Deep Unbiased Embedding Transfer for Zero-Shot Learning

Zhen Jia, Zhang Zhang, Liang Wang, *Fellow, IEEE*, Caifeng Shan, *Senior Member, IEEE*, and Tieniu Tan, *Fellow, IEEE*

***Abstract*— Zero-shot learning aims to recognize objects which do not appear in the training dataset. Previous prevalent mapping-based zero-shot learning methods suffer from the projection domain shift problem due to the lack of image classes in the training stage. In order to alleviate the projection domain shift problem, a deep unbiased embedding transfer (DUET) model is proposed in this paper. The DUET model is composed of a deep embedding transfer (DET) module and an unseen visual feature generation (UVG) module. In the DET module, a novel combined embedding transfer net which integrates the complementary merits of the linear and nonlinear embedding mapping functions is proposed to connect the visual space and semantic space. What's more, the end-to-end joint training process is implemented to train the visual feature extractor and the combined embedding transfer net simultaneously. In the UVG module, a visual feature generator trained with a conditional generative adversarial framework is used to synthesize the visual features of the unseen classes to ease the disturbance of the projection domain shift problem. Furthermore, a quantitative index, namely the score of resistance on domain shift (ScoreRDS), is proposed to evaluate different models regarding their resistance capability on the projection domain shift problem. The experiments on five zero-shot learning benchmarks verify the effectiveness of the proposed DUET model. As demonstrated by the qualitative and quantitative analysis, the unseen class visual feature generation, the combined embedding transfer net and the end-to-end joint training process all contribute to alleviating projection domain shift in zero-shot learning.**

***Index Terms*— Zero-shot learning, image classification, projection domain shift, convolutional neural network, generative adversarial network.**

Manuscript received April 7, 2019; revised August 29, 2019; accepted October 2, 2019. Date of publication October 22, 2019; date of current version November 27, 2019. This work was supported in part by the National Key R&D Program of China under Grant 2016YFB1001000, in part by the Natural Science Foundation of China under Grant 61525306, Grant 61633021, and Grant 61721004, and in part by CAS-AIR. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ming-Ming Cheng. *(Corresponding author: Zhang Zhang.)*

Z. Jia and Z. Zhang are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhen.jia@nlpr.ia.ac.cn; zzhang@nlpr.ia.ac.cn).

L. Wang and T. Tan are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, with the University of Chinese Academy of Sciences, Beijing 100190, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China (e-mail: wangliang@nlpr.ia.ac.cn; tnt@nlpr.ia.ac.cn).

C. Shan is with the Artificial Intelligence Research, CAS (CAS-AIR), Beijing 100190, China (e-mail: shancf@cas-air.cn).

Digital Object Identifier 10.1109/TIP.2019.2947780

I. Introduction

IMAGE classification methods have developed rapidly in the past decade with the progress of convolutional neural networks (CNNs) [1]–[4]. Despite their outstanding capability, a significant limitation of CNN models is that they highly rely on large-scale datasets, such as the ImageNet dataset [5], for better training parameters. In contrast, human beings can recognize image objects with more variations than what they have seen. Moreover, human beings can even recognize the objects they have just heard or read but never seen before.

Researchers try to endow the image classification models with the capability to recognize objects beyond the datasets. Lampert *et al.* [6] introduce the zero-shot learning (ZSL) problem, where the training image classes (so-called seen classes) are disjoint with the test image classes (so-called unseen classes). The target of ZSL is to design a model that can recognize object categories which do not appear in the training dataset.

Recently, researchers have paid much attention to the problem of ZSL [7]–[16]. Particularly, the mapping-based methods [8], [9], [11], [13], [17], [18], which learn a mapping function to project the image samples from visual space to semantic space, become popular. Nevertheless, these methods suffer from the projection domain shift problem [19]. The schematic illustration of this problem is shown in Figure 1. The ZSL models learn a mapping function on the data of seen classes. Then the mapping function is used to project the unseen class images from visual space to semantic space. As shown in Figure 1(A), an ideal unbiased mapping function should force the projected image samples of both seen and unseen classes surround their own semantic features also termed prototypes of the image classes. However, the training classes and test classes are disjoint in the ZSL task, so that the learned mapping function which is unbiased for training classes may produce somewhat derivations for the projected image samples from the semantic features when applied on the test classes, as shown in Figure 1(B). Specifically, the problem is that the mapped image samples of unseen classes will get biased from their prototypes when applying the mapping function learned on seen classes. Thus, the projection domain shift problem is the main challenge of zero-shot classification, which often leads to poor performance when applying a biased mapping function on unseen classes.

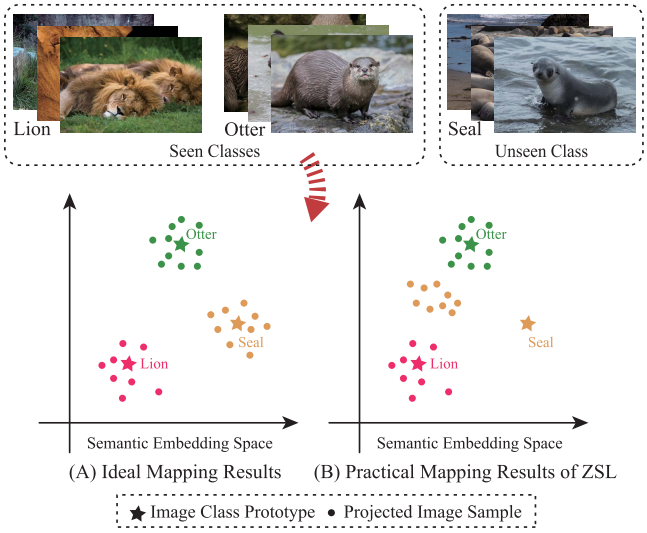


Fig. 1. The schematic illustration of the projection domain shift problem.

In this paper, we propose a deep unbiased embedding transfer (DUET) model for zero-shot image classification. The DUET model is designed to learn an unbiased embedding mapping to alleviate the projection domain shift problem. The DUET model consists of two modules: the deep embedding transfer module and the unseen visual feature generation module. The deep embedding transfer (DET) module consists of a CNN-based visual feature extractor and a combined embedding transfer net. Instead of the separated training manner in the previous work [8], [9], [11], [17], [18], [20], the DET module jointly trains the CNN feature extractor and the embedding transfer net in an end-to-end pipeline. The end-to-end joint training makes the CNN feature extractor and the embedding transfer net promote each other. Due to the ease of optimization and the promising performance, the linear mapping function is prevalent in the existing ZSL methods [8], [11], [13], [17], [18]. Meanwhile, the nonlinear model has better generalization capability which is important for the zero-shot learning task. Here the combined embedding transfer net is proposed to integrate the advantages of linear and nonlinear mapping functions. The unseen visual feature generation (UVG) module based on the Wasserstein generative adversarial network with gradient penalty (WGAN-GP) [21] is implemented to synthesize the visual features of unseen classes. With the training of both the synthesized unseen class visual features and the extracted seen class visual features, an unbiased embedding transfer net can be obtained. Different from previous ZSL methods with generative models [20], [22], [23], the proposed DUET model uses the synthesized visual features of unseen classes to obtain an unbiased embedding mapping between the visual and semantic spaces, instead of training the off-the-shelf classifiers directly.

The framework of the proposed DUET model is shown in Figure 2. Here, the training process is divided into three phases. In Phase I, the DET module is trained. In Phase II, the UVG module learns to generate visual features based on the semantic features. In Phase III, the visual feature generator (G-Net) trained by the UVG module is used to generate visual features of unseen classes and retrain the combined embedding transfer net.

Besides the DUET model, in order to quantitatively demonstrate the effectiveness of the model on projection domain shift, a novel index, namely the score of resistance on domain shift (ScoreRDS), is proposed. Our qualitative and quantitative evaluations verify the end-to-end joint training, the combined embedding transfer net and the synthesized visual features all contribute to alleviating the impact of the domain shift problem.

