantly, the two modules can be jointly optimized. **(2)** We propose a self-adaptive margin center loss (SAMC-loss) to explicitly encourage intra-class compactness and inter-class separability. The SAMC-loss also cooperates with a semantic cycle-consistency constraint to enable FR to learn more discriminative class- and semantically-relevant representations, which are especially important for GZSL. **(3)** Extensive experimental results on five benchmarks, i.e., CUB, SUN, FLO, AWA1, and AWA2, clearly demonstrate the advantages of the proposed FREE over its baseline and current state-of-the-art methods.

2. Related Work

Visual-Semantic Domain Gap. The required knowledge transfer from seen to unseen classes for GZSL relies on semantic embedding. One key task is to bridge the visual and semantic domains [26, 1, 2, 52, 11, 51]. Since visual features in various forms may convey the same concept, the distribution of instances in the visual space is often distinct from that of their underlying semantics in the semantic space. Thus, there is typically a gap between the two domains, known as the visual-semantic domain gap problem of GZSL [52]. Common space learning [7, 67, 56, 52, 15], model parameter transfer [6, 13, 20] and direct mapping [1, 2, 11, 57, 59, 38, 18, 32, 8, 12] are often used to bridge this gap. Early mapping works [1, 2] were carried out by either a classifier or a regression model, depending on the semantic representation adopted. Most recent mapping works, in contrast, are based on generative models (e.g., VAEs [3, 47, 34], GANs [57, 31, 29, 43, 46, 65], and generative flows [49]). These models not only learn visual-semantic mapping, but also generate a great number of feature samples of unseen classes for data augmentation.

Seen-Unseen Bias. GZSL methods inevitably encounter a seen-unseen bias problem [47, 59, 49, 49, 33, 66, 10, 55, 20, 30, 4, 63, 62]. As only seen data is involved during training, most generative GZSL models tend to overfit to seen classes [33, 9, 36], i.e., the unseen data generated tends to have the same distribution as the seen categories. This heavily ham-

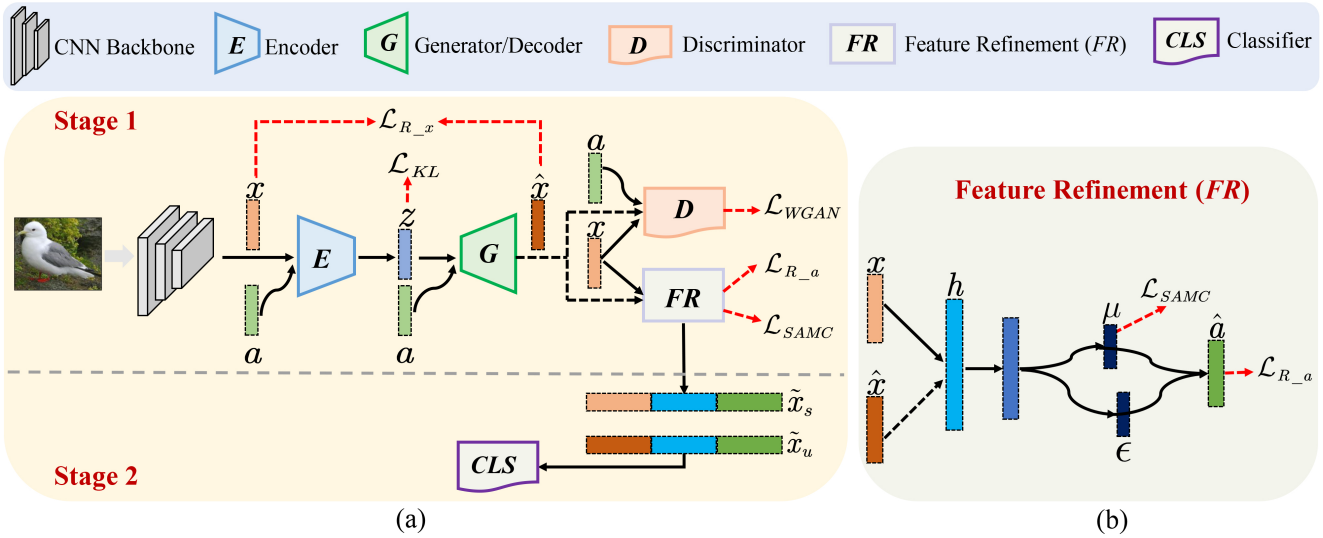


Figure 2. A schematic overview of FREE. (a) FREE consists of a feature generating VAEGAN (f-VAEGAN), a feature refinement (FR) module, and a classifier. In Stage 1, f-VAEGAN aims to learn the *semantic→visual* mapping (G) for visual feature generation, while FR tries to learn discriminative representations for the real/synthesized seen visual features. The two are jointly optimized. In Stage 2, we refine the real seen visual features and the real/synthesized unseen visual features with the trained FR, and they are then fed into a classifier for classification. (b) Proposed FR module. FR learns discriminative features utilizing the SAMC-loss and semantic cycle-consistency loss, and its discriminative features in various layers are then concatenated to extract the fully refined features.

pers the classification performance of unseen classes. In [33], Liu proposed the Deep Calibration Network (DCN) to enable simultaneous calibration of deep networks on the confidence of source classes and the uncertainty of target classes. Zhang *et al.* [66] employed a co-representation network to learn a more uniform visual embedding space, effectively bypassing the bias problems and improving classification. Maximum mean discrepancy (MMD) based methods optimize the distribution between real seen features and synthesized unseen features to tackle the bias problem explicitly [5, 49].

