

Disentangling Semantic-to-visual Confusion for Zero-shot Learning

Zihan Ye, *Student Member, IEEE*, Fuyuan Hu[†], *Member, IEEE*, Fan Lyu, Linyan Li, and Kaizhu Huang[†], *Senior Member, IEEE*

Abstract—Using generative models to synthesize visual features from semantic distribution is one of the most popular solutions to ZSL image classification in recent years. The triplet loss (TL) is popularly used to generate realistic visual distributions from semantics by automatically searching discriminative representations. However, the traditional TL cannot search reliable unseen disentangled representations due to the unavailability of unseen classes in ZSL. To alleviate this drawback, we propose in this work a multi-modal triplet loss (MMTL) which utilizes multi-modal information to search a *disentangled* representation space. As such, all classes can interplay which can benefit learning disentangled class representations in the searched space. Furthermore, we develop a novel model called Disentangling Class Representation Generative Adversarial Network (DCR-GAN) focusing on exploiting the disentangled representations in training, feature synthesis, and final recognition stages. Benefiting from the disentangled representations, DCR-GAN could fit a more realistic distribution over both seen and unseen features. Extensive experiments show that our proposed model can lead to superior performance to the state-of-the-arts on four benchmark datasets. Our code is available at <https://github.com/FouriYe/DCRGAN-TMM>.

Index Terms—Zero-shot Learning, Generative Adversarial Network, Representation Learning, Deep Learning.

I. INTRODUCTION

Classical pattern recognition classifies images into categories only seen in the training stage [1], [2], [3]. In contrast, zero-shot learning (ZSL), one of the most active research topics in multimedia, aims at exploring unseen categories, which has recently drawn much attention [4], [5], [6], [7], [8], [9], [10], [11], [12]. Furthermore, Chao et al. propose the generalized zero-shot learning (GZSL) [13] in a more practical scenario. Different from ZSL, GZSL intends to recognize both seen and unseen classes during test time. Since ZSL/GZSL does not require a vast amount of new data, ZSL models could be utilized as an imitative solution in crucial and life-saving situations, e.g. current COVID-19 literature search [14], autonomous driving planning [15], [16].

To conduct zero-shot classification, researchers usually engage intermediate semantic features to bridge the gap from seen to unseen classes. Intermediate semantic features have

Z. Ye is with Suzhou University of Science and Technology (E-mail: zihhye@outlook.com).

F. Lyu is with Tianjin University.

L. Li is with Suzhou Institute of Trade & Commerce.

[†] Corresponding authors: Prof. Hu is with Suzhou University of Science and Technology, Suzhou, 215009 China (E-mail: fuyuanhu@mail.usts.edu.cn). Prof. Huang is with the Department of Electrical and Electronic Engineering in Xi'an Jiaotong-Liverpool University, Suzhou, 215123 China (E-mail: Kaizhu.Huang@xjtlu.edu.cn).

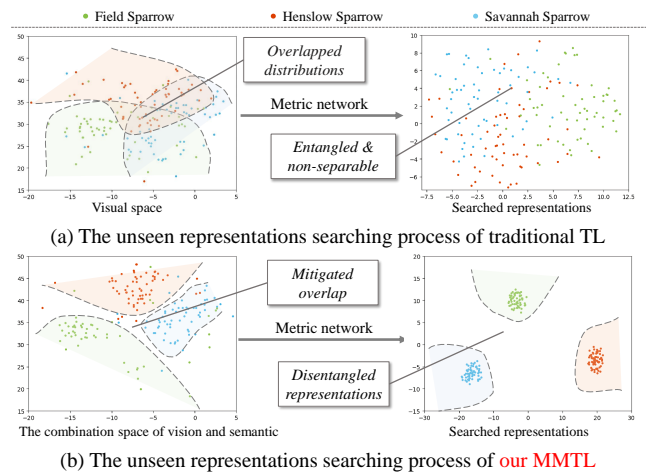


Fig. 1. Comparison between the traditional triplet loss (TL) and our multi-modal triplet loss (MMTL) on three unseen bird classes of CUB. Due to overlapped unseen visual distributions, searched representations by the tradition TL are too entangled. However, MMTL mitigates the overlapped unseen visual feature problem by combining visual and semantic features, and searches disentangled unseen representations consequently.

many alternatives, including attribute annotations [7], text representations from online text corpora [9], and even gaze embedding [17]. Based on these semantic features, researchers have explored two dominated types of ZSL methods, i.e. embedding methods, and generative methods. Embedding methods learn a projection from single modal features to another modal space for similarity measurement [18], [4], [10]. In contrast, generative methods focus on learning realistic unseen visual distributions from semantic features. They take advantages of the expressive power of generative adversarial networks (GANs) [19], [20] to generate plausible visual features for unseen classes [9], [21], [22], [23], [24]. In this way, ZSL can be converted to a conventional classification problem.

The connection between semantic and visual relationships is the key of the most ZSL/GZSL methods. Recently, researchers focus on how to define manually the constraints about the connection. For example, LsrGAN [25] claims that synthesized visual features of different classes should have a similar relationship to their semantic features. Thus, they propose to utilize the semantic relationship for guiding visual feature synthesis. However, semantic features might also be too ambiguous to be classified. Our previous work, SRGAN [22] investigates the visual relationships of different classes and argues that it could be used to rectify their over-smoothing semantics of some classes, and then, synthesize visual fea-

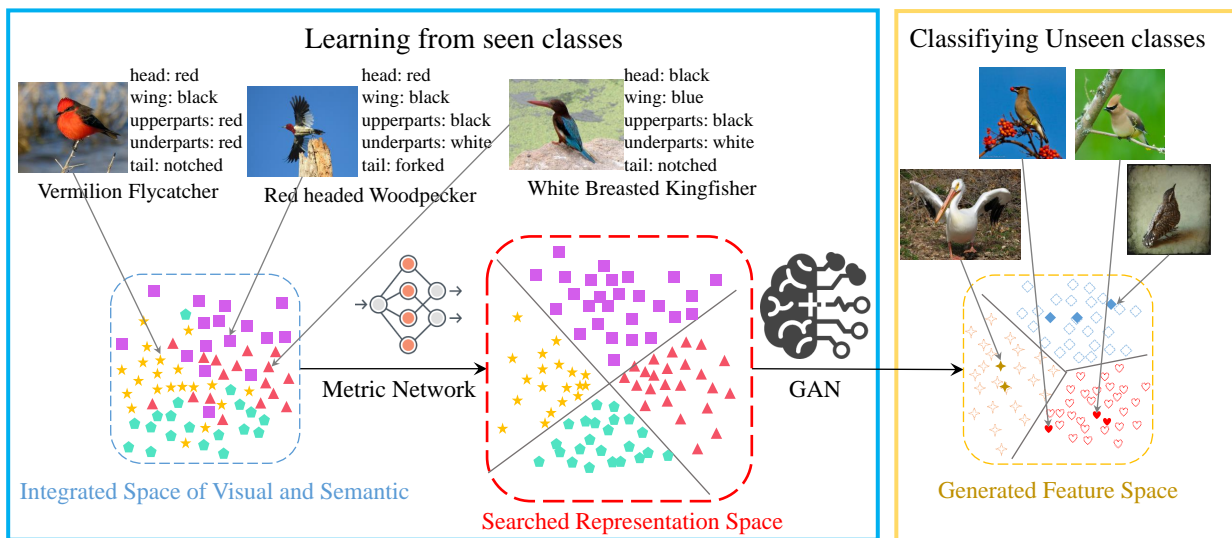


Fig. 2. Illustration of our proposed DCR-GAN approach. We develop the metric network to construct a clear class representation space SRS by the combination of visual and semantic spaces. Unseen class representation $R(\mathbf{a})$ is used to synthesize unseen visual features $\tilde{\mathbf{x}}$. As a result, we can recognize unseen class in the generated feature space.

tures from rectified semantic features. Other researchers also construct a re-representation space to align visual and semantic features simultaneously [12]. These methods focus on using single modal information, semantics or vision. Obviously, single modal information could be incomplete for classification. Using one modality to constrain the other is imperfect and would cause semantic-to-visual confusion. Thus, we pay our attention to two important questions: (1) how to find more disentangled/separable class representations by utilizing both semantic and visual information? (2) and consequently how to make the best use of disentangled representations for visual feature generation and recognition. Our idea is illustrated in Fig. 2.

To answer the first question, we notice that recent studies have designed a series of methods for automatically discriminative representation search [22], [11], [25], [12], [10] where the triplet loss is often used [26], [27], [28], [24], [10]. For example, Latent Discriminative Features Learning (LDF) [10] recognizes unseen samples in semantic and latent semantic space, which is searched by triplet loss (TL). TL is usually considered in the same fashion among these methods, i.e. they train a metric network (MN) and search a representation space from seen visual features by regulating both inter-class and intra-class distances. A well-designed MN would minimize the margin among intra-class samples and maximize the margin among inter-class samples. As a hypothetical result, these works suggest that the unseen classes could also form disentangled searched representations in the searched space.

In this paper, however, we find that TL may lead to a serious problem in ZSL/GZSL due to the inherited nature of ZSL, i.e. the unavailability of unseen visual features. Particularly, both visual feature extraction models and MN cannot access unseen features in ZSL. Since the feature extraction models are not trained for unseen classes, extracted visual distributions of different unseen classes would be overlapped, leading that unseen features will be entangled. Even if MN could

search discriminative seen class representations from seen visual features well, the TL training may be highly fragile to out-of-training-distribution features, which is similar to other margin-based losses [29], [30]. As a consequence, MN would produce non-separable and entangled unseen representations due to the under-fitting of unseen visual features, as shown in Fig. 1 (a). As such, we entitle the problem as *entangled unseen visual features problem*. This problem prevents the ZSL models from achieving the original purpose of using triplet loss, i.e. minimizing the distance among samples of the same classes and maximizing the distance among samples of different classes.