The contributions of this paper can be summarized in four-fold:

- This paper proposes a deep unbiased embedding transfer (DUET) model to alleviate the projection domain shift problem and promote the classification capability for zero-shot learning. In the DUET model, the deep embedding transfer (DET) module trains the visual feature extractor and the embedding transfer net jointly in an end-to-end manner. The unseen visual feature generation (UVG) module is designed to synthesize the visual features of unseen classes to obtain an unbiased embedding mapping.
- This paper proposes a combined embedding transfer net for the embedding mapping process in the model. It associates the advantages of linear and nonlinear mapping functions. The combined embedding transfer net is easy to optimize. Meanwhile, it has better generalization ability compared to the simple linear mapping net. As shown in the experimental results, the combined embedding transfer net endows the DUET model better capability to handle the projection domain shift problem with better classification accuracy.
- This paper proposes a novel quantitative index, the score of resistance on domain shift (ScoreRDS), to evaluate the effectiveness of the ZSL model on the projection domain shift problem. Besides the qualitative analysis, ScoreRDS can provide quantitative evaluations over different zero-shot learning models.
- Extensive experiments over five ZSL datasets, including AwA [6], AwA2 [12], CUB [24], aPY [25] and LAD [26], are conducted to validate the effectiveness of the proposed methods. Compared to state-of-the-art ZSL methods, the superior performance of the DUET model can be achieved. Furthermore, the qualitative and quantitative ablation studies demonstrate that the proposed end-to-end joint training, the combined embedding transfer net and the synthesized visual features all contribute to alleviating the projection domain shift problem of ZSL.

This paper is organized as follows. In Section II, related studies of zero-shot learning and generative adversarial net are introduced. Section III and Section IV describe the proposed DUET model and the ScoreRDS index in detail. Section V and Section VI show the implementation details and experimental results of the model. Finally, Section VII concludes the paper with the future prospect of zero-shot learning.

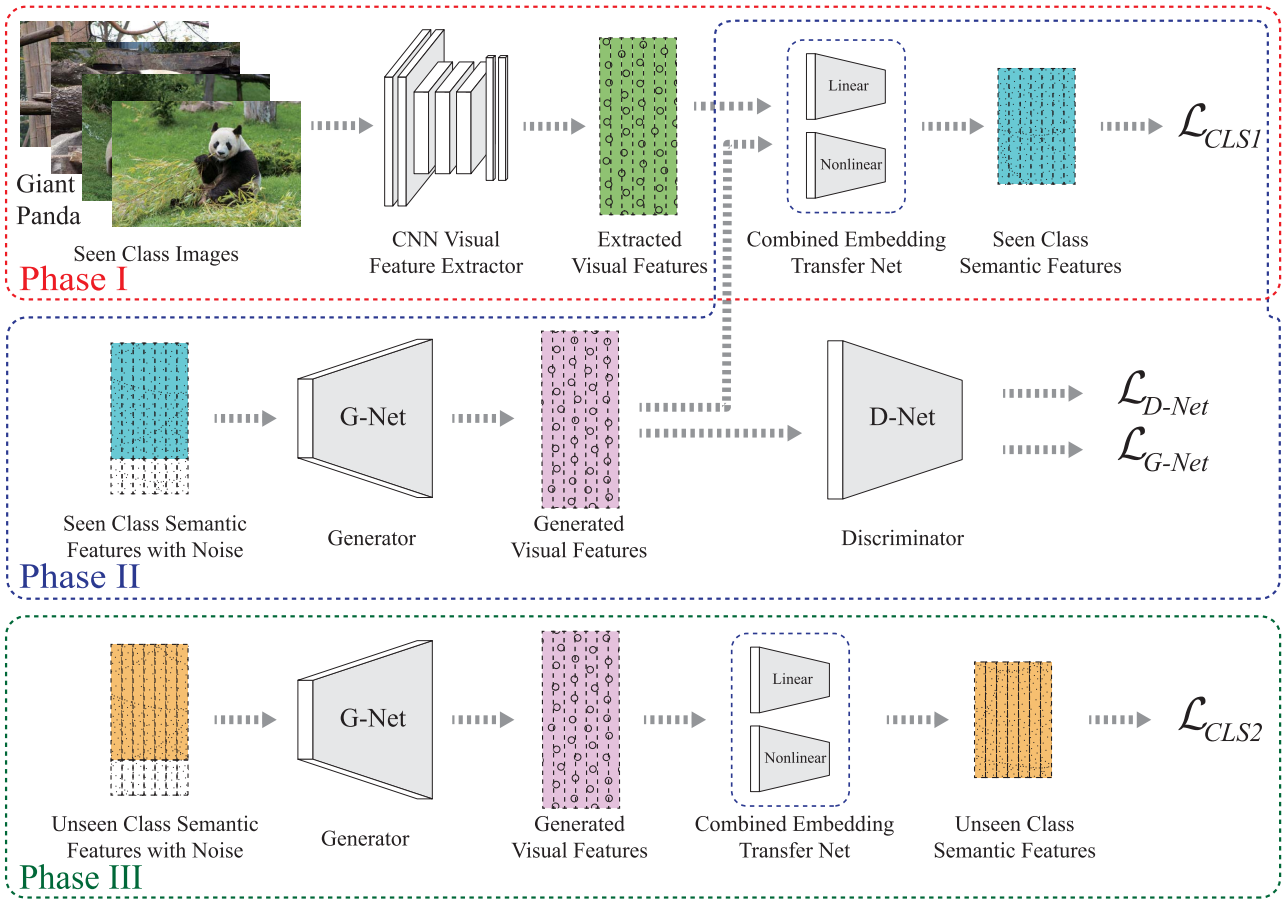


Fig. 2. The framework of the deep unbiased embedding transfer (DUET) model. The training process is divided into three phases. In Phase I, the deep embedding transfer (DET) module is trained. In Phase II, the unseen visual feature generation (UVG) module learns to generate visual features based on the semantic features. In Phase III, the trained G-Net of the UVG module is used to generate visual features of unseen classes to retrain the combined embedding transfer net.

II. Related Work

A. Zero-Shot Learning

Zero-shot learning has drawn lots of attention in recent years. Some studies put forward the models that combine the training classes to synthesize the test classes, termed hybrid models [7], [10], [27]. A more popular trend, termed mapping-based models [8], [9], [11], [13], [17], [18], attempts to learn an embedding mapping function to correlate the visual space and the semantic space. In the semantic space, the category of a test sample is decided by the distance between the image sample and the semantic prototype.

The deep visual-semantic embedding (DeViSE) model [8] is a representative mapping-based method. It trains a mapping matrix by optimizing a hinge ranking loss function. In the attribute label embedding (ALE) model [11], a bi-linear mapping function is learned between different embeddings using a ranking objective function. In the structured joint embedding (SJE) model [17], the mapping function is learned by optimizing the loss function based on the unregularized structured support vector machine (SVM). The embarrassingly simple approach to zero-shot learning (ESZSL) [18] learns a linear mapping matrix through a square loss with three specially designed regularizers. Socher *et al.* put forward a cross-modal transfer (CMT) model [9] using a two-layer neural network to learn a nonlinear mapping function. In particular, Morgado and Vasconcelos [13] make an effort to learn the mapping function with the widespread end-to-end training style in the semantically consistent regularization (SCoRe) model. The SCoRe model trains the mapping matrix as a layer in the CNN classification pipeline and implements semantic constraints to supervise attribute and category classification for zero-shot learning.

The aforementioned methods can be divided into two categories depending on the type of the embedding mapping function. DeViSE [8], ALE [11], SJE [17], ESZSL [18] and SCoRe [13] are linear models. On the other hand, CMT [9] is a nonlinear model.

B. Generalized Zero-Shot Learning

In the task setting of zero-shot learning, the test samples are from the unseen classes and the classification label space just includes the labels from unseen classes. A more realistic setting of ZSL is proposed in [28], which is termed generalized zero-shot learning (GZSL). In GZSL, the test samples come from both seen classes and unseen classes. Naturally, the label space includes not only unseen class labels, but also seen class labels. The schematic illustration of ZSL and GZSL is shown in Figure 3. In [12], the authors further evaluate the