Cross-Dataset Bias. While datasets are expected to resemble the probability distribution of the real world, the data collection procedure can be biased by human and systematic factors, leading to a distribution mismatch between two datasets. Thus, the knowledge transfer from one dataset (e.g., ImageNet) to other new datasets (e.g., GZSL benchmarks) is limited, which results in the extraction of poor-quality visual features on the new datasets. This is known as cross-dataset bias [50]. Domain adaptation techniques can be used to reduce this bias [22, 40], but they are not applicable to GZSL, as the features of GZSL benchmarks are extracted from a pre-trained CNN backbone before sequential learning is conducted. Additionally, although fine-tuning is a typical method to alleviate this problem, it inevitably results in other severer issues, including inefficiency and overfitting [17, 28]. Thus, we attempt to essentially circumvent cross-dataset bias and improve *semantic→visual* mapping in a unified network using feature refinement for GZSL classification.

3. Method

Motivation. As shown in Fig. 1, the cross-dataset bias limits its knowledge transfer from ImageNet to GZSL benchmarks, resulting in poor-quality visual features being extracted from the GZSL benchmarks (e.g., CUB [53]) by a pre-trained CNN backbone. This hampers the *semantic→visual* learning, feature synthesis, and GZSL classification. As a result, the upper bound of the recognition performance of GZSL on both seen and unseen classes is potentially limited. Although fine-tuning may alleviate this issue to a certain degree, it inevitably results in other severer problems [17, 44, 28]. For example, it is difficult to fine-tune on a new small dataset, and it is easy for the model to overfit to seen classes, which ultimately is not conducive to the generalization of GZSL.

These observations prompt us to speculate that the current poor performance of GZSL is closely related to the cross-dataset bias. The experimental results of fine-tuning in [59, 38] further support this claim. In other words, we believe that, by alleviating the cross-dataset bias, the visual features of GZSL benchmarks will be enhanced, enabling us to also further improve the GZSL classification. To this end, we propose a novel method, termed feature refinement for generalized zero-shot learning (FREE). Our strategy is to utilize class label supervision and a semantic cycle-consistency constraint to guide the proposed feature refinement (FR) module to learn class- and semantically-relevant feature representations in a unified network. FR can effectively refine

the visual features and avoids the inefficiency and overfitting risks of fine-tuning.

Overview. The pipeline of FREE is shown in Fig. 2. FREE includes a feature generating VAEGAN (f-VAEGAN) [59], a feature refinement module (FR), and a classifier. In Stage 1, f-VAEGAN aims to learn the *semantic→visual* mapping (G) for visual feature synthesis, while FR tries to learn discriminative representations for the real/synthesized seen visual features. The two are jointly optimized. To learn discriminative features, FR is optimized using our SAMC-loss and a semantic cycle-consistency loss. In Stage 2, we refine the visual features of both seen and unseen classes using the trained FR. The refined real seen visual features and refined synthesized unseen visual features are then fed into a classifier (e.g., softmax) for training, while the refined real unseen visual features are used for testing. Thus, FREE is an *inductive* method.

Notation. We denote seen data as $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{M}$, where x_i is a visual feature, y_i is its class label in $\mathcal{Y}^s$, and M is the number of seen images. Let $\mathcal{Y}^u$ be the set of unseen classes, which is disjoint from the seen class set $\mathcal{Y}^s$, i.e., $\mathcal{Y}^s \cap \mathcal{Y}^u = \varnothing$. Each seen class and unseen class have their own corresponding attribute embedding $a_j \in \mathcal{A}, \forall j \in \mathcal{Y}^s \cup \mathcal{Y}^u$.

3.1. Revisiting f-VAEGAN

f-VAEGAN [59] has achieved impressive performance and thus become a popular baseline for generative GZSL methods. In this paper, we take f-VAEGAN as a baseline for learning *semantic→visual* mapping. f-VAEGAN integrates a VAE [24] and GAN [14] into a unified generative model to simultaneously take advantage of both. Specifically, it comprises a feature generating VAE (f-VAE) and a feature generating WGAN (f-WGAN). The f-VAE consists of an encoder $E(x, a)$ (denoted as E) and a decoder G (shared with the f-WGAN, as a conditional generator $G(z, a)$). The encoder $E(x, a)$ encodes an input seen visual feature x to a latent code z, while the decoder $G(z, a)$ reconstructs visual feature $\hat{x}$ from z. the f-VAE is first optimized by a VAE loss $\mathcal{L}_V$:

$$\begin{aligned}\mathcal{L}_V &= \mathcal{L}_{KL} + \mathcal{L}_{R\_x} \\ &= \mathrm{KL}(E(x,a)\|p(z \mid a)) - \mathbb{E}_{E(x,a)}[\log G(z,a)],\end{aligned} \tag{1}$$

where $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence, $p(z \mid a)$ is a prior distribution assumed to be $\mathcal{N}(0, 1)$, and $\mathcal{L}_{R\_x}$ is the visual feature reconstruction loss represented by $-\log G(z, a)$. The f-WGAN, on the other hand, comprises a generator $G(z, a)$ and a discriminator $D(x, a)$ (denoted as D). The generator $G(z, a)$ synthesizes a visual feature $\hat{x}$ from a random input noise z, whereas the discriminator $D(x, a)$ takes a real visual feature x or a synthesized visual feature $\hat{x}$ and outputs a real value indicating the degree of realness or fakeness. Both G and D are conditioned on the embedding a, optimized by the WGAN loss

$$\begin{aligned}\mathcal{L}_W = &\mathbb{E}[D(x,a)] - \mathbb{E}[D(\hat{x},a)] \\ &- \lambda \mathbb{E}\left[\left(\|\nabla D(x',a)\|_2 - 1\right)^2\right],\end{aligned} \tag{2}$$

where $x' = \tau x + (1-\tau)\hat{x}$ with $\tau \sim U(0, 1)$ and λ is the penalty coefficient.

3.2. Feature Refinement

Our strategy towards circumventing the cross-dataset bias is to refine the visual features of GZSL benchmarks for remedying the limited knowledge transfer with an FR module, which is constrained by the SAMC-loss and semantic cycle-consistency loss. Furthermore, we concatenate the features of various layers in FR to extract the fully refined features for classification.