In this work, we propose to address the entangled unseen visual features problem by mitigating the entangled input condition. To this end, we develop the novel multi-modal triplet loss (MMTL), which combines two modal features, visual and semantic, to form more complete class descriptions. Compared to the traditional TL, our MMTL can utilize multi-modal information which can benefit disentanglement of the feature representations. Concretely, when visual features of samples from different classes are close, MMTL can utilize semantic information of samples to distinguish samples, and vice versa. Therefore, as shown in Fig. 1 (b), our MMTL is capable of searching disentangled representations for all the classes, even when some unseen visual features of different classes are close. In the testing stage for unseen samples, we can take sampling methods to obtain unseen representations. Note that, due to the instability in training GAN and triplet-based loss, we train our MN and our generator separately in this work.

To answer the second question, we further design the Disentangled Class Representation Generative Adversarial Network (DCRGAN), trying to make the best use of searched representations. DCRGAN integrates searched representations in all the stages of the ZSL pipeline, i.e. features synthesis, model training, and final classification. First of all, our gener-

ator synthesizes visual features from semantics and searched representations. Next, for model training, we point out that general GAN-based ZSL adversarial loss is not applicative for ZSL since they adopt a classification loss to make synthesized visual features more discriminative. However, though such classification loss intends to make synthesized features more separable, our investigation indicates that they cause the real features mixed-up together [28]; such learned synthesized features would lead to serious misclassification of real samples which are located in class boundaries. To tackle this problem, we propose our adversarial loss $\mathcal{L}_{WGAN-SR}$ by integrating auxiliary information, i.e. semantic features and searched representations, into our critical loss instead of attaching a classification loss. In the final classification stage, we train three softmax classifiers in visual, semantic, and searched representations spaces, respectively. Our results show that such integration can largely improve the accuracy both in seen and unseen classes.

Overall, this work has three main contributions.

1. We argue that the traditional TL has the inherited shortcoming for ZSL called the *entangled unseen visual feature problem*, i.e., the traditional TL cannot search appropriately disentangled representations for unseen classes.
2. We propose the MMTL to mitigate the entangled unseen feature problem. MMTL could search more disentangled representations than the traditional TL, which can be utilized to generate a more realistic distribution.
3. We propose a novel GAN-based framework named Disentangled Class Representation Generative Adversarial Network (DCR-GAN) for ZSL. DCR-GAN is capable of searching disentangled representations that are readily integrated in *all* the parts of ZSL. DCR-GAN achieves not only a high accuracy for classifying unseen images but also leads to significant improvement for classifying seen images.

II. RELATED WORK

A. Zero-shot Learning

Zero-shot Learning (ZSL) [5], [31], [32], [33], [34], [35], [22], [21], [36], [37] is one active research topic in multimedia, which aims at recognizing images from categories that are not included in the training set. Generalized Zero-Shot Learning proposed in [13] considers a more practical situation, in which both seen and unseen instances are mixed in the test data. One main challenge of ZSL/GZSL is that empirical risk minimization becomes unreliable [38], since unseen visual samples are not available in the training stage. This challenge also occurs in other relevant problems, i.e., Few-Shot Learning [39]. To overcome the limitations, researchers utilize semantics as intermediate representations of unseen classes. Such semantics are often manually defined attributes [4], word vectors [40] and text descriptions [9]. Other works also utilize gaze embedding, that is collected by non-experts, as semantics [17], [41].

Mainstreams ZSL methods can fall into embedding methods and generative methods. Embedding ZSL methods learn a visual-to-semantic embedding [18], [4], [10], a semantic-to-visual embedding [42], or an unified embedding space [43],

[12]. Generative ZSL methods focus on leveraging Generative Adversarial Network [19] and/or Variational Autoencoders (VAE) [44] to synthesize unseen visual features from semantic features.

Obviously, the quality of semantic features are the key of all the ZSL methods. Incomplete semantics would cause confusions of visual features generation. Semantic Rectifying GAN (SRGAN) [22] utilizes manually designed distance functions to rectify over-smoothing semantic features by visual similarities. Some embedding methods [10] and VAE-based methods [27], [28] try to utilize the triplet loss to search automatically more discriminative representations from visual features.

Though previous embedding methods and VAE-based methods have introduced TL to augment class representations, GAN-based methods hardly take concentration on utilizing representations searched by TL or other metric learning (partially due to their notorious training instability).

In this work, we make an attempt to fill the gap. We focus on searching automatically disentangled representations to enhance the fidelity of synthesized features; this is significantly different from the previous ways that manually define constraints between visual and semantic spaces. We also identify a novel problem of the traditional TL, and design a framework to utilize the searched representation in training, synthesizing, and recognition stages.

B. Triplet Loss in ZSL

The traditional TL was discussed by Google in FaceNet to search face representations for recognition [45]. It takes a metric network (MN) to project an anchor feature a , a positive feature p , and a negative feature n into the searched representation space. The anchor and positive features share the same class, while the anchor and negative features belong to different classes. MN aims to tighten up the margin between positive pairs (a, p) , and widen the margin between negative pairs (a, n) .

In previous ZSL works using TL, their MN are all trained by single-modal visual features. For example, one embedding ZSL method, Latent Discriminative Features Learning (LDF) [10], utilizes TL to mine new latent semantic features from visual features. In generative methods, Entropy-based Uncertainty calibration VAE (EUC-VAE) [27] and Over-Complete Distribution VAE (OCD-VAE) [28] integrate TL in VAE to enhance the separability of encoded representations. EUC-VAE designs two TLs trained by visual features and semantic features, respectively. OCD-VAE develops an online batch TL to speed up the process of gradient backward, but it still adopts the same TL formulation as that of LDF.

The above mentioned methods all ignore the entangled unseen visual features problem. Traditional TL is highly fragile for out-of-training-distribution features, which is similar to other margin-based losses [29]. In ZSL, TL is required to search representations not only for seen classes, but also for unseen features. However, unseen visual features are entangled. Traditional TL in ZSL lacks the ability in defense of overlaps among unseen distributions. Differently, in this

paper, we develop the Multi-Modal Triplet Loss (MMTL) that can mitigate the entangled problem by concatenating multi-modal features to form more complete descriptions of unseen classes. Benefiting from other modal information, the unseen distributions do not overlap. Consequently, our MN trained by MMTL can search disentangled representations which are usually entangled in the traditional TL. As such, MMTL can better meet the intention of using margin-based losses, i.e. maximizing inter-class variation and minimizing intra-class variation.

III. DISENTANGLED CLASS REPRESENTATION GENERATIVE ADVERSARIAL NETWORK

The training pipeline of our model follows two steps:

- (1) Pre-training Metric Network (MN, or in short M) for searching disentangled representations, and Semantic Rectify Network (SRN or in short R) for sampling searched representations from the semantic space.

This step will be described in Section III-B and be summarized in Algorithm 1.

- (2) Training a visual feature generator G with a discriminator D to synthesize pseudo visual features. We also utilize two regressors F_1 and F_2 to enhance the multi-modal consistencies among visual space, semantic space, and searched representation space.

This step will be introduced in Section III-C and III-D, and be summarized in Algorithm 2.

Once G is trained, we can train the final ZSL classifier with synthesized unseen features. Previous generative ZSL methods only train a visual ZSL classifier. Differently in this work, we also train a semantic and a searched representation classifier to make the best of auxiliary information. The test step will be presented in Section III-E.

A. Notations

Given an image I , the proposed model can recognize it as a specific class c even if it is unseen during training. We take instance $\{\mathbf{x}, \mathbf{a}^s, c^s\}$ as input in the training stage, where \mathbf{x} describes the instance-level visual feature in the visual feature space \mathcal{V} , \mathbf{a}^s in the seen semantic space, \mathcal{A}^s is class-level semantic extracted from attributes or other description information, and c^s denotes the corresponding seen class label. \mathcal{C}^s is the set of seen class labels. In the testing stage, given an image, ZSL and GZSL will recognize it as an unseen class \mathcal{C}^u or a class c^{s+u} (either seen or unseen). The unseen semantic space and the whole semantic space are denoted as \mathcal{A}^u , and $\mathcal{A} = \mathcal{A}^s \cup \mathcal{A}^u$ respectively.

B. Multi-modal Triplet Loss and Sampling Strategy

1) *Triplet loss in ZSL*: One primary obstacle of generative methods for ZSL mainly comes from the incomplete class semantic features. Such semantic features would confuse the model, as well as generating less reliable visual features. To search more comprehensive class representations, previous works try to engage TL to search discriminative class representations. Given an anchor visual feature \mathbf{x}_a^s with its label

Algorithm 1 Training algorithm of MN and SRN.

Require: The batch size, m_1 ; the number of epoch of training MN, n_M ; the number of epoch of training SRN, n_R ; initial MN parameters, θ_M ; the margin parameter, m ; initial SRN parameters, θ_R ; Adam hyper-parameters, α , β_1 , and β_2 .

- 1: **for** iter = 1, \dots , n_M **do**
- 2: Sample a minibatch of seen visual features \mathbf{x}^s , matching semantic features \mathbf{a}^s .
- 3: Compute the MMTL loss \mathcal{L}_{MMTL} using Eq. 3.
- 4: $\theta_M \leftarrow \text{Adam}(\nabla_{\mathcal{L}_{MMTL}}, \theta_M, \alpha, \beta_1, \beta_2)$
- 5: **end for**
- 6: **for** iter = 1, \dots , n_R **do**
- 7: **for** $t = 1, \dots, |\mathcal{C}^s|$ **do**
- 8: Sample all seen visual features \mathbf{x}^s , and matching semantic features \mathbf{a}^s of a certain class in \mathcal{C}^s .
- 9: Compute the sampling loss \mathcal{L}_{sam} using Eq. 4.
- 10: $\theta_R \leftarrow \text{Adam}(\nabla_{\mathcal{L}_{sam}}, \theta_R, \alpha, \beta_1, \beta_2)$
- 11: **end for**
- 12: **end for**
- 13: fix θ_M and θ_R

Algorithm 2 Training algorithm of feature generator.