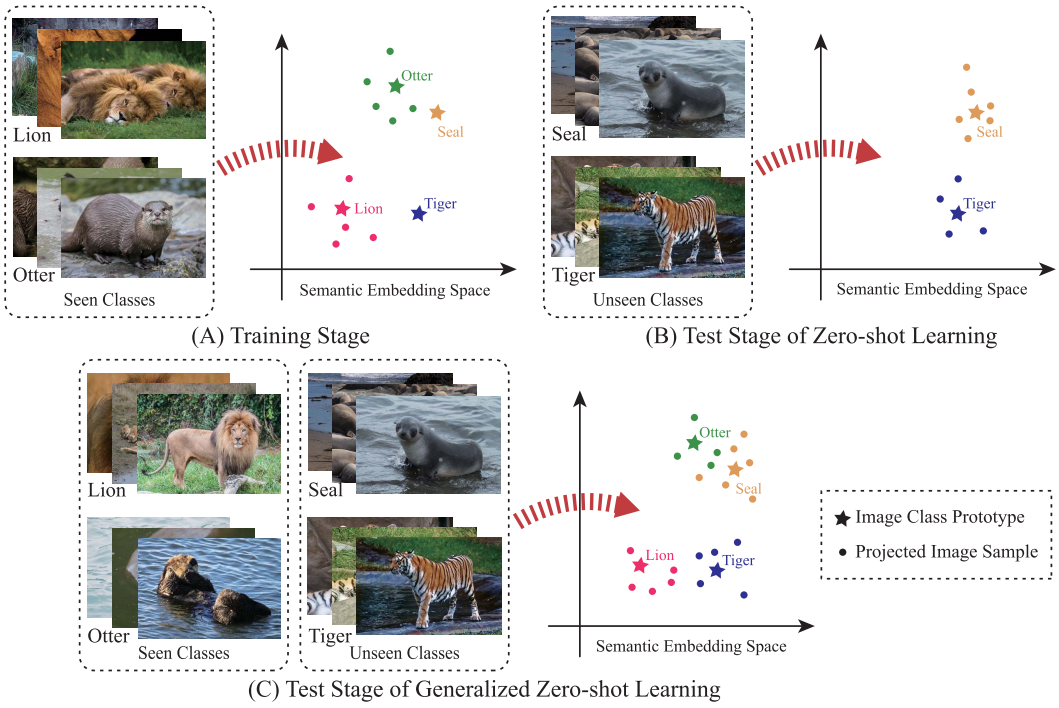


Fig. 3. The schematic illustration of zero-shot learning(ZSL) and generalized zero-shot learning (GZSL).

performance of different ZSL models on the task setting of GZSL. The harmonic mean of accuracies on seen and unseen classes is proposed as the evaluation criteria of GZSL [12]. In this paper, we report the performance of the proposed DUET model on the task settings of both ZSL and GZSL.

C. Generative Adversarial Nets for ZSL

Goodfellow *et al.* [29] propose the Generative Adversarial Net (GAN) in which the generator can be used to generate instances for specific tasks. However, the original GAN has some shortcomings that restrict its application, e.g., the difficulties of training, the lack of varieties of generated samples and the poor directivity of the loss function of the generator and the discriminator. To overcome these shortcomings, Wasserstein GAN (WGAN) [30] is proposed to use Wasserstein distance for optimization. However, the 1-Lipschitz constraint of Wasserstein distance is approximated by the weight clipping operation in practice. It leads to optimization difficulties. Further, Gulrajani *et al.* put forward the Wasserstein GAN with gradient penalty (WGAN-GP) [21] which contributes to a much more stable training of the model.

The current development of GAN methods provides a new approach to zero-shot learning. Several studies [20], [22], [23] have been proposed to address the ZSL task using the GAN models. Bucher *et al.* [22] propose to use auxiliary classifier GAN (AC-GAN) [31] to synthesize visual representations for unseen classes. Then the generated visual representations are used to train the classifier in the same way with supervised learning. In the generative adversarial approach for ZSL (GAZSL) [23], the authors propose to use visual pivots to constrain the training of the generator in GAN. Recently, Xian *et al.* [20] put forward the f-CLSWGAN model to train WGAN-GP with an auxiliary classification loss. The classification loss is used to reinforce the discriminative ability of the generated features. Similar to [22], [23], f-CLSWGAN trains a softmax classifier with the generated samples that turns the ZSL task into a supervised learning task.

With the synthesized data of unseen classes to train the ZSL models, the problem of projection domain shift can be alleviated to some degree. However, these methods utilize GAN to transform the ZSL task to supervised classification by training off-the-shelf classifiers with the generated unseen visual features directly. What's more, the training processes of GAN and the classifiers are separated from each other, which will influence the ZSL classification capability. Different from them, the UVG module in the proposed DUET model uses the generated visual features to train the combined embedding transfer net to connect the visual and semantic spaces. The training process maintains an end-to-end pipeline rather than the separated training manner. With the training of both the generated unseen class visual features and the extracted seen class visual features, an unbiased embedding transfer net can be obtained for the ZSL task.

III. Methodology

Similar to the aforementioned mapping-based methods, the DUET model utilizes an embedding mapping function to connect the visual embedding space and the semantic embedding space. Formally,

$$F(x, y; \mathbf{W}) = \langle \mathbf{W}^{\mathrm{T}}\mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{y}, \tag{1}$$

where $\mathbf{x}$ is the CNN visual representation of the image x and $\mathbf{y}$ is the semantic feature of the image category y. The inner product $\langle \cdot \rangle$ is the operation to calculate the compatibility score between $\mathbf{x}$ and $\mathbf{y}$. $\mathbf{W}$ is the mapping matrix to be

obtained. Under the task setting of ZSL, $\mathbf{W}$ is trained by the seen class data $\{\mathcal{X}_s, \mathcal{Y}_s\}$. Then the model is tested on the unseen class data $\{\mathcal{X}_u, \mathcal{Y}_u\}$. $\mathcal{X}$ and $\mathcal{Y}$ indicate the sets of visual and semantic representations of the image classes respectively. The subscripts s and u denote the seen and unseen classes respectively. Meanwhile, the model is tested on both the unseen class data $\{\mathcal{X}_u, \mathcal{Y}_u\}$ and the seen class data $\{\mathcal{X}_s, \mathcal{Y}_s\}$ under the task setting of GZSL. According to the problem definition of ZSL, the training categories and test categories are disjoint, i.e., $\mathcal{Y}_s \cap \mathcal{Y}_u = \varnothing$.

A. Deep Embedding Transfer (DET)

The DET module is one of the backbones in the proposed DUET model. It is designed to train the proper visual feature extractor and the embedding transfer net for zero-shot learning.

1) End-to-End Joint Training Process: Most previous ZSL methods [8], [9], [11], [17], [18] use a fixed visual feature extractor, such as the CNN model pre-trained on the ImageNet dataset. Then the training of embedding mapping function is separate from the image feature extraction, which is suboptimal for zero-shot classification.

In this paper, an end-to-end deep embedding transfer module is proposed to train the visual feature extractor and the embedding transfer net simultaneously. Compared to the fixed visual feature extractor, the joint training process contributes to extract better visual features for the specific ZSL task. What's more, the visual feature extractor and the embedding transfer net can promote each other in the training stage.

The framework of the DET module is shown in the Phase I part of Figure 2. The DET module is based on a CNN image classification network which is responsible to extract the visual features of input images. Then the visual features go through the combined embedding transfer net, after which the visual features are mapped into the semantic embedding space. In the semantic embedding space, the inner product is used to calculate the compatibility score between the image samples and the semantic features of the classes, i.e., the prototypes of the classes, in the form of Equation (1). The DET module is trained by the supervision of a cross-entropy loss function with softmax activation. Formally,

$$\mathcal{L}_{CLS1} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(\mathbf{x}_s^{i\mathrm{T}}\mathbf{W}\mathbf{y}_s^{c(i)})}{\sum_j^{C_s}\exp(\mathbf{x}_s^{i\mathrm{T}}\mathbf{W}\mathbf{y}_s^{j})}, \tag{2}$$

where N is the number of training samples and C_s is the number of seen classes. $c(i)$ indicates the category of the image sample $\mathbf{x}_i$.

2) Combined Embedding Transfer Net: Regarding embedding transfer, most of the previous zero-shot learning models use either linear [8], [11], [13], [17], [18] or nonlinear [9] methods according to the type of the embedding mapping function. Due to the convenience of optimization, most previous methods are based on the linear mapping function. However, the nonlinear mapping function can fit a larger embedding space and maintain more information during the embedding transfer process. Meanwhile, the nonlinearity can bring better generalization capability to the model.

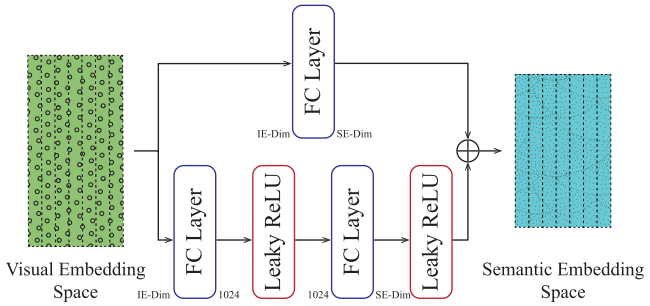


Fig. 4. The framework of the combined embedding transfer net. The input and output feature dimensions are marked on the left and right side of the *FC Layer* blocks. "IE-Dim" and "SE-Dim" indicate the dimensions of visual and semantic embedding features respectively.

To integrate the superiority of both linear and nonlinear mapping functions, a novel combined embedding transfer net is proposed in the DUET model. In the combined embedding transfer net, the linear and nonlinear mappings are performed separately. Then the embedding transfer results are summarized to represent the image samples in the semantic embedding space. Formally,

$$\mathbf{W} = \alpha\mathbf{W}_{Linear} + (1-\alpha)\mathbf{W}_{Nonlinear}, \tag{3}$$

where $\alpha \in [0, 1]$ is the hyperparameter to control the balance of the two mapping functions.