Self-Adaptive Margin Center Loss. To encourage FR to learn more discriminative class-relevant representations for visual features, we propose the self-adaptive margin center loss (SAMC-loss, $\mathcal{L}_{SAMC}$) to constraint FR. There are four reasons of why and how we conduct SAMC-loss: (1) Since class label information is available, we introduce $\mathcal{L}_{SAMC}$ to explicitly encourage intra-class compactness and inter-class separability, and guide FR to learn discriminative class-relevant features. (2) $\mathcal{L}_{SAMC}$ has the advantages of the center loss [54] and triplet loss [48], e.g., avoid artificial sampling, and simultaneously learn intra-class compactness and inter-class separability. (3) Considering the intra-class compactness and inter-class separability are differently sensitive to various datasets, i.e., coarse-grained and fine-grained datasets, $\mathcal{L}_{SAMC}$ takes a balance factor (γ) to balance the inter-class separability and intra-class compactness adaptively. (4) $\mathcal{L}_{SAMC}$ is conducted on the intermediate encoded features μ in FR, which is directly beneficial for improving the discriminability of features of the shallower layers in FR. Furthermore, the class centers in $\mathcal{L}_{SAMC}$ are dynamically updated during training, enabling the discriminative feature learning to be more effective. $\mathcal{L}_{SAMC}$ is formulated as:

$$\begin{aligned}\mathcal{L}_{SAMC}(\hat{a}, y, y') = \max\Big(0, \Delta + \gamma \left\|\mu - \mathbf{c}_y\right\|_2^2 \\ -(1-\gamma)\left\|\mu - \mathbf{c}_{y'}\right\|_2^2\Big),\end{aligned} \tag{3}$$

where $\mathbf{c}_y$ is the yth (the label of seen visual feature x) class center of semantic embedding, $\mathbf{c}_{y'}$ is the y'th (a randomly selected class label other than y) class center, Δ represents the margin that controls the distance between intra- and inter-class pairs, μ is the encoded feature in FR, and $\gamma \in [0, 1]$ is used for balancing the inter-class separability and intra-class compactness, which are adaptable to various datasets.

We use a large γ for fine-grained datasets (e.g., CUB, SUN), and a small γ for coarse-grained datasets (e.g., AWA1 [26], AWA2 [58]). This is done because (1) when the classes

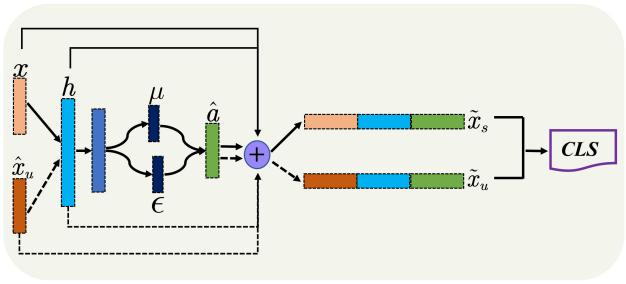


Figure 3. Extracting fully refined features in FR. Best viewed in color.

are ambiguous in a fine-grained dataset, we can more easily distinguish them by encouraging intra-class compactness, as shown in Fig. 7(a); (2) when the classes are confused in a coarse-grained dataset, we can effectively seperate them by enlarging the inter-class separability, as shown in Fig. 7(b). Note that our SAMC-loss is different from the max-margin loss [1, 2, 45, 12]. Our SAMC-loss enables the model to learn visual-semantic embedding and refined features simultaneously. In contrast, the max-margin loss is typically employed to learn a compatibility function [1, 2, 45] or a scoring function [12] between images and textual side-information to represent visual-semantic label embedding in conventional ZSL, resulting in limited performance caused by the cross-dataset bias.

Semantic Cycle-Consistency Loss. The last layer of FR reconstructs the semantic embedding $\hat{a}$ from $\hat{x}$ or x using the reparametrization trick [24]. To further guide FR to effectively learn semantically-relevant representations, we apply a semantic cycle-consistency loss ($\mathcal{L}_{R\_a}$) [11, 38] to the reconstructed semantic embeddings to ensure that the synthesized semantic vectors $\hat{a}$ are transformed to be the same embeddings that generated them. To this end, semantically-relevant features are learned using FR. The semantic cycle-consistency loss $\mathcal{L}_{R\_a}$ is achieved using the ℓ_1 reconstruction loss, formulated as follows:

$$\mathcal{L}_{R\_a} = \mathbb{E}\left[\|\hat{a}_{real} - a\|_1\right] + \mathbb{E}\left[\|\hat{a}_{syn} - a\|_1\right], \quad (4)$$

where $\hat{a}_{real}$ are the semantically-relevant features synthesized from x using FR, and $\hat{a}_{syn}$ are those synthesized from $\hat{x}$. Note that $\hat{a} = \hat{a}_{real} \cup \hat{a}_{syn}$, and a is the semantic embedding corresponding to visual features x or $\hat{x}$.

Extracting Fully Refined Features. After training in Stage 1, we extract the fully refined features $\tilde{x}_s$ and $\tilde{x}_u$ from FR to refine real seen visual features x and real/synthesized unseen visual features $x_u/\hat{x}_u$ into discriminative features. FR transforms the original high-dimensional features into low-dimensional features, inevitably discarding some discriminative information, which may hamper the GZSL classification performance. Using the residual information [16], we concatenate the visual features x and $\hat{x}$ with the corresponding latent embedding $h_s, h_u \in \mathcal{H}$ and semantically-relevant embedding $\hat{a}_s, \hat{a}_u \in \mathcal{A}$ learned from FR as fully refined features, as shown in Fig. 3. They are formulated as:

$$\tilde{x}_s = x \oplus h_s \oplus \hat{a}_s \quad (5)$$

$$\tilde{x}_u = \hat{x}_u \oplus h_u \oplus \hat{a}_u \quad (6)$$

where $\oplus$ is a *concatentation* operation, and $\tilde{x}_s$ and $\tilde{x}_u \in \tilde{\mathcal{X}}$. Thus, the visual features are fully refined as discriminative features, which are class- and semantically-relevant for reducing ambiguities among feature instances of different categories. Note that the refined real seen features and refined synthetic unseen features are used to train the classifier, while the refined real unseen features are used for testing.