Require: The maximal loops N_{loop} ; the batch size m ; the iteration number of discriminator in a loop N_d ; the iteration number of generator N_g ; initial generator parameters θ_G ; initial discriminator parameters θ_D ; the trained semantic rectifying Network R ; the gradient penalty hyper-parameter λ ; the two reconstruction parameters λ_1 and λ_2 ; Adam hyper-parameters α , β_1 , and β_2 .

- 1: **for** iter = 1, \dots , N_{loop} **do**
- 2: **for** iter = 1, \dots , N_d **do**
- 3: Sample a mini-batch of seen visual features \mathbf{x}^s , matching semantic features \mathbf{a}^s , and random noise \mathbf{z}
- 4: Compute the discriminator loss \mathcal{L}_D using Eq. 9.
- 5: $\theta_D \leftarrow \text{Adam}(\nabla_{\mathcal{L}_D}, \theta_D, \alpha, \beta_1, \beta_2)$
- 6: **end for**
- 7: **for** iter = 1, \dots , N_g **do**
- 8: Sample a mini-batch of seen visual features \mathbf{x}^s , corresponding semantic features \mathbf{a}^s , and random noise \mathbf{z}
- 9: Compute the reconstruction loss \mathcal{L}_{F_1} using Eq. 7.
- 10: $\theta_{F_1} \leftarrow \text{Adam}(\nabla_{\mathcal{L}_{F_1}}, \theta_{F_1}, \alpha, \beta_1, \beta_2)$
- 11: Compute the reconstruction loss \mathcal{L}_{F_2} using Eq. 8.
- 12: $\theta_{F_2} \leftarrow \text{Adam}(\nabla_{\mathcal{L}_{F_2}}, \theta_{F_2}, \alpha, \beta_1, \beta_2)$
- 13: Compute the generator loss \mathcal{L}_G using Eq. 9.
- 14: $\theta_G \leftarrow \text{Adam}(\nabla_{\mathcal{L}_G}, \theta_G, \alpha, \beta_1, \beta_2)$
- 15: **end for**
- 16: **end for**

c_a^s , the traditional TL in ZSL attempts to find a positive visual feature \mathbf{x}_p^s with the same label $c_p^s = c_a^s$ and a negative visual feature \mathbf{x}_n^s with $c_n^s \neq c_a^s$. Then, TL trains a MN by minimizing the distance of the positive pair $d(M(\mathbf{x}_a^s), M(\mathbf{x}_p^s))$ and maximizing the distance of the negative pair $d(M(\mathbf{x}_a^s), M(\mathbf{x}_n^s))$ by a margin m . MN would search a vector according to the input: $M(\mathbf{x}) \in \mathcal{R}^L$, where L is a hyper-parameter to control the size

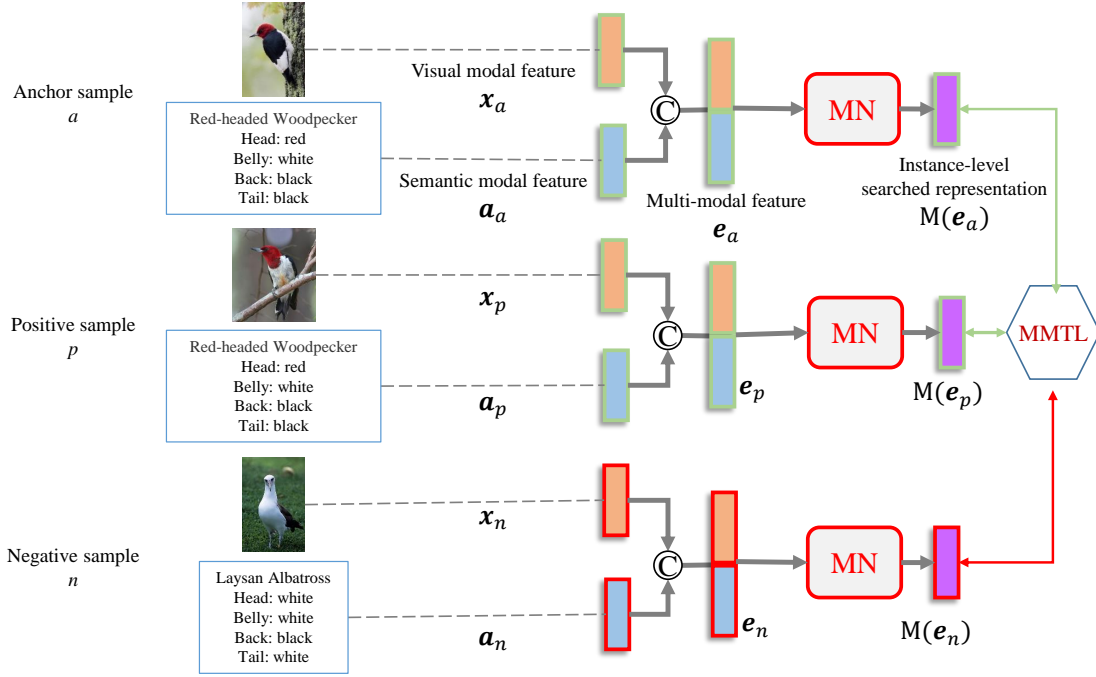


Fig. 3. Illustration of our proposed Multi-Modal Triplet Loss (MMTL), which consists of one positive pair (e_a, e_p) and one negative sample pair (e_a, e_n). Metric Network (MN) trained by our proposed MMTL would minimize the distance of all positive pairs ($M(e_a), M(e_p)$) and maximize the distance of all negative pairs ($M(e_a), M(e_n)$). The MN is instance-level.

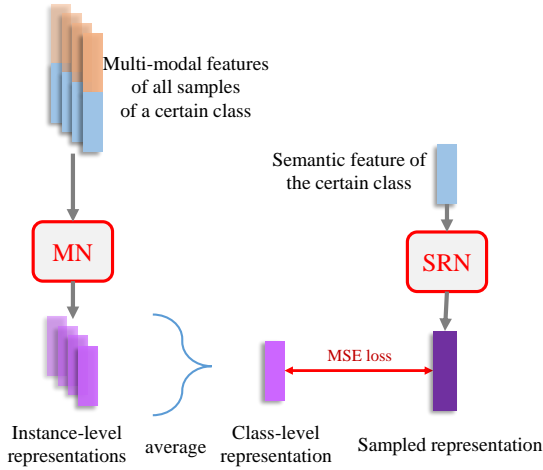


Fig. 4. Illustration of the sampling strategy for unseen representations. We need train another network Semantic Rectifying Network (SRN). The class-level searched representation is an average of all instance-level representations searched by MN. Then, we can train our SRN to learn the mapping: *semantics* \rightarrow *searched representations*. After that, we can obtain unseen searched representations that only need take unseen semantic features as input of SRN.

of searched vector. $d(\cdot, \cdot)$ is usually the Euclidean distance. It can be formulated as:

$$\mathcal{L}_{TL} = \max(0, m + d(M(\mathbf{x}_a^s), M(\mathbf{x}_p^s)) - d(M(\mathbf{x}_a^s), M(\mathbf{x}_n^s))). \quad (1)$$

The searched representation $M(\mathbf{x}^s)$ is instance-level representation. By averaging on all samples of the same class, we can obtain class-level representations.

2) *Limitations*: The tradition TL does not consider the *entangled unseen visual features problem* caused by overlapped unseen visual distributions in ZSL. MN cannot access any unseen features during the training stage. For example, there exist some very similar unseen visual features implying $\mathbf{x}_i^u \approx \mathbf{x}_j^u$, where i and j belong to two different classes. Although their semantic features, denoted as \mathbf{a}_i^u and \mathbf{a}_j^u , are different, visual features \mathbf{x}_i^u and \mathbf{x}_j^u are too smoothing to be distinguished by MN. Such entangled unseen visual features would confuse MN, and finally undermine the discriminativeness of searched unseen representations, i.e. $M(\mathbf{x}_i^u) \approx M(\mathbf{x}_j^u)$.

3) *Multi-modal Triplet Loss*: To tackle the entangled problem, we try to combine two modal features, i.e. both visual and semantic features, to form more complete class descriptions, as shown in Fig. 3. Mathematically, our proposed multi-modal triplet loss can be represented as:

$$\mathcal{L}_{MMTL} = \max(0, m + d(M(\mathbf{e}_a^s), M(\mathbf{e}_p^s)) - d(M(\mathbf{e}_a^s), M(\mathbf{e}_n^s))), \quad (2)$$

where \mathbf{e} is a concatenated feature from multiple modalities, e.g. vision, semantics, and/or gaze embedding. In this work, we concatenate visual and semantic modalities, i.e. $\mathbf{e}^s = [\mathbf{x}^s, \mathbf{a}^s]$ where $[\cdot, \cdot]$ denotes the concatenation operation.

Compared to the traditional TL, our MN can utilize multi-modal information to search a latent space which is sharp enough to distinguish different unseen classes. Obviously, when visual features of samples from different classes are close, MN can utilize semantic information of samples to distinguish samples, and vice versa. Additionally, we also use the weight decay to prevent over-fitting. The total loss of training MN is:

$$\mathcal{L}_{MN} = \mathcal{L}_{MMTL} + \|\theta_M\|_2. \quad (3)$$

Then, we can use sampling methods to get unseen representations from MN for generator training.