The framework of the combined embedding transfer net is shown in Figure 4. The linear mapping of the net is achieved by a fully connected layer (FC layer). The nonlinear mapping is composed of two FC layers with Leaky ReLU nonlinear activations.

Actually, with the mapping matrix $\mathbf{W}$ learned in the DET module, the zero-shot classification can be implemented conveniently. However, due to the lack of the unseen image classes in the training stage, the learned $\mathbf{W}$ is biased and suffers from the projection domain shift problem. Thus, the UVG module is proposed to deal with the problem.

B. Unseen Visual Feature Generation (UVG)

The UVG module is another backbone of the DUET model. This module is adopted to generate image features of unseen categories, which is used to adjust the embedding mapping for an unbiased projection. The UVG module is based on the WGAN-GP method due to its favorable characteristics on model training and feature generation.

The framework of the UVG module is shown in the Phase II part of Figure 2. The UVG module contains a generator (G-Net) and a discriminator (D-Net). The G-Net is trained to generate visual features with convincing similarity to real visual features. The D-Net is trained to discriminate whether the visual features are real or not. We adopt the conditional version WGAN-GP method by cascading the visual feature $\mathbf{x}$ and semantic feature $\mathbf{y}$ as the input of the D-Net. For the G-Net, the input is the semantic feature $\mathbf{y}$ concatenated with a Gaussian noise $\mathbf{z}$. The Gaussian noise is used to increase the distribution variance of the generated features.

The loss function of the discriminator is as follows.

$$\mathcal{L}_{D\text{-}Net} = \mathbb{E}[D(cat(\tilde{\mathbf{x}}_s, \mathbf{y}_s))] - \mathbb{E}[D(cat(\mathbf{x}_s, \mathbf{y}_s))] + \lambda\mathbb{E}[(\|\nabla_{\hat{\mathbf{x}}} D(cat(\hat{\mathbf{x}}, \mathbf{y}_s))\|_2 - 1)^2]. \quad (4)$$

$\tilde{\mathbf{x}}_s = G(cat(\mathbf{y}_s, \mathbf{z}))$ is the generated visual features on the basis of the corresponding semantic feature of a seen class $\mathbf{y}_s$. $D(\cdot)$ and $G(\cdot)$ are the operations of D-Net and G-Net respectively. $cat(\cdot)$ represents the concatenation operation. The third term in Equation (4) is the gradient penalty of WGAN-GP and λ is the coefficient. $\hat{\mathbf{x}} = \epsilon\mathbf{x}_s + (1-\epsilon)\tilde{\mathbf{x}}_s$, where $\epsilon \sim \mathcal{U}[0, 1]$, indicates that $\hat{\mathbf{x}}$ is sampled randomly from the connection line between $\mathbf{x}_s$ and $\tilde{\mathbf{x}}_s$.

Besides the WGAN-GP loss of the generator, the UVG module also applies a cross-entropy classification loss for the G-Net. The classification loss can enforce the generator to synthesize the visual features which are separable in the embedding space and suitable for the combined embedding transfer net. Formally,

$$\mathcal{L}_{G\text{-}Net} = -\mathbb{E}[D(cat(\tilde{\mathbf{x}}_s, \mathbf{y}_s))] - \frac{\sigma}{\tilde{N}} \sum_{i}^{\tilde{N}} \log \frac{\exp(\tilde{\mathbf{x}}_s^{i\,\mathrm{T}} \mathbf{W} \mathbf{y}_s^{c(i)})}{\sum_j^{C_s} \exp(\tilde{\mathbf{x}}_s^{i\,\mathrm{T}} \mathbf{W} \mathbf{y}_s^{j})}, \quad (5)$$

where $\tilde{\mathbf{x}}_s$ still indicates the synthesized visual features of the seen class from G-Net. σ is the coefficient of the classification loss and $\tilde{N}$ is the number of generated features.

After the UVG module is trained on the seen classes, the G-Net is able to generate the visual features according to the semantic features of the corresponding classes. Afterwards, the G-Net is taken out from the UVG module to generate the visual features of unseen classes, i.e., $\tilde{\mathbf{x}}_u$ independently as demonstrated in Phase III of Figure 2. The generated visual features of unseen classes are adopted to retrain the combined embedding transfer net. The process is guided by the classification loss function as follows.

$$\mathcal{L}_{CLS2} = -\frac{1}{M} \sum_{i}^{M} \log \frac{\exp(\tilde{\mathbf{x}}_u^{i\,\mathrm{T}} \mathbf{W} \mathbf{y}_u^{c(i)})}{\sum_j^{C_u} \exp(\tilde{\mathbf{x}}_u^{i\,\mathrm{T}} \mathbf{W} \mathbf{y}_u^{j})}, \quad (6)$$

where $\tilde{\mathbf{x}}_u = G(cat(\mathbf{y}_u, \mathbf{z}))$ is the generated visual features of unseen classes. M indicates the number of features and C_u indicates the number of unseen classes.

Since the embedding transfer net is trained with both the seen class visual features $\mathbf{x}_s$ and the generated unseen class visual features $\tilde{\mathbf{x}}_u$, the projection domain shift problem can be alleviated effectively.

C. Zero-Shot Classification

Following the training of the DUET model, the zero-shot classification is carried out on the images of unseen classes under the task setting of ZSL. For a test image of unseen classes, x_u, the visual feature $\mathbf{x}_u$ is extracted by the DET module. Afterwards, $\mathbf{x}_u$ is mapped to the semantic embedding space by the combined embedding transfer net. Then the inner product is calculated to determine the classification result of x_u. Formally,

$$f(x_u) = \arg\max_{y_u \in \mathcal{Y}_u} \langle \mathbf{W}^{\mathrm{T}} \mathbf{x}_u, \mathbf{y}_u \rangle. \quad (7)$$

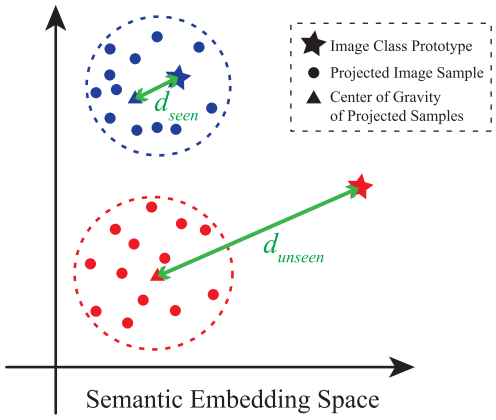


Fig. 5. The schematic illustration of the distance between image class prototype and the center of gravity of mapped samples.

Under the task setting of GZSL, the classification is carried out with the test images come from both seen classes and unseen classes. The classification label space contains all the image categories in the dataset. Formally,

$$f(x) = \arg\max_{y \in \mathcal{Y}_s \cup \mathcal{Y}_u} \langle \mathbf{W}^{\mathrm{T}} \mathbf{x}, \mathbf{y} \rangle. \quad (8)$$

In this paper, we evaluate the proposed DUET model under the task settings of both ZSL and GZSL.

IV. Score of Resistance on Domain Shift (ScoreRDS)

Zero-shot learning methods suffer from the projection domain shift problem inherently. Because the images of unseen classes do not appear in the training stage, the mapped unseen image samples will get biased from its class prototype in the test stage. The resistance capability of the model to the projection domain shift problem determines the model's performance on ZSL. Previous work usually demonstrates the model's capability on the projection domain shift problem indirectly by the zero-shot classification performance or qualitative visualization [32]–[34], which lacks a quantitative measure. Here we propose a novel index, the score of resistance on domain shift (ScoreRDS), to evaluate the model quantitatively.

The proposed ScoreRDS is based on the distance between the mapped image samples and the class prototypes in the semantic embedding space. As shown in Figure 1, the mapped image samples of seen classes will surround their own prototypes, but the samples of unseen classes will get biased from their prototypes in the practical ZSL situation.