3.3. Optimization

We jointly train the encoder (E), generator (G), discriminator (D), and feature refinement (FR) to optimize the overall objective, which involves a weighted sum of the following losses:

$$\mathcal{L}_{total}(E, G, D, FR) = \mathcal{L}_V + \mathcal{L}_W + \lambda_{SAMC}\mathcal{L}_{SAMC} + \lambda_{R\_a}\mathcal{L}_{R\_a}, \quad (7)$$

where λ_{SAMC} and λ_{R_a} are weights that control the importance of the related loss terms. Similar to the alternative updating policy for GANs, we alternately train E, G before the generated visual features and E, D and FR after the generated visual features. This is a joint learning framework that couples *semantic→visual* mapping and visual feature refinement in a unified network. There is also an online interaction between them that mutually benefits both tasks for GZSL. However, fine-tuning is limited in this.

3.4. Classification

After obtaining the training data, we train a supervised classifier in the refined feature space as the final GZSL classifier. GZSL aims to learn the classifier $f_{gzsl} : \tilde{\mathcal{X}} \rightarrow \mathcal{Y}^s \cup \mathcal{Y}^u$. During testing, the seen/unseen test features are refined as new features by FR, and then used for further testing.

4. Experiments

Datasets. We evaluate our method on five benchmark datasets, i.e., CUB (Caltech UCSD Birds 200) [53], SUN (SUN Attribute) [42], FLO (Oxford Flowers) [39], AWA1 (Animals with Attributes 1) [26], and AWA2 (Animals with Attributes 2)[58]. Among these, CUB, SUN and FLO are fine-grained datasets, whereas AWA1 and AWA2 are coarse-grained. We use the same seen/unseen splits and class embeddings, as [58], which are summarized in Table 2.

Evaluation Protocols. During testing, we use the unified evaluation protocol proposed from [58] to facilitate direct

Table 1. State-of-the-art comparison on five datasets. The best and second-best results are marked in **red** and **blue**, respectively.

	Method	AWA1			AWA2			CUB			SUN			FLO		
		U	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*
Non-generative	DCN [33]	25.5	**84.2**	39.1	–	–	–	28.4	60.7	38.7	25.5	37.0	30.2	–	–	–
	SP-AEN [8]	23.3	**90.9**	37.1	–	–	–	34.7	70.6	46.6	24.9	38.6	30.3	–	–	–
	AREN [60]	–	–	–	15.6	**92.9**	26.7	38.9	**78.7**	52.1	19.0	**38.8**	25.5	–	–	–
	CRnet [66]	58.1	74.7	65.4	–	–	–	45.5	56.8	50.5	34.1	36.5	35.3	–	–	–
	GAFE [34]	25.5	76.6	38.2	26.8	78.3	40.0	22.5	52.1	31.4	19.6	31.9	24.3	–	–	–
	PQZSL [30]	–	–	–	31.7	70.9	43.8	43.2	51.4	46.9	35.1	35.3	35.2	–	–	–
	MLSE [10]	–	–	–	23.8	83.2	37.0	22.3	**71.6**	34.0	20.7	36.4	26.4	–	–	–
	TCN [20]	49.4	76.5	60.0	**61.2**	65.8	63.4	52.6	52.0	52.3	31.2	37.3	34.0	–	–	–
	DVBE [36]	–	–	–	**63.6**	70.8	**67.0**	53.2	60.2	**56.5**	45.0	37.2	40.7	–	–	–
Generative	SE-GZSL [3]	56.3	67.8	61.5	58.3	68.1	62.8	41.5	53.3	46.7	40.9	30.5	34.9	–	–	–
	f-CLSWGAN [57]	57.9	61.4	59.6	–	–	–	43.7	57.7	49.7	42.6	36.6	39.4	59.0	73.8	65.6
	cycle-CLSWGAN [11]	56.9	64.0	60.2	–	–	–	45.7	61.0	52.3	**49.4**	33.6	40.0	59.2	72.5	65.1
	CADA-VAE [47]	57.3	72.8	64.1	55.8	75.0	63.9	51.6	53.5	52.4	47.2	35.7	40.6	–	–	–
	f-VAEGAN [59]	–	–	–	57.6	70.6	63.5	48.4	60.1	53.6	45.1	38.0	41.3	56.8	74.9	64.6
	LisGAN [29]	52.6	76.3	62.3	–	–	–	46.5	57.9	51.6	42.9	37.8	40.2	57.7	**83.8**	68.3
	GMN [46]	**61.1**	71.3	**65.8**	–	–	–	**56.1**	54.3	55.2	**53.2**	33.0	40.7	–	–	–
	E-PGN [65]	–	–	–	52.6	**83.5**	64.6	52.0	61.1	56.2	–	–	–	**71.5**	82.2	**76.5**
	OCD-CVAE [21]	–	–	–	59.5	73.4	65.7	44.8	59.9	51.3	44.8	**42.9**	**43.8**	–	–	–
	LsrGAN [51]	54.6	74.6	63.0	–	–	–	48.1	59.1	53.0	44.8	37.7	40.9	–	–	–
	FREE (Ours)	**62.9**	69.4	**66.0**	60.4	75.4	**67.1**	**55.7**	59.9	**57.7**	47.4	37.2	**41.7**	**67.4**	**84.5**	**75.0**

Table 2. Statistics of the CUB, SUN, FLO, AWA1 and AWA2 datasets, including the dimensions of semantic vectors per class (*Att*), seen/unseen class size (*Seen/Unseen*), and total number of images (*Img*).