4) *Sampling Unseen Representation Strategy*: In recognizing unseen samples, we cannot know their labels before recognition. However, for a certain unseen class that has N_c samples, we need to recognize all visual features of the class to produce its class-level representation $\frac{1}{N_c} \sum_{i=1}^{N_c} M(\mathbf{e}_i^u)$. It is completely reversed with the training process of MN, whether using TL or MMTL. Thus, we need to design a sampling representation strategy to get unseen representations.

The traditional LDF [10] method designs a sampling method by training a relationship matrix \mathcal{W} that maps all seen semantics \mathcal{A}^s to all unseen semantics \mathcal{A}^u , i.e. $\mathcal{A}^u \approx \mathcal{W} \cdot \mathcal{A}^s$. Then, unseen representations $M(\mathbf{a}^u)$ can be obtained from the matrix and searched seen representations, i.e. $M(\mathcal{A}^u) = \mathcal{W} \cdot M(\mathcal{A}^s)$. Such sampling strategy has several drawbacks: (1) It is a transductive method since it takes unseen semantics to train. (2) It may bring incorrect semantic relationships into the searched representation space. For example, semantics of some classes are too smoothing to be distinguished [22]. (3) It forces that the searched representations have the same dimension with semantic features.

In contrast, instead of learning the whole seen-unseen relationships from semantics, we train another network, Semantic Rectifying Network (SRN), for directly mapping semantics, no matter whether seen or unseen, to the searched representation space, as shown in Fig. 4. Our sampling strategy need not unseen semantics in training. Thus it is inductive. Moreover, it can determine flexibly the dimensions of the searched representations. Our experiments also verify that the dimensions of searched representations can affect the ZSL classification.

Specifically, MN will be fixed after we finish its training. For any seen class, we minimize the l_2 loss between the average searched class representation and rectified semantic feature by SRN with the weight decay, i.e.

$$\mathcal{L}_{sam} = \left\| \frac{1}{N_c} \sum_{i=1}^{N_c} M(\mathbf{e}_i^s) - R(\mathbf{a}_i^s) \right\|_2 + \|\theta_R\|_2, \quad (4)$$

where N_c is the number of all samples of the chosen seen class. Then, for an unseen class u , we can directly get its searched class representation by rectifying its semantic feature, i.e. $R(\mathbf{a}_i^u)$. SRN and MN consist of a multi-layer perceptron (MLP) activated by Leaky ReLU, and the output layer does not apply any activation.

C. Visual Feature Synthesis with Search Representation

Generative Adversarial Network has demonstrated its usefulness for ZSL [9], [21], [22], [23], [24], [46], due to its promising ability to generate visual features from semantic features. The most popular generative ZSL methods are based on the conditional WGAN architecture with gradient penalty, which consists of a generator G , a discriminator D , and a classifier. The generator G synthesizes visual features from semantic features and normal distribution $z \sim \mathcal{N}(0, 1)$. The discriminator D distinguishes the synthesized samples \mathbf{x}_{fake} from real samples \mathbf{x} . The classifier predicts the probabilities of their label, $\log P(y|\mathbf{x}_{fake})$ and $\log P(y|\mathbf{x})$. The classifier, G

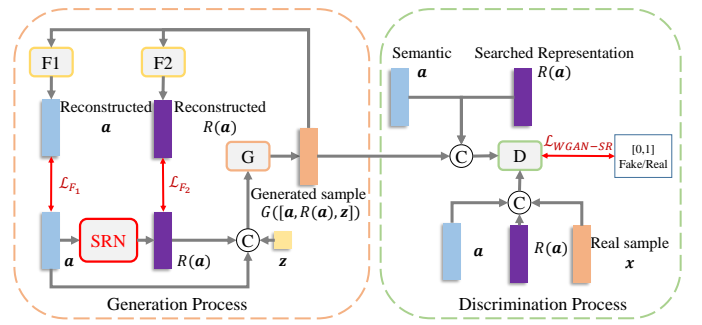


Fig. 5. Detailed illustration of the training process for DCR-GAN. G and D represent respectively the generator/discriminator of our feature GAN. The SRN projects semantic space to searched representation space. The $F1$ and $F2$ are two regress networks that respectively projects fake visual features $G(\mathbf{a}, R(\mathbf{a}), \mathbf{z})$ to semantic space and searched class representation space for the inter-class diversity.

and D are trained at the same time by the following minimax objective:

$$\begin{aligned} \min_G \max_D \mathcal{L}_{WGAN} = & \mathbf{E}_{\mathbf{z} \sim p_z} [D(\mathbf{x}_{fake})] - \mathbf{E}_{\mathbf{x} \sim p_{data}} [D(\mathbf{x})] \\ & + \mathbf{E}[\log P(y|\mathbf{x}_{fake})] - \mathbf{E}[\log P(y|\mathbf{x})] \\ & + \lambda(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2, \end{aligned} \quad (5)$$

where $\mathbf{E}(\cdot)$ denotes the expected value, $\mathbf{x}_{fake} = G(\mathbf{a}, \mathbf{z})$, λ is a parameter, and the last term $\lambda(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2$ is the gradient penalty to enforce the Lipschitz constraint [47], in which $\hat{\mathbf{x}} = \mu \mathbf{x} + (1 - \mu) \mathbf{x}_{fake}$ with $\mu \sim U(0, 1)$.

However, indiscriminately feeding vague semantic features into a generator may undermine the generated visual features. By a pre-trained SRN model, we can easily obtain more distinguished class representations. Therefore, we design a feature GAN model that translates these rectified semantic features into visual features.

\mathcal{L}_{WGAN} has two limitations: (1) It does not consider that semantic and visual space are heteroid. Some information might be missing completely in the other modal space. (2) Due to the entangled unseen visual features problem, many visual features are in the boundary area of other classes. In order to reduce the classification risk, G only generates samples that are far from the boundaries of classification, and thus, does not generate hard samples.

To address this problem, we propose the WGAN with the searched representations loss $\mathcal{L}_{WGAN-SR}$:

$$\begin{aligned} \mathcal{L}_{WGAN-SR} = & \mathbf{E}_{\mathbf{z} \sim p_z} [D([\mathbf{x}_{fake}, \mathbf{a}, R(\mathbf{a})])] \\ & - \mathbf{E}_{\mathbf{x} \sim p_{data}} [D([\mathbf{x}, \mathbf{a}, R(\mathbf{a})])] \\ & + \lambda(\|\nabla_{\hat{\mathbf{x}}} D([\hat{\mathbf{x}}, \mathbf{a}, R(\mathbf{a})])\|_2 - 1)^2, \end{aligned} \quad (6)$$

where $\mathbf{x}_{fake} = G(\mathbf{a}, R(\mathbf{a}), \mathbf{z})$. The training process of our DCR-GAN is described in Fig. 5. Our $\mathcal{L}_{WGAN-SR}$ enjoys two differences with \mathcal{L}_{WGAN} : (1) We integrate searched representations to align two modalities. (2) We remove the classifier and leverage auxiliary information, i.e. semantic features and searched representations, to train a class-sensitive discriminator D . With the integrated auxiliary information, interlaced class boundaries are pushed off. In this case, our generator G does not worry about the classification risk for hard samples.

D. Feature Reconstruction

With the above process, our model is able to synthesize proper visual features to some extent. However, there exists a significant problem, e.g. the generated visual features have poor consistencies with the input semantics and searched representations. Accordingly, we utilize two regression networks to keep the consistencies of $semantic \rightarrow visual \rightarrow semantic$ space and $searched\ representation \rightarrow visual \rightarrow searched\ representation$ space, respectively. Specifically, the regression network F_1 , keeping in step with the generator G , takes the generated feature \mathbf{x}_{fake} as input, and builds consistency losses between original semantics and reconstructed semantics from visual features. The regression network F_2 works in the same way for searched representations. The two regression loss for F_1 and F_2 can be computed by

$$\mathcal{L}_{F_1} = \|F_1(\mathbf{x}_{fake}) - \mathbf{a}\|_1, \quad (7)$$

$$\mathcal{L}_{F_2} = \|F_2(\mathbf{x}_{fake}) - R(\mathbf{a})\|_1. \quad (8)$$

Obviously, $G \circ F_1$ and $G \circ F_2$ can be considered as two Auto-Encoders [44], where $A \circ B$ denotes the composite of two mappings. The reconstruction $G \circ F_1$ enhances the relationship between the synthetic visual features and the corresponding class semantics by minimizing the difference between the reconstructed and original semantic features. The searched representation reconstruction $G \circ F_2$ works in the same way.

Finally, by integrating the reconstruction losses, the new objective of our DCR-GAN can be modified as:

$$\min_{G, F_1, F_2} \max_D \mathcal{L}_{DCR-GAN} = \mathcal{L}_{WGAN-SR} + \lambda_1 \mathcal{L}_{F_1} + \lambda_2 \mathcal{L}_{F_2}, \quad (9)$$

where λ_1 and λ_2 are two reconstruction parameters for semantic and searched representation, respectively.