When evaluating the model's capability to resist the projection domain shift problem, the mapped unseen class samples should be close to their own prototypes if the model can handle the projection domain shift problem sufficiently. As illustrated in Figure 5, we calculate the mean of the mapped features of a class, i.e., the center of gravity (CG) of the features, in the semantic space. Then the distance between the CG and the prototype of the class can be obtained. In Figure 5, the seen class and the unseen class are marked in blue and red respectively. The mapped image sample, the prototype and the CG of features are marked as dot,

pentacle and triangle respectively. The distance from the CG to the prototype of a seen class and an unseen class is tagged as d_{seen} and d_{unseen} respectively. Because different ZSL models will generate different data distributions in the semantic space, the magnitude of d_{seen} and d_{unseen} from different models may vary drastically. Consequently, it is unreasonable to just compare the d_{unseen} to determine models' resistant capability on domain shift. In another perspective, one model should have the same order of magnitude on one dataset for both d_{seen} and d_{unseen}. Therefore, we can use the distance of seen classes as the standard to evaluate the magnitude of the distance of unseen classes. In another word, d_{unseen}/d_{seen} can be used as the criteria to evaluate whether the model could map the unseen class samples to their prototypes as close as seen classes, i.e., could handle the projection domain shift problem properly.

According to the above analysis, we define the ScoreRDS index as follows. First of all, the distance from the CG of mapped features to the prototype of image class o, i.e., $Dist_o$, is calculated as follows.

$$Dist_o = \left\| \frac{1}{n} \sum_{k=1}^{n} \mathbf{W}^{\mathrm{T}} \mathbf{x}_{ok} - \mathbf{y}_o \right\|_2, \quad o \in \mathcal{Y}_s \cup \mathcal{Y}_u, \tag{9}$$

where n is the number of the image samples of class o.

We define the ScoreRDS as the quotient of the average distance between the unseen class CGs and their prototypes and the average distance between the seen class CGs and their prototypes in the semantic space. Formally,

$$ScoreRDS = \frac{\sum_{p}^{m_{unseen}} Dist_p / m_{unseen}}{\sum_{q}^{m_{seen}} Dist_q / m_{seen}}, \quad p \in \mathcal{Y}_u, \ q \in \mathcal{Y}_s, \tag{10}$$

where m_{unseen} and m_{seen} are the numbers of unseen and seen classes respectively.

As to different ZSL models, the smaller ScoreRDS indicates that the unseen class image samples are closer to their own class prototypes in the semantic embedding space compared to the model's performance on seen classes. Thus, a smaller ScoreRDS indicates the better capability of the model to resist the projection domain shift problem. But we should also notice that, the proposed ScoreRDS only considers the data distributions of image samples in the semantic embedding space. The performance of ScoreRDS can only evaluate the model's ability on alleviating the projection domain shift problem. It can not indicate the model's ability of zero-shot classification directly.

V. Datasets and Implementation Details

A. Datasets

In this paper, five zero-shot learning benchmarks are used to test the proposed DUET model. The statistics of the datasets are summarized in Table I.

- *Animals with Attributes (AwA):* The AwA dataset [6] contains 30,475 animal images of 50 classes with attribute annotations. Usually, 40 classes of the images are used for training and 10 classes are used for test. But in the standard split, 6 of 10 test classes are among the 1,000 classes in the ImageNet dataset which is often used to pre-train the CNN based visual feature extractor. In other words, using the standard class split and the CNN model pre-trained on ImageNet simultaneously will violate the task setting of ZSL. In this paper, the experiments of both ZSL task and GZSL task follow the new class split proposed by Xian *et al.* [12] which avoids the setting violation problem. The 85-dimension continuous attribute features are used as the semantic features.
- *Animals With Attributes 2 (AwA2):* The AwA2 dataset [12] is proposed in the same class structure and attribute annotations as the AwA dataset. It contains 37,322 new images to avoid the copyright resistance of the AwA dataset. In the experiments, the same class split and attribute features as AwA are used for the AwA2 dataset.
- *Caltech-UCSD Birds-200-2011 (CUB):* The CUB dataset [24] is a fine-grained image classification dataset with 200 bird classes. It contains 11,788 images and 312-dimension attribute annotations. 150 classes are used as the seen classes and the remaining 50 classes are used as the unseen classes. The proposed class split in [12] is used in this paper. The continuous class-level attribute features are used as semantic features of classes.
- *A-Pascal and a-Yahoo (aPY):* The aPY dataset [25] contains 20 categories of images from the Pascal VOC 2008 dataset and 12 categories of images from the Yahoo image search. It contains 15,399 images with 64-dimension attribute annotations. The original data split of the aPY dataset also violates the task setting of ZSL. The class split in [12] is used for the experiments of ZSL and GZSL in this paper.
- *Large-Scale Attribute Dataset (LAD):* The LAD dataset [26] is a recently proposed large-scale attribute dataset with 78,017 images and 359-dimension attribute annotations. It contains 230 image categories which are divided into five different super-classes. The super-classes include animals, fruits, vehicles, electronics and hairstyles. The experiments should be implemented on every super-class respectively. What's more, the LAD dataset provides five different seen/unseen class splits which all should be tested in the experiments. We should notice that the provided class splits do not guarantee that there is no overlap between the unseen classes and the 1,000 classes of the ImageNet dataset.

TABLE I
STATISTIC INFORMATION OF ZSL DATASETS

Dataset	Instances	Prototype Dimensions	Class Numbers (Seen/Unseen)
AwA	30,475	85	40/10
AwA2	37,322	85	40/10
CUB	11,788	312	150/50
aPY	15,399	64	20/12
LAD	78,017	359	184/46

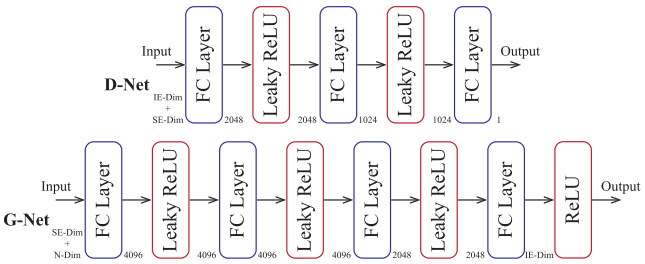


Fig. 6. Network structure of the UVG module. The input and output feature dimensions are marked on the left and right side of the *FC Layer* blocks. "IE-Dim", "SE-Dim" and "N-Dim" indicate the dimensions of visual features, semantic features and Gaussian noise respectively.

B. Implementation Details

1) Visual Feature Extractor of DET: The CNN visual feature extractor is based on the ResNet-101 [4] model in the proposed DUET model for a fair comparison with the previous work. The parameters in the model is pre-trained on the ImageNet dataset with 1,000 image categories. The 2048-dimensional visual features before the fully connected layer are used as the visual embedding representation. The images are resized to 224×224 as the inputs for all the five datasets.

2) Network Structure of UVG: The UVG module is composed of the generator (G-Net) and the discriminator (D-Net). The D-Net in the UVG module is a multilayer perceptron (MLP) with three FC layers. The Leaky ReLU layers are used as the nonlinear activations. In order to expand the generation space, we design a deeper MLP which has four FC layers as the G-Net. Similarly, three Leaky ReLU layers are implemented between the FC layers as nonlinear activations. Since the output features of G-Net should be non-negative, the same as the visual features extracted by the DET module, a ReLU layer is used as the final activation. Intuitively, the networks of D-Net and G-Net are shown in Figure 6.

3) Experimental Parameters: In the experiments, the dimension of the Gaussian noise $\mathbf{z}$ is set as 50 for the AwA, AwA2 and aPY datasets. Meanwhile, $\mathbf{z}$ is set as 300-dimension for the CUB and LAD dataset. All the noise features are sampled from the Gaussian distribution $\mathcal{N}(0, 0.1)$. In the training process of the UVG module, the G-Net is trained for 1 iteration after the D-Net is trained for every 5 iterations. In all experiments, we set the hyperparameter $\alpha = 0.5$ in Equation (3). For the coefficient of the gradient penalty in Equation (4), we set $\lambda = 10$ which is the common setting of the WGAN-GP method. The coefficient of the cross entropy classification loss, σ, is set as 1 in Equation (5) for all the experiments.