	Att	*Seen/Unseen*	*Img*
CUB	312	150/50	11788
SUN	102	645/72	14340
FLO	1024	82/20	8189
AWA1	85	40/10	30475
AWA2	85	40/10	37322

comparison. Since the test set is composed of seen classes ($\mathcal{Y}^s$) and unseen classes ($\mathcal{Y}^u$), we evaluate the top-1 accuracies on both, denoted as S and U, respectively. Furthermore, their harmonic mean (defined as $H = (2 \times S \times U)/(S+U)$) is also used for evaluating the performance of GZSL.

Implementation Details. In FREE, the encoder, generator and discriminator are multilayer perceptrons (MLPs) containing a 4096-unit hidden layer with LeakyReLU activation. FR is also an MLP that has two hidden layers with 4096-unit and $2 \times |\hat{a}|$-unit activated by LeakyReLU, followed an encoding layer where the second hidden layer is encoded into two feature vectors with the size of $|\hat{a}|$. Its output layer $\hat{a}$ is learned by the reparametrization trick [24] and corresponds to the semantic vector of the datasets (e.g., $|\hat{a}| = 312$ for CUB). We use the Adam optimizer [23] with β_1 = 0.5, β_2 = 0.999. The visual features are extracted from the 2048-dimimentional top-layer pooling units of a ResNet-101 pretrained on ImageNet, following [57]. The penalty coefficient λ is set to 10. We empirically set the loss weights λ_{SAMC} and λ_{R_a} to 0.5 and 0.1/0.001 respectively for SUN/others datasets. The balance factor γ is set to 0.8 and 0.1 for fine- and coarse-grained datasets, respectively.

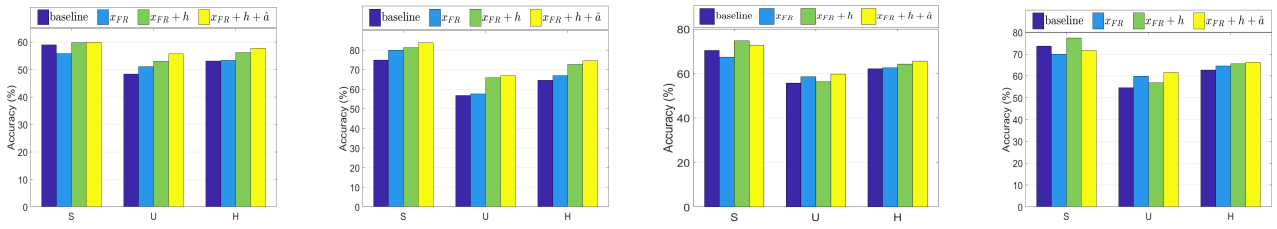
4.1. Comparison with State of the Arts

Since FREE is an inductive method, we compare it with other state-of-the-art inductive models for a fair comparison. We categorize the compared methods into generative and non-generative methods.

Table 1 shows the top-1 accuracies of different methods on unseen classes (U), seen classes (S) and their harmonic mean (H). The results show that FREE consistently attains the best performance for harmonic mean on three benchmarks, i.e., 66.0 on AWA1, 67.1 on AWA2, and 57.7 on CUB. Meanwhile, FREE achieves the second-best results for harmonic mean with 41.7 and 75.0 on SUN and FLO, respectively. These results indicate that the refined features are discriminative and generic for seen/unseen classes on both coarse- and fine-grained datasets. Notably, unlike the compared state-of-the-art methods which only achieve good performance on either seen or unseen classes, FREE attains promising results on both. This reveals that FREE maintains a good balance between seen and unseen classes, benefitting from the unified model jointly trained for *semantic→visual* mapping and FR. Specifically, the joint training enables the two modules to encode complementary information of categories and encourages them to learn discriminative representations by avoiding cross-dataset bias.

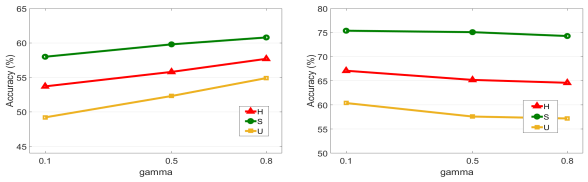
4.2. Ablation Study

To provide further insight into FREE, we conduct ablation studies to evaluate the effect of different model built and



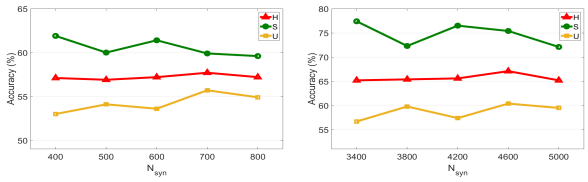
(a) CUB (b) FLO (c) AWA1 (d) AWA2

Figure 4. The effectiveness of various visual feature components refined by FR. Best viewed in color.



(a) CUB (b) AWA2

Figure 5. The effectiveness of the balance factor γ for the SAMC-loss.



(a) CUB (b) AWA2

Figure 6. The impact of the number of synthetic visual features N_{syn} in each unseen class.

Table 3. Ablation studies for different components of FREE on CUB and AWA2. The best results are marked in **boldface**.

Method	CUB			FLO			AWA2		
	U	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*
f-VAEGAN [59]	48.4	60.1	53.6	56.8	74.9	64.6	57.6	70.6	63.5
baseline	48.3	58.9	53.1	57.2	75.1	64.9	54.6	73.6	62.7
baseline+FR($\mathcal{L}_{R\_a}$)	50.4	**62.2**	55.7	63.2	73.8	68.1	57.1	73.8	64.4
baseline+FR($\mathcal{L}_{SAMC}$)	53.2	60.2	56.5	64.0	82.4	72.0	59.1	72.7	65.2
baseline+FR($\mathcal{L}_{SAMC}$+$\mathcal{L}_{R\_a}$)	**54.9**	60.8	**57.7**	**67.4**	**84.5**	**75.0**	**60.4**	**75.4**	**67.1**

feature components. Since FREE is based on f-VAEGAN [59], we re-implement this method as our *baseline*.