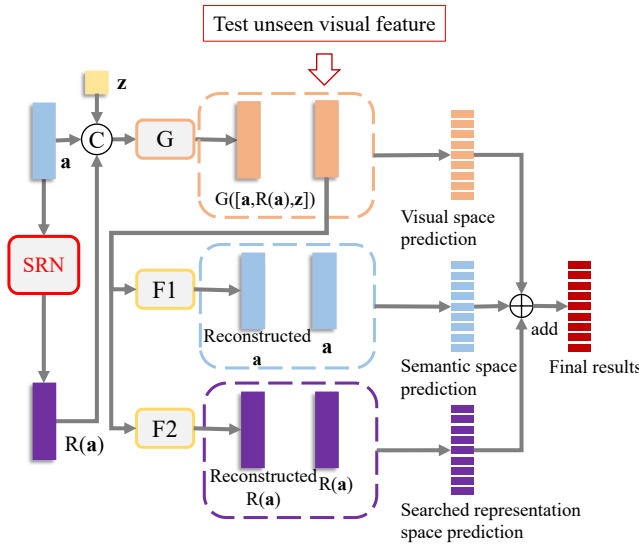


Fig. 6. Overview of zero-shot classification. We use our trained generator for unseen visual feature synthesis. Then, the synthesized feature is used to train a softmax classifier. Analogously, a semantic classifier and searched representation classifier are trained. We also use F_1 and F_2 to map all the real unseen visual features into the semantic space and searched representation space, respectively. We integrate the final results from visual, semantic, and searched representation space.

E. Zero-shot Classification with Searched Representation

By Eq. 9, we can train a GAN generator G which is able to synthesize virtual visual features for unseen categories. Then, we use the synthesized features to train a classifier, e.g. softmax, to recognize the real unseen instances. In other words, zero-shot learning is converted into a supervised classification problem that is performed in the visual space. As shown in Fig. 6, we train a softmax classifier by synthesized unseen visual features. We also use F_1 and F_2 to map all the real unseen visual features into the semantic space and searched representation space, respectively. Analogously, a semantic classifier and searched representation classifier are trained. Thus, in our model, we take full advantages of the visual space prediction score f_{VS} , semantic space prediction score f_{SS} , and searched representation space prediction score f_{SRs} . Finally, we get the final classification scores as:

$$f = f_{VS} + \omega_1 f_{SS} + \omega_2 f_{SRs}, \quad (10)$$

where ω_1 and ω_2 are two parameters used to balance the three terms, as shown in Fig. 6.

For GZSL, the main steps are the same as ZSL. The only difference lies in the final classification process. Specifically, the testing data of GZSL are from both the seen and unseen categories. Thus, we need to train a classifier on both real seen features and synthesized unseen features.

IV. EXPERIMENTS

A. Datasets

We evaluate our approach on four benchmark datasets for ZSL and GZSL: (1) Caltech-UCSD-Birds 200-2011 (CUB) [48] consisting of 11,788 images of 200 classes of birds annotated with 312 binary attributes; (2) Animals with Attributes (AWA) [7] consisting of 30,475 images of 50 animals classes with 85 attributes; (3) Attribute Pascal and Yahoo (APY) [49] containing of 15,339 images, 32 classes and 64 attributes from both PASCAL VOC 2008 dataset and Yahoo image search engine; (4) SUN Attribute (SUN) [50] annotating 102 attributes on 14,340 images from 717 types of scene. For the four datasets, we use the widely-used ZSL and GZSL split proposed in [5]. For clarity, the statistics of these datasets are summarized in Table I.

We adopt the evaluation metrics proposed in [5]. For ZSL, we measure *average per-class top-1 accuracy* ($T1$) of unseen classes C_u . It is defined as follows:

$$T1 = \frac{1}{\|C_u\|} \sum_i^{\|C_u\|} \frac{\# \text{ of correction predictions in } i}{\# \text{ of samples in } i}. \quad (11)$$

For GZSL, we compute the average per-class top-1 accuracy of seen classes C_s , denoted by S , the average per-class top-1 accuracy of unseen classes C_u , denoted by U , and their harmonic mean, i.e. $H = 2 \times (S \times U) / (S + U)$.

B. Comparison to State-of-the-arts

We compare our model with the recent state-of-the-arts published in the last few years. Embedding methods include DEVISE [18] (NeurIPS13), DAP [4] (TPAMI14), SSE

TABLE I
STATISTICS OF DATASETS.

Dataset	#attributes	#seen classes (train+val)	#unseen classes	#images (total)	#images (train+val)	#images (test unseen/seen)
AWA1 [7]	85	27+13	10	30475	19832	4958/5685
CUB [48]	312	100+50	50	11788	7057	2679/1764
APY [49]	64	15+5	12	15339	5932	7924/1483
SUN [50]	102	580+65	72	14340	10320	1440/2580

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON FOUR DATASETS.

Approach	Zero-shot Learning				Generalized Zero-shot Learning											
	AWA1	CUB	APY	SUN	AWA1			CUB			APY			SUN		
	T1	T1	T1	T1	U	S	H	U	S	H	U	S	H	U	S	H
Embedding approaches																
DEVISE [18]	54.2	52.0	39.8	56.5	13.4	68.7	22.4	23.8	53.0	32.8	4.9	76.9	9.2	16.9	27.4	20.9
DAP [4]	44.1	40.0	33.8	39.9	0.0	88.7	0.0	1.7	67.9	3.3	4.8	78.3	9.0	4.2	25.1	7.2
SSE [43]	60.1	43.9	34.0	51.5	7.0	80.5	12.9	8.5	46.9	14.4	0.2	78.9	0.4	2.1	36.4	4.0
SJE [34]	65.6	53.9	32.9	53.7	11.3	74.6	19.6	23.5	59.2	33.6	3.7	55.7	6.9	14.7	30.5	19.8
ESZSL [51]	58.2	53.9	38.3	54.5	6.6	75.6	12.1	12.6	63.8	21.0	2.4	70.1	4.6	11.0	27.9	15.8
ALE [6]	59.9	54.9	39.7	58.1	16.8	76.1	27.5	23.7	62.8	34.4	4.6	73.7	8.7	21.8	33.1	26.3
LATEM [52]	55.1	49.3	35.2	55.3	7.3	71.7	13.3	15.2	57.3	24.0	0.1	73.0	0.2	14.7	28.8	19.5
SYNC [53]	54.0	55.6	23.9	56.3	8.9	87.3	16.2	11.5	70.9	19.8	7.4	66.3	13.3	7.9	43.3	13.4
SAE [32]	53.0	33.3	8.3	40.3	1.8	77.1	3.5	7.8	54.0	13.6	0.4	80.9	0.9	8.8	18.0	11.8
PSR [54]	-	56.0	38.4	61.4	-	-	-	24.6	54.3	33.9	13.5	51.4	21.4	20.8	37.2	26.7
CDL [12]	69.9	54.5	43.0	63.6	28.1	73.5	40.6	23.5	55.2	32.9	19.8	48.6	28.1	21.5	34.7	26.5
CRNet [55]	-	-	-	-	58.1	74.7	65.4	45.5	56.8	50.5	32.4	68.4	44.0	34.1	36.5	35.3
DVBE (fixing) [56]	-	-	-	-	-	-	-	53.2	60.2	56.5	32.6	58.3	41.8	45.0	37.2	40.7
Generative approaches																
f-CLSWGAN [21]	68.2	57.3	40.5	60.8	57.9	61.4	59.6	43.7	57.7	49.7	32.9	61.7	42.9	42.6	36.6	39.4
cycle-CLSWGAN [23]	66.3	58.6	-	59.9	56.9	64.0	60.2	45.7	61.0	52.3	-	-	-	49.4	33.6	40.0
DASCN [57]	-	-	-	-	59.3	68.0	63.4	45.9	59.0	51.6	39.7	59.5	47.6	42.4	38.5	40.3
AFC-GAN [58]	69.1	62.9	45.5	63.3	58.2	66.8	62.2	53.5	59.7	56.4	36.5	62.6	46.1	49.1	36.1	41.6
GDAN [24]	-	-	-	-	-	-	-	39.3	66.7	49.5	30.4	75.0	43.4	38.1	89.9	53.4
GAZSL [9]	68.2	55.8	41.1	61.3	19.2	86.5	31.4	23.9	60.6	34.3	14.2	78.6	24.0	21.7	34.5	26.7
SRGAN [22]	72.0	55.4	44.0	62.3	41.5	83.1	55.3	31.3	60.9	41.3	22.3	78.4	34.8	22.1	38.3	27.4
OCD-VAE [28]	-	60.3	-	63.5	-	-	-	44.8	59.9	51.3	-	-	-	44.8	42.9	43.8
EUC-VAE [27]	65.7	61.7	39.1	63.5	60.4	70.4	65.1	50.8	55.1	52.9	44.1	36.8	40.1	35.0	62.7	44.9
LsrGAN [25]	66.4	60.3	-	62.5	54.6	74.6	63.0	48.1	59.1	53.0	-	-	-	44.8	37.7	40.9
DCR-GAN (ours)	71.0	61.0	48.0	63.7	62.7	73.3	67.6	55.8	66.8	60.8	37.2	71.7	49.0	47.1	38.5	42.4

TABLE III

ABLATION STUDY IN THE GZSL SETTING. THE RESULTS ARE REPORTED AS AVERAGE PER-CLASS TOP-1 ACCURACY OF UNSEEN CLASSES (U), SEEN CLASSES (S) AND THE HARMONIC MEAN (H). VS, SS AND SRS REPRESENTS CLASSIFIERS IN VISUAL, SEMANTIC, AND SEARCHED REPRESENTATION SPACE, RESPECTIVELY.