4) Evaluation Criteria: The multi-way classification accuracy, i.e., the average per-class top-1 classification accuracy is used for all the five datasets for the task setting of ZSL. For the task setting of GZSL, the multi-way classification accuracy results on seen classes and unseen classes, i.e., Acc_s and Acc_u, are reported respectively. Besides the accuracy results, the harmonic mean H is used to evaluate whether the model could achieve satisfactory performance on both seen and unseen classes. Formally,

$$H = \frac{2 * Acc_s * Acc_u}{Acc_s + Acc_u}. \tag{11}$$

TABLE II
QUANTITATIVE COMPARISONS OF AVERAGE PER-CLASS ZSL CLASSIFICATION ACCURACY (%)

Method	AwA	AwA2	CUB	aPY	LAD
DAP (2009) [6]	44.1	46.1	40.0	33.8	-
IAP (2009) [6]	35.9	35.9	24.0	36.6	-
ConSE (2013) [7]	45.6	44.5	34.3	26.9	31.4
DeViSE (2013) [8]	54.2	59.7	52.0	39.8	-
CMT (2013) [9]	39.5	37.9	34.6	28.0	-
ESZSL (2015) [18]	58.2	58.6	53.9	38.3	39.6
SJE (2015) [17]	65.6	61.9	53.9	32.9	49.9
SSE (2015) [35]	60.1	61.0	43.9	34.0	-
SynC (2016) [10]	54.0	46.6	55.6	23.9	48.0
ALE (2016) [11]	59.9	62.5	54.9	39.7	-
LatEm (2016) [36]	55.1	55.8	49.3	35.2	49.7
PSRZSL (2018) [37]	-	63.8	56.0	38.4	-
DeViSE* (2018)	66.9	-	60.3	-	-
ESZSL* (2018)	63.9	-	54.7	-	-
SJE* (2018)	66.9	-	58.4	-	-
ALE* (2018)	68.2	-	61.5	-	-
LatEm* (2018)	69.9	-	60.8	-	-
f-CLSWGAN (2018) [20]	68.2	-	57.3	-	-
SE-ZSL (2018) [38]	69.5	69.2	59.6	-	-
DET (Ours)	69.5	67.8	69.0	40.6	54.1
DUET (Ours)	**71.7**	**72.6**	**72.4**	**41.9**	**58.6**

* indicates that the model is strengthened with the f-CLSWGAN method [20].

VI. EXPERIMENTAL RESULTS

A. Comparisons With the State-of-the-Art Methods on ZSL

The quantitative comparisons of the proposed DUET model with the state-of-the-art methods on the task setting of ZSL are presented in Table II. The table is separated into three parts. In the upper part (from DAP [6] to PSRZSL [37]), we show the performance of the classic ZSL methods. For the convenience of description, we call them *classic methods* for short. In the middle part (form DeViSE* to SE-ZSL [38]), the performance of methods which use generative models to synthesize visual features of unseen classes is manifested. Similarly, we call these methods the *generative methods*. The lower part shows the performance of the proposed DUET model.

In order to avoid violating the task configuration of ZSL, the newly proposed class splits in [12] are implemented on AwA, AwA2, CUB and aPY datasets. The five class splits provided by [26] are used for LAD dataset. We should notice that the unseen classes of LAD have overlaps with the 1,000 classes of the ImageNet dataset. The experimental results of AwA, AwA2, CUB and aPY datasets, from DAP [6] to LatEm [36], are provided by [12] since the authors have tested previous ZSL methods on the new class split with the ResNet-101 [4] visual features. The experimental results of LAD dataset are provided by [26], in which the authors also use the ResNet model to extract visual features. The recently published PSRZSL [37] uses the ResNet-101 model and is tested on the new class split. As to the generative method, f-CLSWGAN [20] trains the generator in the WGAN-GP model to generate visual features of unseen classes. Then the generated visual features are used to train a linear softmax classifier, which turns the ZSL task into a supervised learning style. In f-CLSWGAN, the authors also use the learned generator

TABLE III
QUANTITATIVE EXPERIMENTAL RESULTS OF LAD DATASET

Method	Super-class					Average
	Animals	Fruits	Vehicles	Electronics	Hairstyles	
ConSE (2013) [7]	36.9	29.8	37.5	28.3	24.6	31.4
ESZSL (2015) [18]	50.2	37.2	45.8	32.8	31.8	39.6
SJE (2015) [17]	61.9	46.4	63.0	39.5	38.5	39.6
SynC (2016) [10]	61.6	51.4	54.9	43.0	29.1	48.0
LatEm (2016) [36]	63.9	44.2	60.9	40.7	38.5	49.7
DET (Ours)	69.4	50.1	66.1	44.6	40.0	54.1
DUET (Ours)	**73.6**	**55.4**	**70.3**	**47.3**	**46.6**	**58.6**

to enhance several classic methods. These models are trained with both the seen class images and the generated unseen class visual features. We demonstrate the enhanced performance of these methods with the indications of asterisk(*). Instead of using the GAN model, the recent SE-ZSL method [38] is built on a variational autoencoder. Similar to f-CLSWGAN, the generated image examples of SE-ZSL are also used to train the off-the-shelf classifier. For the evaluation of the AwA dataset, the SE-ZSL model utilizes the VGG-19 [2] visual features. On the AwA2 and CUB datasets, the SE-ZSL model uses ResNet visual features.

In the bottom part of Table II, the zero-shot classification results of the DET module and the whole DUET model are provided. For a fair comparison with previous methods, we also utilize the ResNet-101 CNN visual feature extractor in the DUET model.

As we can see from Table II, the performance of the DET module already surpasses the performance of all the previous classic methods on all the five datasets. What's more, the DET module also performs much better than all the generative methods on the CUB dataset (69.0% of DET to 61.5% of ALE*). On the AwA dataset, the DET module also attains the similar performance (69.5% of DET to 69.9% of LatEm*). The promising performance of the DET module suggests the important effect of the combined embedding transfer net and the end-to-end joint training process for the visual feature extractor and the embedding mapping function. In contrast, all the classic and generative methods in Table II use the fixed image extractor which lacks the feature adaptation for the specific ZSL tasks. Moreover, these methods just use singular linear or nonlinear mapping functions.

The performance of the whole DUET model achieves further improvements than DET's as demonstrated (2.2% on AwA, 4.8% on AwA2, 3.4% on CUB, 1.3% on aPY and 4.5% on LAD). The visual features of unseen classes generated by the UVG module contribute a lot to the alleviation of the projection domain shift problem.

According to the guidance of the LAD dataset [26], the ZSL models should be evaluated on each super-class separately. What's more, the experiments should be implemented on five different class splits and the average accuracy should be reported as the zero-shot classification result of the specific super-class. In Table II, we present the average accuracy of all super-classes as the general performance on LAD. In Table III, the experimental results of every super-class are demonstrated.

As presented in Table III, the performance of the LAD dataset is promoted by the proposed DUET model with a large margin. Since the visual feature extractors of all models in Table III are pre-trained on the ImageNet dataset which contains nearly 400 animal classes and more than 50 vehicle classes, the proper visual features can be extracted for the Animals and Vehicles super-classes. The models achieve relatively high performance on these two super-classes. The DET module achieves much better performance than the classic models on four super-classes (Animals, Vehicles, Electronics and Hairstyles), which demonstrates the superiority of the combined embedding transfer net and the end-to-end joint training. The entire DUET model achieves further improvement on the basis of the DET module. As to the Hairstyles super-class, the ImageNet dataset does not contain any visually or semantically similar image classes. Therefore, we can consider it as the most difficult super-class for zero-shot classification in the LAD dataset. On the Hairstyles super-class, the DUET model achieves the most significant improvement based on the DET module (6.6%). The improvement shows that the UVG module has effective generalization capability even for the difficult classes.

B. Comparisons With the State-of-the-Art Methods on GZSL

The quantitative comparisons of the proposed DUET model with the state-of-the-art methods on the task setting of GZSL are demonstrated in Table IV. Because the LAD dataset does not provide the class split for GZSL task, the remaining four datasets are used for the comparisons. The table is also separated into three parts as Table II. Compared to the classification accuracy on seen classes and unseen classes, i.e., Acc_s and Acc_u, the harmonic mean H is a more important criteria for GZSL, because we want to achieve a high accuracy on both seen and unseen classes in the task of GZSL. Only when Acc_s and Acc_u are high simultaneously, we can get a satisfying value of H.