Analysis of Model Components. As shown in Table 3, our FR provides significant improvements over the baseline with various model components (i.e., FR($\mathcal{L}_{R\_a}$), FR($\mathcal{L}_{SAMC}$) and both combined). We first evaluate the two components independently. FR($\mathcal{L}_{SAMC}$) and FR($\mathcal{L}_{R\_a}$) individually outperform the baseline in harmonic mean on CUB (by 3.4% and 2.6%), FLO (by 7.1% and 3.2%) and AWA2 (by 2.5% and 1.7%). This shows the effectiveness of FR. Interestingly, since the cross-dataset bias on fine-grained datasets (e.g., CUB, FLO) is larger than on coarse-grained datasets (e.g., AWA2), our FR can achieve greater improvements on fine-grained datasets than on coarse-grained datasets. Furthermore, FR($\mathcal{L}_{SAMC}$) performs better than FR($\mathcal{L}_{R\_a}$), which shows the effectiveness of our SAMC-loss. The complete version of FREE gives the highest results on all datasets, achieving an impressive accuracy gain of 4.6%, 10.1% and 4.4% in harmonic mean on CUB, FLO and AWA2, respectively. This indicates that the SAMC-loss and semantic cycle-consistency loss are mutually complementary for FR. These results prove that FR is good for feature enhancement in GZSL, and thus the cross-dataset bias is alleviated.

Analysis of Feature Components. We study the effectiveness of various visual feature components refined by FR. As shown in Fig. 4, all feature components in FR substantially improve the performance. When only taking the real seen features x and the synthesized unseen visual features $\hat{x}_u$ for classification, labeled as $\boldsymbol{x_{FR}}$, FREE provides consistent improvement of harmonic mean over the baseline. This proves that FR contributes positively to the *semantic→visual* mapping. We also concatenate seen/unseen visual features with the hidden features h (labeled as $\boldsymbol{x_{FR}+h}$), and both h and the learned semantically-relevant features $\hat{a}$ (labeled as $\boldsymbol{x_{FR}+h+\hat{a}}$) for classification. Further improvement is achieved. Since there is a more obvious cross-dataset bias for fine-grained datasets (e.g., FLO), FR consistently achieves improvements on all evaluation protocols with various visual feature components. As for coarse-grained datasets (e.g., AWA2) with small cross-dataset bias, FR also improves the performances of unseen accuracy and Harmonic mean. This is attributed to the fact that FR can reduce the cross-dataset bias for improving the quality of visual features.

4.3. Hyperparameter Analysis

Balance Factor γ. We study the balance factor γ in Eq. 3 to determine its influence on the module. As can be seen from Fig. 5, as γ grows, S, U and H gain consistent improvement on the fine-grained datasets (e.g., CUB). Nevertheless, S, U and H consistently decrease when γ increases on coarse-grained datasets (e.g., AWA2). These results are explained as follows: (1) On the fine-grained datasets, increasing the intra-class compactness provides larger gains when the classes are

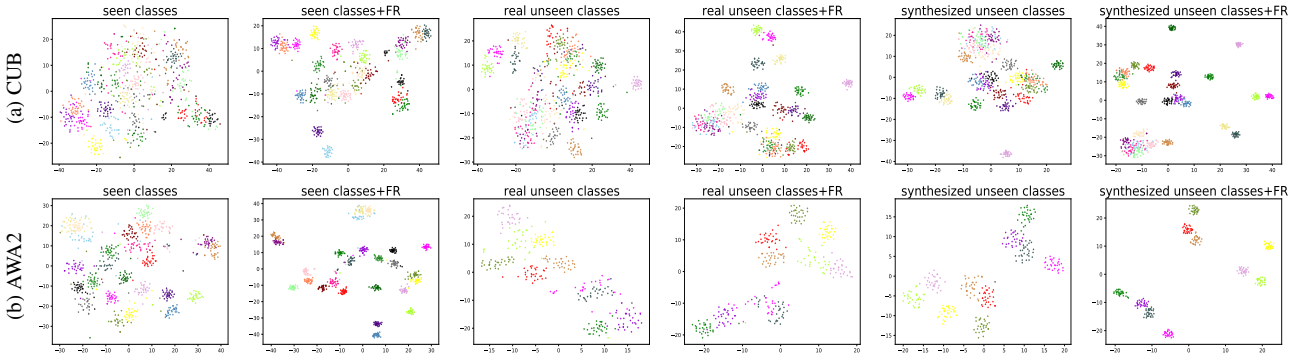


Figure 7. t-SNE visualization [35] of visual features for real seen classes, real/synthesized unseen classes, and their features refined by FR on (a) CUB and (b) AWA2. The real/synthesized unseen class features are represented with the same color. Best viewed in color.

confused. (2) On the coarse-grained datasets, increasing the inter-class separability significantly benefits the classification of ambiguous classes.

Number of Synthesized Visual Features N_{syn}. We evaluate the impact of the number of synthetic visual features per unseen class N_{syn}. As shown in Fig. 6, FREE is generally insensitive to N_{syn} on all datasets. When increasing the number of synthetic features, the seen class accuracy drops slightly and the unseen class accuracy improves. This demonstrates that FREE can also alleviate the seen-unseen bias problem. Since there exists an upper bound on synthetic diversity, all results will decrease if N_{syn} is set to too large. As such, we set N_{syn} to 4600, 4600, 2400, 700 and 300 for AWA1, AWA2, FLO, CUB and SUN, respectively.