Variant	Loss	Classifier	Generalized Zero-Shot Learning											
			AWA1			CUB			APY			SUN		
			U	S	H	U	S	H	U	S	H	U	S	H
A	$\mathcal{L}_{WGAN} + \mathcal{L}_{F1}$ ($\mathbf{x}_{fake} = G(\mathbf{a}, \mathbf{z})$)	VS	55.1	64.7	59.5	30.5	51.0	38.2	21.6	55.2	31.1	34.1	37.3	35.6
B	$\mathcal{L}_{WGAN} + \mathcal{L}_{F1} + \mathcal{L}_{F2}$ ($\mathbf{x}_{fake} = G(\mathbf{a}, R(\mathbf{a}), \mathbf{z})$)	VS	58.5	62.0	60.2	38.5	50.7	43.8	22.7	54.0	32.0	36.3	35.1	35.7
C	$\mathcal{L}_{DCRGAN} =$ $\mathcal{L}_{WGAN-SR} + \mathcal{L}_{F1} + \mathcal{L}_{F2}$	(C-1) VS	54.1	64.7	58.9	44.0	54.6	48.7	29.8	60.6	39.9	40.5	36.0	38.1
		(C-2) SS	56.3	75.7	64.6	54.4	34.4	42.1	31.1	44.9	36.8	48.0	8.6	15.0
		(C-3) SRS	49.9	62.5	57.1	33.6	27.3	30.1	32.3	71.1	44.4	24.0	9.8	14.0
		(C-4) VS+SRS	56.1	65.9	60.6	47.1	54.8	50.6	37.9	66.3	48.2	42.1	36.0	38.8
		(C-5) VS+SS+SRS	62.7	73.3	67.6	55.8	66.8	60.8	37.2	71.7	49.0	47.1	38.5	42.4

[43] (ICCV15), SJE [34] (CVPR15), ESZSL [51] (ICML15), ALE [6] (TPAMI16), LATEM [52] (CVPR16), SYNC [53] (CVPR16), SAE [32] (CVPR17), CRNet [55] (ICML19) and DVBE [56] (CVPR20); generative methods include GAZSL (CVPR18) [9], PSR (CVPR18) [54], f-CLSWGAN [21] (CVPR18), CDL [12] (ECCV18), SRGAN [22] (ICME19), GDAN [24] (CVPR19), DASCN [57] (NeurIPS19), AFC-GAN [58] (ACM MM19), OCD-VAE [28] (CVPR20), EUC-VAE [27] (arxiv21) and LsrGAN [25] (ECCV20). The results of ZSL and GZSL are reported in Table II. We first take an overall comparison with the state-of-the-arts in Section IV-B1 for ZSL and GZSL. Then, we compare our approach with others in Section IV-B2 and Section IV-B3 from two aspects, GAN-based and triplet-loss-based viewpoints, respectively.

1) *(Generalized) Zero-shot Learning*: For ZSL, from the results reported in Table II, we can see that we obtain 2.5%, 0.1% improvements on APY and SUN, respectively, against the previous state-of-the-art.

For GZSL, we follow previous work [5] and report the harmonic mean that can avoid the effects of extreme values. For instance, we can see from the results that SAE gets 1.8% for unseen and 77.1% for seen on CUB. Although the accuracy of seen classes is the best, the harmonic mean is only 3.5% due to the extremely low result on unseen categories. In a nutshell, the harmonic mean is high only if the accuracies on both seen and unseen categories are high. From the results, we can observe that our method achieves overall the best harmonic mean on all of the evaluations besides SUN. It indicates that our DCR-GAN is a stable method which can work well for both seen and unseen instances. In details, DCR-GAN outperforms the state-of-the-art LsrGAN with 4.6%, 7.8% and 1.5% improvements on AWA1, CUB and SUN, respectively. Notably, our method achieves the best on AWA1, CUB of the unseen categories, which significantly outperforms AFC-GAN by 4.5% and 2.3%. It is worth mentioning that we do not use any explicit constraint to avoid “train bias” problem, but our proposed model can still surpass AFC-GAN that uses the boundary loss to compel synthesized unseen features to be far away from seen features.

2) *Comparison with GAN-based Approaches*: The GAN-based methods, f-CLSWGAN, cycle-CLSWGAN, AFC-GAN, GAZSL, SRGAN and LsrGAN, share the same basic loss \mathcal{L}_{WGAN} . Benefiting from the prior semantic features to generate missing data, these GAN-based methods can obtain better performance than those earlier embedding approaches though one recent embedding method, i.e. DVBE also demonstrates excellent performance. In addition, GDAN is a new method that unifies generative, embedding, and metric learning as a basic architecture. Benefiting from the change in the basic architecture, GDAN demonstrates an excellent score S for seen classification in SUN, which indicates a better basic loss may be necessary.

Our approach DCR-GAN adopts the proposed loss $\mathcal{L}_{WGAN-SR}$. Comparing our DCR-GAN with other GAN-based methods, we observe that our method leads to competitive performance with others both in ZSL and GZSL. More concretely, just SRGAN and our DCR-GAN can recognize more than 70% unseen samples of AWA1 in ZSL. Although

DCR-GAN cannot beat GDAN on SUN in the GZSL setting, it attains 63.7% and shows much better performance than GDAN and all the other methods on SUN for ZSL. These outstanding performance of our DCR-GAN demonstrates the effectiveness of our proposed basic loss $\mathcal{L}_{WGAN-SR}$.

It is noted that GDAN appears to perform much better than all the other methods including DCR-GAN on SUN in terms of H . Exploiting an adversarial loss \mathcal{L}_{GDAN} to train the model, it is observed that GDAN may overfit seen samples almost on all the datasets. This results in its excellent performance S for seen classes (particularly $S = 89.9\%$ in SUN), while the accuracy of U appears much worse: lower than almost all the other GAN-based methods. Such drawback may limit its application in ZSL/GZSL where recognition of unseen samples may be more crucial. In contrast, our proposed adversarial loss $\mathcal{L}_{WGAN-SR}$ appears to achieve an outstanding balance for both seen and unseen classes. As a matter of fact, our method outperforms GDAN in CUB and APY in both U and H ; it is also better than GDAN in terms of U in SUN.

3) *Comparison with Triplet-loss-based Approaches*: Besides our DCR-GAN, the methods OCD-VAE and EUC-VAE also introduce TL for searching discriminative latent features. Our DCR-GAN exploits the proposed MMTL, while the other methods use the traditional single-modal TL. As observed, for ZSL, DCR-GAN outperforms the others on APY and SUN; for GZSL, the proposed method demonstrates the best performance on AWA1, CUB and APY. It is noted that, although DCR-GAN performs not as well as them on SUN for GZSL, our model generates better performance for unseen recognition. Namely, our DCR-GAN recognizes 47.1% unseen samples in SUN, while OCD-VAE and EUC-VAE only recognize 44.8% and 35.0%, respectively.

C. Ablation Study

To further verify the effectiveness of our approach, we take an ablation study on the searched class representations. Table III reports three variants in the setting of GZSL, respectively. We also provide convergence curve of three visual classifiers in Fig. 7. The variant A indicates the traditional WGAN-based ZSL method trained by:

$$\mathcal{L}_A = \mathcal{L}_{WGAN} + \lambda_1 \mathcal{L}_{F1}, \quad (12)$$

where $\mathbf{x}_{fake} = G(\mathbf{a}, \mathbf{z})$. To fairly verify the effectiveness of the searched representations, we also train the variant B . We only change \mathbf{x}_{fake} as $G(\mathbf{a}, R(\mathbf{a}), \mathbf{z})$, and add in its reconstruction loss. In a word, the variant B is trained by:

$$\mathcal{L}_B = \mathcal{L}_{WGAN} + \lambda_1 \mathcal{L}_{F1} + \lambda_2 \mathcal{L}_{F2}, \quad (13)$$

where $\mathbf{x}_{fake} = G(\mathbf{a}, R(\mathbf{a}), \mathbf{z})$. Another variant C presents the complete DCR-GAN model trained by \mathcal{L}_{DCRGAN} . VS, SS and SRS represents classifiers in the visual, semantic and searched representation space, respectively. Comparing the results, we highlight our discussion as follows:

- (1) **Effectiveness of our searched representation**: Comparing Table III A and B, we can find that the harmonic mean accuracies H are significantly improved in both

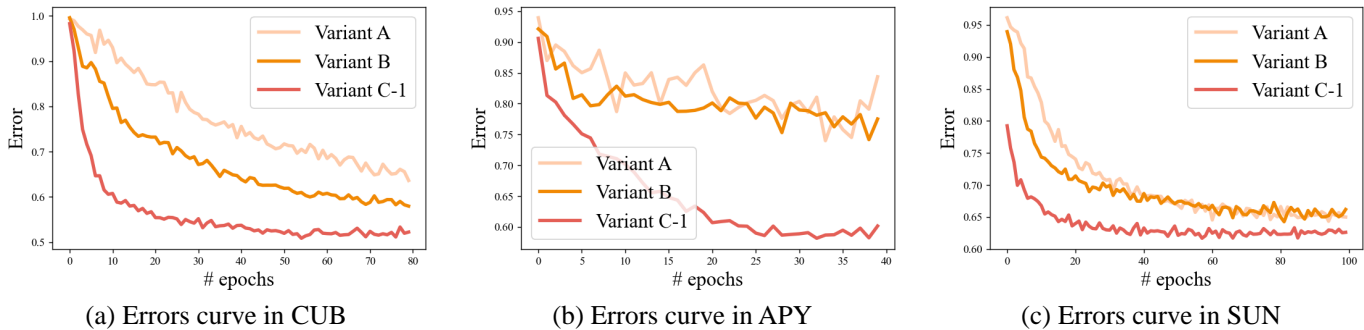


Fig. 7. Comparison between the visual softmax classifiers from DCRGAN and two Variants.

CUB and APY; unseen accuracies U are greatly improved on AWA1 and SUN. These results show that our searched representations can help G to fit more realistic distributions.

Then, comparing the variants C-1 and C-4, we can find that accuracies H in all the datasets are improved by integrating the classification results in the searched representation space. Specifically, H increases most notably in APY, from 39.9% to 48.2%. Even in the most indistinctive dataset SUN, our classifier SRS can still bring 1.6% improvement for unseen recognition.