As shown in Table IV, the proposed DUET model achieves considerable promotion than the state-of-the-art methods on the harmonic mean H (0.7 % on AwA, 0.6 % on AwA2, 3.4 % on CUB and 9.9 % on aPY). It can be observed that all the generative methods perform much better than classic methods on the H value. This phenomenon indicates that the synthesized visual features of unseen classes are extremely important and useful for ZSL and GZSL methods. As to the

TABLE IV
QUANTITATIVE COMPARISONS OF AVERAGE PER-CLASS GZSL CLASSIFICATION ACCURACY (%)

Method	AwA			AwA2			CUB			aPY		
	Acc_u	Acc_s	H	Acc_u	Acc_s	H	Acc_u	Acc_s	H	Acc_u	Acc_s	H
DAP (2009) [6]	0.0	88.7	0.0	0.0	84.7	0.0	1.7	67.9	3.3	4.8	78.3	9.0
IAP (2009) [6]	2.1	78.2	4.1	0.9	87.6	1.8	0.2	72.8	0.4	5.7	65.6	10.4
ConSE (2013) [7]	0.4	88.6	0.8	0.5	90.6	1.0	4.6	72.2	3.1	0.0	91.2	0.0
DeViSE (2013) [8]	13.4	68.7	22.4	17.1	74.7	27.8	23.8	53.0	32.8	4.9	76.9	9.2
CMT (2013) [9]	0.9	87.6	1.8	0.5	90.0	1.0	7.2	49.8	12.6	1.4	85.2	2.8
ESZSL (2015) [18]	6.6	75.6	12.1	5.9	77.8	11.0	12.6	63.8	21.0	2.4	70.1	4.6
SJE (2015) [17]	11.3	74.6	19.6	8.0	73.9	14.4	23.5	59.2	33.6	3.7	55.7	6.9
SSE (2015) [35]	7.0	80.5	12.9	8.1	82.5	14.8	8.5	46.9	14.4	0.2	**78.9**	0.4
SynC (2016) [10]	8.9	87.3	16.2	10.0	90.5	18.0	11.5	70.9	19.8	7.4	66.3	13.3
ALE (2016) [11]	16.8	76.1	27.5	14.0	81.8	23.9	23.7	62.8	34.4	4.6	73.7	8.7
LatEm (2016) [36]	7.3	71.7	13.3	11.5	77.3	20.0	15.2	57.3	24.0	0.1	73.0	0.2
PSRZSL (2018) [37]	-	-	-	20.7	73.8	32.3	24.6	54.3	33.9	13.5	51.4	21.4
DeViSE* (2018)	35.0	62.8	45.0	-	-	-	52.2	42.4	46.7	-	-	-
ESZSL* (2018)	31.1	72.8	43.6	-	-	-	36.8	50.9	43.2	-	-	-
SJE* (2018)	37.9	70.1	49.2	-	-	-	48.1	37.4	42.1	-	-	-
ALE* (2018)	47.6	57.2	52.0	-	-	-	40.2	59.3	47.9	-	-	-
LatEm* (2018)	33.0	61.5	43.0	-	-	-	**53.6**	39.2	45.3	-	-	-
f-CLSWGAN (2018) [20]	**57.9**	61.4	59.6	-	-	-	43.7	57.7	49.7	-	-	-
SE-ZSL (2018) [38]	56.3	67.8	61.5	**58.3**	68.1	62.8	41.5	53.3	46.7	-	-	-
DUET (Ours)	47.5	**90.1**	**62.2**	48.2	**90.2**	**63.4**	39.7	**80.1**	**53.1**	**21.8**	55.6	**31.3**

* indicates that the model is strengthened with the f-CLSWGAN method [20].

TABLE V
EFFECTS OF THE COMBINED EMBEDDING TRANSFER NET ON ZSL CLASSIFICATION (%)

Mapping Function	AwA2		CUB	
	DET	DUET	DET	DUET
Linear	67.0	70.8	67.6	70.1
Nonlinear	40.4	67.8	60.0	66.4
Combined	**67.8**	**72.6**	**69.0**	**72.4**

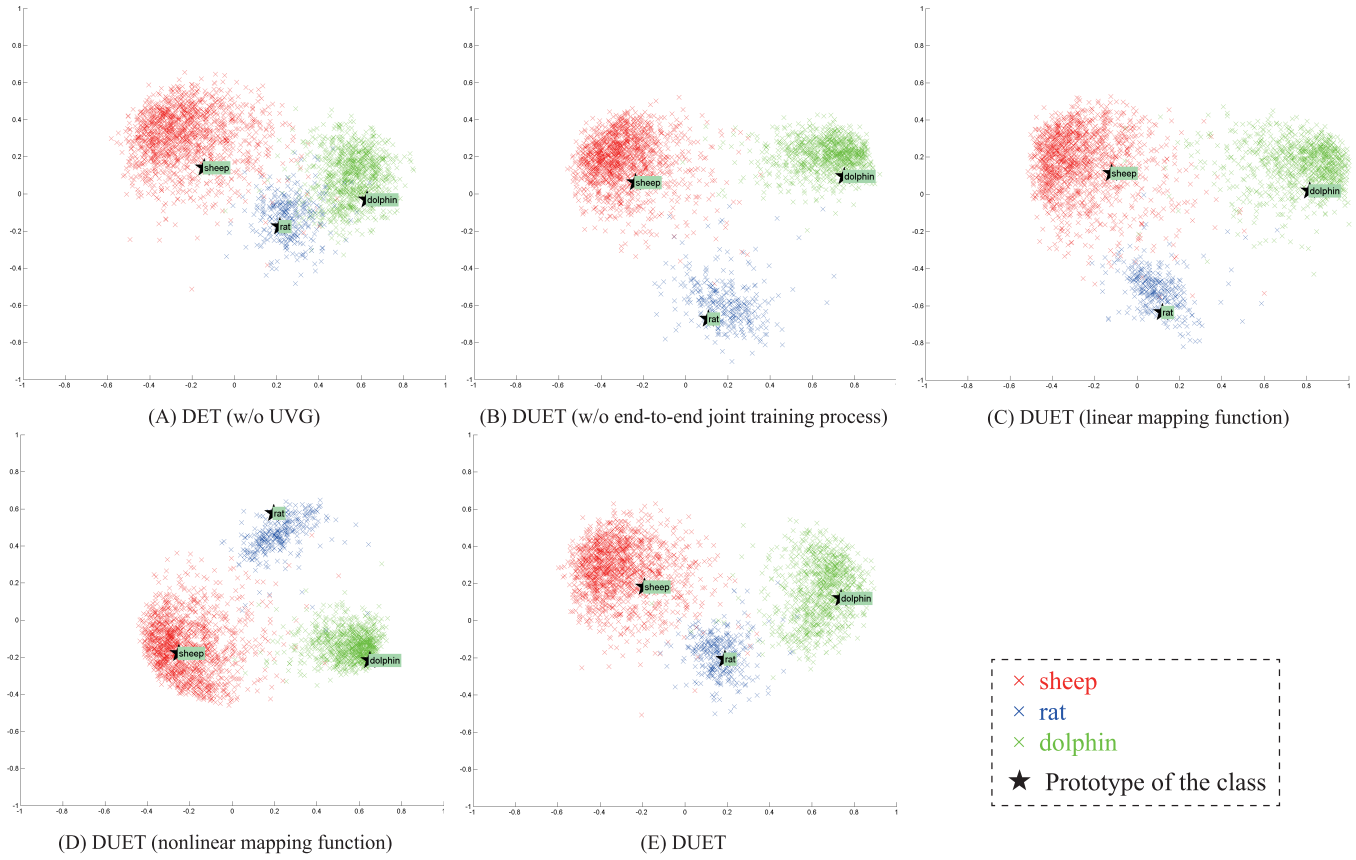
performance of the DUET model, the classification accuracy on seen classes is much better than the accuracy on unseen classes. The reason is that although the combined embedding transfer net is trained with seen class visual features and synthesized unseen class visual features, the CNN visual feature extractor in the DET module is still just trained with the seen class images. Therefore, the performance on seen classes is much better than the performance on unseen classes. It is also the direction to further improve the DUET model.

C. Effects of the Combined Embedding Transfer Net

In order to quantitatively evaluate the contribution of the proposed combined embedding transfer net in the DUET model, we evaluate the model with the linear, nonlinear and the combined embedding mapping functions respectively on the AwA2 and the CUB datasets. In the experiments, one FC layer is used to represent the linear mapping function. As to the nonlinear mapping function, two FC layers and Leaky ReLU nonlinear activations are implemented. The combined embedding mapping function is realized with the proposed combined embedding transfer net, as shown in Figure 4.

As demonstrated in Table V, the combined mapping function achieves the best performance in both DET module and DUET model on the AwA2 and the CUB datasets. The performance of the linear mapping function is close to the proposed combined mapping function. The close performance indicates that the linear mapping function is relatively easy to optimize and achieve promising ZSL classification results. This is also a reason that most of the previous mapping-based ZSL methods choose the linear mapping function to connect the visual and semantic embedding space. Compared to the linear mapping function, the performance of DET with nonlinear mapping function is much inferior. The phenomenon denotes it is hard for the nonlinear mapping function to infer unseen classes with the learning of seen classes. But when the UVG module is implemented in the model, the performance gets promotion rapidly (27.4% on AwA2 and 6.4% on CUB). The performance gets close to the linear mapping function. The promotion illustrates that the nonlinear mapping function has huge information capacity and can learn visual knowledge from the generated unseen class visual features effectively. It is in compliance with the awareness that the nonlinear model has better capability of learning and generalization. As to the combined embedding mapping function, the performance improvements from the DET module to the whole DUET model (4.8% on AwA2 and 3.4% on CUB) are larger than the improvements of linear mapping function (3.8% on AwA2 and 2.5% on CUB). It also demonstrates the effect of the nonlinearity for learning from the synthesized unseen class features and knowledge generalization.