5. Discussion

FR for Cross-Dataset Bias. As displayed in Fig. 1, we intuitively show that the cross-dataset bias results in poor-quality visual features, which potentially limits recognition performance for both seen and unseen classes of GZSL. The experimental results of fine-tuning in [59, 38] well support this claim. In this paper, we attempt to enhance the visual features of seen/unseen classes to address the limited knowledge transfer caused by this bias, using feature refinement (FR), which is encouraged to learn class- and semantically-relevant feature representations. As shown in Fig. 7, FR significantly improves the visual features of seen/unseen classes, reducing the ambiguity between different categories. These results intuitively show the effectiveness of FR. Interestingly, the synthesized unseen features share similar class relationships with the real unseen features, which proves that FR helps FREE to learn a promising *semantic→visual* mapping. As a result, FREE achieves an impressive performance gain over current state-of-the-art methods and its baseline.

Feature Refinement vs Fine-tuning. As analyzed in Section 1, although fine-tuning may alleviate the cross-dataset bias for GZSL to a degree, it leads to ineffectiveness and overfitting (e.g., seen-unseen bias is 29.1% on FLO shown in Table 4). Nevertheless, our proposed FR combines itself with *semantic→visual* mapping, the two of which are mutually beneficial. Meanwhile, the SAMC-loss and semantic cycle-consistency loss guide FR to learn discriminative feature representations to refine the visual features, and thus the limited knowledge transfer is alleviated and the cross-dataset bias is reduced. As shown in Table 4, FR achieves competitive results over fine-tuning on FLO and AWA2, which well support our claims.

Table 4. Feature refinement vs fine-tuning on FLO and AWA2. The best results are marked in **boldface**.

Method	FLO			AWA2		
	U	*S*	*H*	*U*	*S*	*H*
f-VAEGAN [59]	56.8	74.9	64.6	57.6	70.1	63.5
f-VAEGAN+finetuned [59]	63.3	**92.4**	**75.1**	57.1	**76.1**	65.2
f-VAEGAN+FR	**67.4**	84.5	75.0	**60.4**	75.4	**67.1**

6. Conclusion

In this paper, we propose a joint learning framework, termed FREE, that couples *semantic→visual* mapping and FR to alleviate the cross-dataset bias. We further introduce a SAMC-loss that cooperates with the semantic cycle-consistency constraint to encourage FR to learn class- and semantically-relevant feature representations. Meanwhile, we extract the features of various layers in FR as fully refined features for classification. Competitive results on five popular benchmarks demonstrate the superiority and great potential of our approach. Since cross-dataset bias is a major bottleneck for improving the GZSL performance and FR is an effective approach to address it, we believe FR may work well in other generative GZSL methods.

Acknowledgements

This work is partially supported by NSFC (61772220) and Key R&D Plan of Hubei Province (2020BAB027).

References

[1] Zeynep Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1425–1438, 2016. 1, 2, 5

[2] Zeynep Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015. 1, 2, 5

[3] Gundeep Arora, V. Verma, Ashish Mishra, and P. Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, pages 4281–4289, 2018. 1, 2, 6

[4] Y. Atzmon and Gal Chechik. Adaptive confidence smoothing for generalized zero-shot learning. In *CVPR*, pages 11663–11672. 2

[5] M. Bucher, S. Herbin, and F. Jurie. Generating visual representations for zero-shot classification. In *ICCV Workshop*, pages 2666–2673, 2017. 3

[6] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016. 2

[7] Soravit Changpinyo, Wei-Lun Chao, and F. Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *ICCV*, pages 3496–3505, 2017. 2

[8] Long Chen, Hanwang Zhang, Jun Xiao, W. Liu, and S. Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR*, pages 1043–1052, 2018. 1, 2, 6

[9] X. Chen, X. Lan, Fu-Chun Sun, and N. Zheng. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *ECCV*, 2020. 2

[10] Z. Ding and Hongfu Liu. Marginalized latent semantic encoder for zero-shot learning. In *CVPR*, pages 6184–6192, 2019. 2, 6

[11] Rafael Felix, B. V. Kumar, I. Reid, and G. Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, 2018. 2, 5, 6

[12] Andrea Frome, G. S. Corrado, Jonathon Shlens, S. Bengio, J. Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013. 2, 5

[13] Chuang Gan, Ming Lin, Y. Yang, Y. Zhuang, and A. Hauptmann. Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In *AAAI*, 2015. 2

[14] Ian J. Goodfellow, Jean Pouget-Abadie, M. Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 4

[15] Zongyan Han, Zhenyong Fu, and Jian Yang. Learning the redundancy-free features for generalized zero-shot object recognition. In *CVPR*, pages 12862–12871, 2020. 2

[16] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 5

[17] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *ICML*, 2019. 2, 3

[18] H. Huang, C. Wang, Philip S. Yu, and Chang-Dong Wang. Generative dual adversarial network for generalized zero-shot learning. In *CVPR*, pages 801–810, 2019. 2

[19] D. Huynh and E. Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, pages 4482–4492, 2020. 1

[20] Huajie Jiang, R. Wang, S. Shan, and X. Chen. Transferable contrastive network for generalized zero-shot learning. In *ICCV*, pages 9764–9773, 2019. 2, 6

[21] Rohit Keshari, R. Singh, and Mayank Vatsa. Generalized zero-shot learning via over-complete distribution. In *CVPR*, pages 13297–13305, 2020. 1, 6

[22] A. Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *ECCV*, 2012. 3

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[24] Diederik P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 4, 5, 6

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

[26] Christoph H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:453–465, 2014. 1, 2, 5

[27] H. Larochelle, D. Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, pages 646–651, 2008. 1

[28] Hao Li, Pratik Chaudhari, Hao Yang, Michael Lam, A. Ravichandran, Rahul Bhotika, and Stefano Soatto. Rethinking the hyperparameters for fine-tuning. In *ICLR*, 2020. 2, 3