- (2) **Effectiveness of our $\mathcal{L}_{WGAN-SR}$:** The variant B and C-1 are different only in terms of the generation loss. B is trained by \mathcal{L}_{WGAN} , while C-1 is trained by our proposed $\mathcal{L}_{WGAN-SR}$. Inspecting Table III B and C-1, we can find that though H is slightly decreased in AWA1, H is increased notably in the remained three datasets. For example, in the mid-scale dataset SUN, H is lifted up by 2.4% (from 35.7% to 38.1%). These results validate the effectiveness of our proposed $\mathcal{L}_{WGAN-SR}$.

D. Visualization of Feature Space

In order to provide a qualitative evaluation on our proposed DCR-GAN, we first visualize two kinds of unseen class representations as shown in Fig. 8, where (a) is one kind of augmented semantic learned by a triplet loss only from the visual space, i.e. the semantic augmentation method of LDF [10]. Clearly, this augmented semantic is muddled. Various categories are mixed in the representation space. Fig. 8 (b) shows the visualization of our searched class representations, which are learned both from visual and semantic spaces. As observed, by utilizing the original semantic information, our model decouples searched class representations where boundaries between each category are clear. Note that our model does *not* use any unseen features in training. This confirms our idea—the original semantic information is also necessary in the augmented semantic searching process.

In addition, we visualize synthetic image features along with real image features. These results are illustrated in Fig. 9. Since the numbers of categories of CUB and SUN are too large to visualize, we only show the results of seen classes of APY and unseen classes of APY, AWA1 and CUB. For

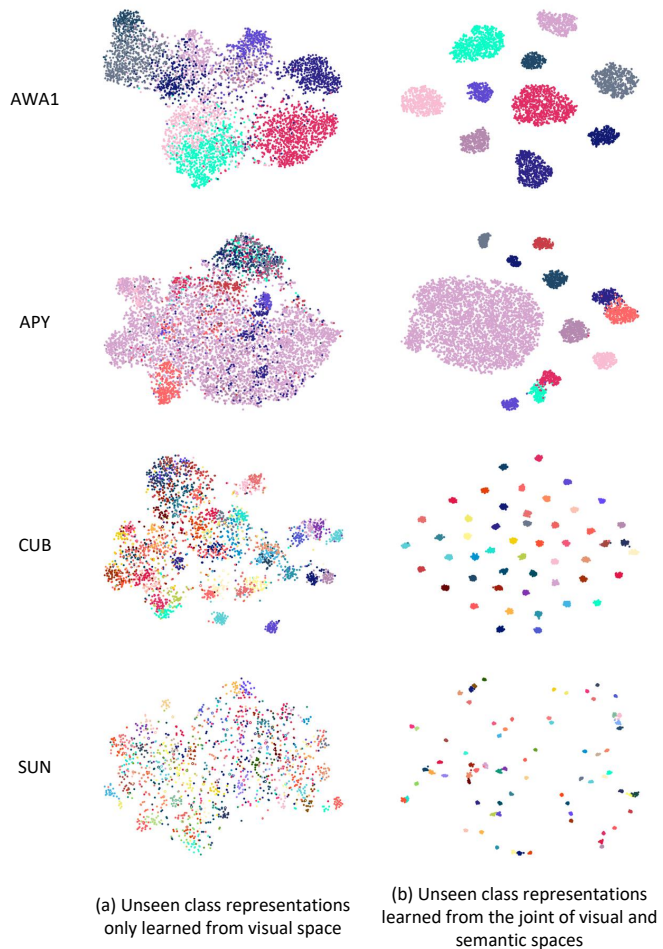


Fig. 8. Visualization of unseen class representations. (a) is searched by the traditional TL (Eq. 1); (b) is searched by the MMTL (Eq. 2). Different colors indicate different classes.

each class we synthesize 100 features, and then we use t-SNE [59] to reduce the dimension to two for visualization. The synthesized features of i -th class are marked by f_i , and real features are marked by r_i . It is evident that our searched class representations successfully help DCRGAN to synthesize more realistic visual features than those of the baseline. Some synthesized features by our DCRGAN are almost the same as real features, e.g. 14-th seen class in APY, 1-st unseen class in APY, 5-th unseen class in AWA1 and 28-th of

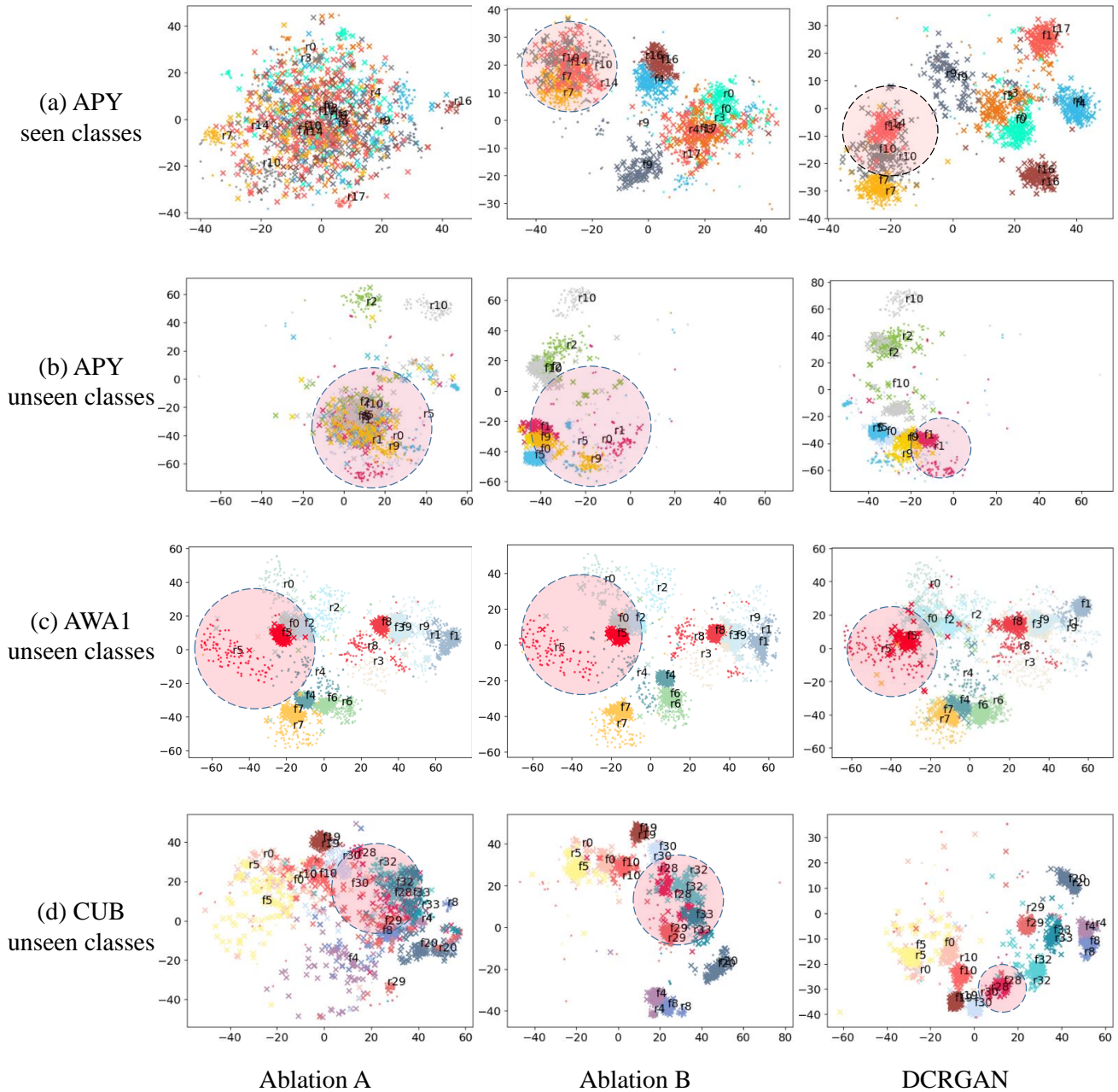


Fig. 9. Visualization of synthesized visual features. Points denote real features and crosses denote synthesized features. Different colors and numbers indicate different classes.

unseen classes in CUB. This visualization again indicates the identified *entangled unseen visual feature problem*. More importantly, our searched class representation can help generative models to fit more realistic distribution.

V. CONCLUSION

In this paper, we argue that the *entangled unseen visual feature problem* exists in the current Zero-shot Learning (ZSL) and Generalized ZSL. We propose our generative framework DCR-GAN to address the problem. DCR-GAN contains two novel loss: a multi-modal triplet loss (MMTL), and an adversarial loss $\mathcal{L}_{WGAN-SR}$. Compared with the traditional TL, MMTL is capable of searching more decoupled unseen class

representations. Our designed $\mathcal{L}_{WGAN-SR}$ can reduce high risks of hard sample generation. Benefiting from our MMTL and $\mathcal{L}_{WGAN-SR}$, our model learns more realistic distribution and generates more disentangled features. With the searched class representations, our DCR-GAN synthesizes visual features from semantic features and searched class representation. Given synthesized visual features, we train a softmax classifier for the visual space. Additionally, we ensemble the semantic and searched class representation softmax with the visual one. Experimental results show that the proposed approach achieves state-of-the-art performance on ZSL task and boosts the performance by a great margin for Generalized ZSL.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Nos. 61876121, 62002254, 61801323, 61876155), the Jiangsu Provincial Key Research and Development Program (Nos. BE2017663, BE2020006-4B), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 19KJB520054).

REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [2] F. Lyu, Q. Wu, F. Hu, Q. Wu, and M. Tan, "Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks," *IEEE Transactions on Multimedia (TMM)*, vol. 21, no. 8, pp. 1971–1981, 2019.
- [3] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking neural activations for fine-grained recognition," *IEEE Transactions on Multimedia (TMM)*, vol. 19(12), pp. 2736–2750, 2017.
- [4] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, no. 3, pp. 453–465, 2014.
- [5] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning — the good, the bad and the ugly," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3077–3086.
- [6] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 38, no. 7, pp. 1425–1438, 2016.
- [7] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 951–958.
- [8] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 49–58.
- [9] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, "A generative adversarial approach for zero-shot learning from noisy texts," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1004–1013.
- [10] Y. Li, J. Zhang, J. Zhang, and K. Huang, "Discriminative learning of latent features for zero-shot recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7463–7471.
- [11] J. Song, C. Shen, J. Lei, A.-X. Zeng, K. Ou, D. Tao, and M. Song, "Selective zero-shot classification with augmented attributes," in *European Conference on Computer Vision (ECCV)*, September 2018, pp. 474–490.
- [12] H. Jiang, R. Wang, S. Shan, and X. Chen, "Learning class prototypes via structure alignment for zero-shot recognition," in *The European Conference on Computer Vision (ECCV)*, September 2018, pp. 121–138.
- [13] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 52–68.
- [14] M. Sean, C. Arman, and G. Nazli, "Sledge-z: A zero-shot baseline for covid-19 literature search," *arXiv preprint arXiv:2010.05987*, 2020.
- [15] A. Filos, P. Tigkas, R. Mcallister, N. Rhinehart, S. Levine, and Y. Gal, "Can autonomous vehicles identify, recover from, and adapt to distribution shifts?" in *The International Conference on Machine Learning (ICML)*, vol. 119, 2020, pp. 3145–3153.
- [16] L. Yu, Q. Feng, Y. Qian, W. Liu, and A. G. Hauptmann, "Zero-virus: Zero-shot vehicle route understanding system for intelligent transportation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020, pp. 2534–2543.
- [17] N. Kaessli, Z. Akata, B. Schiele, and A. Bulling, "Gaze embeddings for zero-shot image classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 6412–6421.
- [18] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013, pp. 2121–2129.
- [19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.
- [20] Z. Ye, F. Lyu, L. Li, Y. Sun, Q. Fu, and F. Hu, "Unsupervised object transfiguration with attention," *Cognitive Computation*, vol. 11, pp. 869–878, 2019.
- [21] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5542–5551.
- [22] Z. Ye, F. Lyu, L. Li, Q. Fu, J. Ren, and F. Hu, "Sr-gan: Semantic rectifying generative adversarial network for zero-shot learning," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 85–90.
- [23] R. Felix, B. G. V. Kumar, I. D. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 21–37.
- [24] H. Huang, C. Wang, P. S. Yu, and C. Wang, "Generative dual adversarial network for zero-shot learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 801–810.
- [25] M. R. Vyas, H. Venkateswara, and S. Panchanathan, "Leveraging Seen and Unseen Semantic Relationships for Generative Zero-Shot Learning," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 70–86.
- [26] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2005, p. 1473–1480.
- [27] Z. Chen, Z. Huang, J. Li, and Z. Zhang, "Entropy-based uncertainty calibration for generalized zero-shot learning," *arXiv preprint arXiv:2101.03292*, 2021.
- [28] R. Keshari, R. Singh, and M. Vatsa, "Generalized zero-shot learning via over-complete distribution," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 13 297–13 305.
- [29] Y. Yuan, W. Chen, Y. Yang, and Z. Wang, "In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020, pp. 1454–1463.
- [30] F. Lyu, S. Wang, W. Feng, Z. Ye, F. Hu, and S. Wang, "Multi-domain multi-task rehearsal for lifelong learning," in *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021)*, 2021.
- [31] Z. Ding, M. Shao, and Y. Fu, "Generative zero-shot learning via low-rank embedded semantic dictionary," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 12, pp. 2861–2874, 2019.
- [32] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4447–4456.
- [33] Z. Zhang and V. Saligrama, "Zero-shot learning via joint latent similarity embedding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 6034–6042.
- [34] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 2927–2936.
- [35] Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, and H. T. Shen, "Zero-shot hashing via transferring supervised knowledge," in *Proceedings of the 24th ACM International Conference on Multimedia (ACM MM)*, 2016, p. 1286–1295.
- [36] G. Qi, W. Liu, C. C. Aggarwal, and T. S. Huang, "Joint intermodal and intramodal label transfers for extremely rare or unseen classes," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 7, pp. 1360–1373, 2017.
- [37] X. Shu, G. J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proceedings of the 23rd ACM Conference on Multimedia (ACM MM)*.
- [38] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys*, vol. 53(3), pp. 1–34, 2020.
- [39] Y. Zhu, W. Min, and S. Jiang, "Attribute-guided feature learning for few-shot image recognition," *IEEE Transactions on Multimedia (TMM)*, pp. 1200–1209, 2020.
- [40] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013, pp. 3111–3119.

- [41] Y. Liu, L. Zhou, X. Bai, Y. Huang, L. Gu, J. Zhou, and T. Harada, "Goal-oriented gaze estimation for zero-shot learning," *arXiv preprint arXiv:2103.03433*, 2021.
- [42] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto, "Ridge regression, hubness, and zero-shot learning," in *Joint European conference on machine learning and knowledge discovery in databases*, 2015, pp. 135–151.
- [43] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4166–4174.
- [44] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114*, 2013.
- [45] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.
- [46] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han, "From zero-shot learning to conventional supervised classification: Unseen visual data synthesis," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6165–6174.
- [47] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, p. 214–223.
- [48] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.
- [49] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, 2009, pp. 1778–1785.
- [50] G. Patterson, C. Xu, H. Su, and J. Hays, "The sun attribute database: Beyond categories for deeper scene understanding," *International Journal of Computer Vision (IJCV)*, vol. 108, no. 1–2, p. 59–81, 2014.
- [51] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *The International Conference on Machine Learning (ICML)*, vol. 37, 2015, pp. 2152–2161.
- [52] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 69–77.
- [53] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 5327–5336.
- [54] Y. Annadani and S. Biswas, "Preserving semantic relations for zero-shot learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 7603–7612.
- [55] F. Zhang and G. Shi, "Co-representation network for generalized zero-shot learning," in *The International Conference on Machine Learning (ICML)*, vol. 97, 2019, pp. 7434–7443.
- [56] S. Min, H. Yao, H. Xie, C. Wang, Z.-J. Zha, and Y. Zhang, "Domain-aware visual bias eliminating for generalized zero-shot learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 12 661–12 670.
- [57] J. Ni, S. Zhang, and H. Xie, "Dual adversarial semantics-consistent network for generalized zero-shot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 6146–6157.
- [58] J. Li, M. Jing, K. Lu, L. Zhu, Y. Yang, and Z. Huang, "Alleviating feature confusion for generative zero-shot learning," in *Proceedings of the 27th ACM International Conference on Multimedia (ACM MM)*, 2019, p. 1587–1595.
- [59] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research (JMLR)*, pp. 2579–2605, 2018.



Zihan Ye is an undergraduate student from Suzhou University of Science and Technology and a student member of the IEEE. His research interests include zero-shot learning, generative adversarial network, computer vision, and deep learning. He has served as a reviewer for IEEE SPL. In 2020 Oct, he took an oral presentation at IEEE International Conference on Image Processing (ICIP). In 2019, he took an oral presentation at IEEE International Conference on Multimedia & Expo (ICME). In 2018, he received the best poster paper award at The 8th International Conference on Brain Inspired Cognitive Systems (BICS). (Email: zihye@outlook.com)



Fuyuan Hu (corresponding author) was a postdoctoral researcher at Vrije Universiteit Brussel, Belgium, a Ph.D. student at Northwestern Polytechnical University, and a visiting Ph.D. student at the City University of Hong Kong. He is a professor at Suzhou University of Science and Technology. His research interests include machine learning, graphical models, structured learning, and tracking. (Email: fuyuanhu@mail.usts.edu.cn)



Fan Lyu is an undergraduate student from Suzhou University of Science and Technology and a student member of the IEEE. received the BS and MS degree in Electronic & Information Engineering, Suzhou University of Science and Technology, China, in 2015 and 2018. He is working toward the PhD degree in the College of Intelligence and Computing, Tianjin University, China. His research interests include lifelong learning, vision-language understanding, low-shot learning, etc.



Linyan Li is currently an associate Professor at Suzhou Institute of Trade & Commerce, China. She obtained her Master degree from Wuhan University in 2007. She has been working in machine learning, neural information processing, and pattern recognition.



Kaizhu Huang (corresponding author) is currently a Professor at Xi'an Jiaotong-Liverpool University, China. Prof. Huang obtained his PhD degree from Chinese University of Hong Kong (CUHK) in 2004. He worked in Fujitsu Research Centre, CUHK, University of Bristol, National Laboratory of Pattern Recognition, Chinese Academy of Sciences from 2004 to 2012. Prof. Huang has been working in machine learning, neural information processing, and pattern recognition. He was the recipient of 2011 Asia Pacific Neural Network Society Young Researcher Award. He received best paper or book award five times. Until September 2020, he has published 9 books and over 190 international research papers (70+ international journals) e.g., in journals (JMLR, Neural Computation, IEEE T-PAMI, IEEE T-NNLS, IEEE T-BME, IEEE T-Cybernetics) and conferences (NeurIPS, IJCAI, SIGIR, UAI, CIKM, ICDM, ICML, ECML, CVPR). He serves as associated editors/advisory board members in a number of journals and book series. He was invited as keynote speaker in more than 20 international conferences or workshops. (E-mail: Kaizhu.Huang@xjtlu.edu.cn)