[29] J. Li, Mengmeng Jing, K. Lu, Z. Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *CVPR*, pages 7394–7403, 2019. 2, 6

[30] J. Li, X. Lan, Y. Liu, L. Wang, and N. Zheng. Compressing unknown images with product quantizer for efficient zero-shot classification. In *CVPR*, pages 5458–5467, 2019. 2, 6

[31] K. Li, Martin Renqiang Min, and Yun Fu. Rethinking zero-shot learning: A conditional visual classification perspective. In *ICCV*, pages 3582–3591, 2019. 1, 2

[32] Y. Li, D. Wang, Huanhang Hu, Yuetan Lin, and Yueting Zhuang. Zero-shot recognition using dual visual-semantic mapping paths. In *CVPR*, pages 5207–5215, 2017. 2

[33] Shichen Liu, Mingsheng Long, J. Wang, and Michael I. Jordan. Generalized zero-shot learning with deep calibration network. In *NeurIPS*, 2018. 1, 2, 3, 6

[34] Yang Liu, D. Xie, Quanxue Gao, Jungong Han, S. Wang, and X. Gao. Graph and autoencoder based feature extraction for zero-shot learning. In *IJCAI*, 2019. 1, 2, 6

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

[36] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Z. Zha, and Yongdong Zhang. Domain-aware visual bias eliminating for generalized zero-shot learning. In *CVPR*, pages 12661–12670, 2020. 1, 2, 6

[37] Ashish Mishra, M. K. Reddy, A. Mittal, and H. Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *CVPR Workshop*, pages 2269–2277, 2018. 1

[38] Sanath Narayan, A. Gupta, F. Khan, Cees G. M. Snoek, and L. Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *ECCV*, 2020. 1, 2, 3, 5, 8

[39] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729, 2008. 5

[40] M. Oquab, L. Bottou, I. Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, pages 1717–1724, 2014. 3

[41] Mark Palatucci, D. Pomerleau, Geoffrey E. Hinton, and Tom Michael Mitchell. Zero-shot learning with semantic output codes. In *NeurIPS*, pages 1410–1418, 2009. 1

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

[43] Akanksha Paul, Narayanan C. Krishnan, and Prateek Munjal. Semantically aligned bias reducing zero shot learning. In *CVPR*, pages 7049–7058, 2019. 2

[44] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In *NeurIPS*, 2019. 3

[45] Scott Reed, Zeynep Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016. 1, 5

[46] Mert Bülent Sariyildiz and Ramazan Gokberk Cinbis. Gradient matching generative networks for zero-shot learning. In *CVPR*, pages 2163–2173, 2019. 2, 6

[47] Edgar Schönfeld, S. Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. In *CVPR*, pages 8239–8247, 2019. 1, 2, 6

[48] Florian Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 4

[49] Yuming Shen, J. Qin, and L. Huang. Invertible zero-shot recognition flows. In *ECCV*, 2020. 1, 2, 3

[50] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 2, 3

[51] M. R. Vyas, Hemanth Venkateswara, and S. Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *ECCV*, 2020. 1, 2, 6

[52] Q. Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *International Journal of Computer Vision*, 124:356–383, 2017. 1, 2

[53] P. Welinder, S. Branson, T. Mita, C. Wah, Florian Schroff, Serge J. Belongie, and P. Perona. Caltech-ucsd birds 200. *Technical Report CNS-TR-2010-001, Caltech,*, 2010. 2, 3, 5

[54] Y. Wen, Kaipeng Zhang, Z. Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 4

[55] Jiamin Wu, Tianzhu Zhang, Z. Zha, Jiebo Luo, Yongdong Zhang, and Feng Wu. Self-supervised domain-aware generative network for generalized zero-shot learning. In *CVPR*, pages 12764–12773, 2020. 2

[56] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016. 2

[57] Yongqin Xian, T. Lorenz, B. Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018. 1, 2, 6

[58] Yongqin Xian, B. Schiele, and Zeynep Akata. Zero-shot learning — the good, the bad and the ugly. *CVPR*, pages 3077–3086, 2017. 1, 2, 5

[59] Yongqin Xian, Saurabh Sharma, B. Schiele, and Zeynep Akata. F-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, pages 10267–10276, 2019. 2, 3, 4, 6, 7, 8

[60] Guo-Sen Xie, L. Liu, Xiaobo Jin, F. Zhu, Zheng Zhang, J. Qin, Yazhou Yao, and L. Shao. Attentive region embedding network for zero-shot learning. In *CVPR*, pages 9376–9385, 2019. 1, 6

[61] Guo-Sen Xie, L. Liu, Xiaobo Jin, F. Zhu, Zheng Zhang, Yazhou Yao, J. Qin, and L. Shao. Region graph embedding network for zero-shot learning. In *ECCV*, 2020. 1

[62] Guo-Sen Xie, Xu-Yao Zhang, Yazhou Yao, Zheng Zhang, Fang Zhao, and Ling Shao. Vman: A virtual mainstay alignment network for transductive zero-shot learning. *IEEE Transactions on Image Processing*, 30:4316–4329, 2021. 2

[63] Guo-Sen Xie, Zheng Zhang, Guoshuai Liu, Fan Zhu, Li Liu, Ling Shao, and Xuelong Li. Generalized zero-shot learning with multiple graph adaptive generative networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2

[64] H. Yu and B. Lee. Zero-shot learning via simultaneous generating and learning. In *NeurIPS*, 2019. 1

[65] Y. Yu, Zhong Ji, J. Han, and Z. Zhang. Episode-based prototype generating network for zero-shot learning. In *CVPR*, pages 14032–14041, 2020. 1, 2, 6

[66] F. Zhang and G. Shi. Co-representation network for generalized zero-shot learning. In *ICML*, 2019. 1, 2, 3, 6

[67] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, pages 6034–6042, 2016. 2