Consequently, the combined embedding transfer net can associate the superiority of linear and nonlinear embedding mapping functions. On one hand, the combined embedding transfer net is easy to get convergence and obtain the promising performance. On the other hand, it can closely cooperate with generative models with better generalization capability and keep more knowledge during the embedding transfer process. When the generated unseen class visual features are



(A) DET (w/o UVG) (B) DUET (w/o end-to-end joint training process) (C) DUET (linear mapping function)

(D) DUET (nonlinear mapping function) (E) DUET

Fig. 7. Qualitative analysis of the DUET model on alleviating projection domain shift.

introduced into the training process, the combined embedding transfer net learns more information details and obtains better ZSL classification performance.

D. Ablation Studies of DUET on the Projection Domain Shift Problem

In this section, we demonstrate the effect of the proposed DUET model on alleviating the projection domain shift problem by qualitative visualization and quantitative ScoreRDS evaluation. As introduced, there are two backbones, i.e., the DET module and the UVG module in the DUET model. There are three key points in the two backbones. The first one is the unseen class visual feature generation in the UVG module. The second one is the combined embedding transfer net in the DET module. Last but not the least, the end-to-end joint training process in the DET module is the third key point of the DUET model. The qualitative and quantitative analysis of every key point's effects on domain shift will explain why the DUET model achieves the performance better than the state-of-the-art methods. For the convenience of description, the four model variants are proposed as follows.

- Variant A The DUET model without the UVG module, i.e., the DET module.
- Variant B The DUET model without the end-to-end joint training process.
- Variant C The DUET model with only the linear mapping function in the embedding transfer process.
- Variant D The DUET model with only the nonlinear mapping function in the embedding transfer process.

Model variants A – D are implemented to compare with the intact DUET model respectively.

1) Qualitative Visualization: The projection domain shift problem is that the mapped visual samples are biased from their corresponding class prototypes in the semantic embedding space. In Figure 7, we demonstrate the effect of the DUET model on alleviating the domain shift problem qualitatively. In order to make the demonstration clear, three unseen classes of the AwA2 dataset, *sheep*, *rat* and *dolphin*, are used for the visualization. The attribute features of the classes are used as the prototypes in the semantic embedding space. The image samples are mapped into the semantic embedding space by the four model variants of the DUET model and the model itself.

The visualization results of the variant models are shown in Figure 7 (A) – (D) respectively. The result of the whole DUET model is shown in Figure 7 (E). The principal components analysis (PCA) [39] is used for dimension reduction. The black pentacles represent the prototypes of the classes. The label name of the class is marked beside the pentacle. The mapped image samples are marked by crosses of different colors.

As shown in the figure, the mapping results of classes *rat* and *dolphin* with the model variants A – D clearly derivate from the class prototypes, where the prototypes are all located

TABLE VI
SCORERDS AND ZSL CLASSIFICATION ACCURACY (%) ON AWA2 AND CUB

Model	AwA2		CUB	
	ScoreRDS	*Acc*	ScoreRDS	*Acc*
Variant A (DET, w/o UVG)	0.854	67.8	1.029	69.0
Variant B (w/o joint training process)	0.913	68.1	1.058	62.0
Variant C (linear mapping function)	0.831	69.0	0.993	70.1
Variant D (nonlinear mapping function)	0.882	67.8	1.024	66.4
DUET model	**0.800**	72.6	**0.981**	72.4

on the edge of the image sample clusters of the corresponding image classes. As to the class *sheep*, the variant D performs better than the other model variants. As to the proposed DUET model (Figure 7 (E)), the projection domain shift problem is alleviated significantly, where the class prototypes of classes *rat* and *dolphin* are situated close to the center of each image sample cluster. And the mapping result of class *sheep* is also better than model variants A, B and C.

2) Quantitative ScoreRDS Evaluation: Besides the qualitative visualization, the quantitative evaluation with the proposed ScoreRDS is implemented in this section. As demonstrated, the ScoreRDS is proposed to evaluate the effect of the model on alleviating the projection domain shift problem. The smaller ScoreRDS indicates the model has superior capability on alleviating domain shift than others. In Table VI, the ScoreRDS performance of the DUET model and its four variants on the AwA2 and CUB datasets is presented. For the convenience of reference, the ZSL classification accuracy of the DUET model and its four variants is also presented in Table VI.

As demonstrated in Table VI, the proposed DUET model achieves the lowest ScoreRDS on both the AwA2 and CUB datasets, which indicates that the DUET model has the superior capability to alleviate the domain shift problem than other model variants. It confirms that all the three key points in the DUET model contribute to reduce the effects of the projection domain shift problem. On both the AwA2 and the CUB datasets, the variant B achieves the highest score which implies that the end-to-end joint training process for the visual feature extractor and the embedding transfer net plays the most important role in the DUET model for the domain shift problem. According to the results in Table VI, the three key points arranged in a descending order of the importance for alleviating the domain shift problem are the joint training process, the combined embedding transfer net and the UVG module on the AwA2 dataset. On the CUB dataset, the order is the joint training process, the UVG module and the combined embedding transfer net. If we focus on the combined embedding transfer net, i.e., the performance of variant C and variant D in Table VI, the results reveal that the linear mapping function contributes more than the nonlinear mapping function for the projection domain shift problem on the two datasets. From the table, we also find that the classification accuracy is approximately inversely proportional to the ScoreRDS, which also indicates the rationality and validity of the proposed DUET model for alleviating the projection domain shift problem.

By the quantitative analysis of ScoreRDS, we prove that all the key points of the DUET model are effective to alleviating the projection domain shift problem.

VII. CONCLUSION

In this paper, a deep unbiased embedding transfer (DUET) model is proposed for the zero-shot learning task. The DUET model is composed of the deep embedding transfer (DET) module and the unseen visual feature generation (UVG) module. In the DET module, a novel combined embedding transfer net is proposed. Besides that, an end-to-end joint training process is implemented to train the CNN visual feature extractor and the embedding transfer net simultaneously. In the UVG module, synthesized visual features of unseen classes are generated to ease the projection domain shift problem of the zero-shot learning task. After the training of the DUET model, an unbiased embedding transfer can be obtained for the ZSL task. What's more, the ScoreRDS is proposed to quantitatively evaluate the model's effect on the projection domain shift problem. By the evaluation, all the key points in the proposed model contribute to alleviating the projection domain shift problem. The proposed DUET model achieves the state-of-the-art performance over five zero-shot learning benchmarks.

In the DUET model, the linear and nonlinear mapping functions are combined in a simple manner. More network structures of the combined embedding transfer net will be explored in the future work.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[3] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[5] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[6] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 951–958.
[7] M. Norouzi *et al.*, "Zero-shot learning by convex combination of semantic embeddings," Dec. 2013, *arXiv:1312.5650*. [Online]. Available: https://arxiv.org/abs/1312.5650
[8] A. Frome *et al.*, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[9] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 935–943.

[10] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5327–5336.

[11] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1425–1438, Jul. 2016.

[12] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," Jul. 2017, *arXiv:1707.00600*. [Online]. Available: https://arxiv.org/abs/1707.00600

[13] P. Morgado and N. Vasconcelos, "Semantically consistent regularization for zero-shot recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2037–2046.

[14] F. Shen, X. Zhou, J. Yu, Y. Yang, L. Liu, and H. T. Shen, "Scalable zero-shot learning via binary visual-semantic embeddings," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3662–3674, Jul. 2019.

[15] L. Niu, J. Cai, A. Veeraraghavan, and L. Zhang, "Zero-shot learning via category-specific visual-semantic mapping and label refinement," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 965–979, Feb. 2019.

[16] Y. Guo, G. Ding, J. Han, and Y. Gao, "Zero-shot learning with transferred samples," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3277–3290, Jul. 2017.

[17] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2927–2936.

[18] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2152–2161.

[19] Y. Fu, T. M. Hospedales, T. T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2332–2345, Nov. 2015.

[20] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5542–5551.

[21] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[22] M. Bucher, S. Herbin, and F. Jurie, "Generating visual representations for zero-shot classification," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2666–2673.

[23] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, "A generative adversarial approach for zero-shot learning from noisy texts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1004–1013.

[24] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA: Tech. Rep. CNS-TR-2011-001, 2011.

[25] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1778–1785.

[26] B. Zhao, Y. Fu, R. Liang, J. Wu, Y. Wang, and Y. Wang, "A large-scale attribute dataset for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1–11.

[27] Z. Jia, J. Zhang, K. Huang, and T. Tan, "Encyclopedia enhanced semantic embedding for zero-shot learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1287–1291.

[28] 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 *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 52–68.

[29] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[30] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," Jan. 2017, *arXiv:1701.07875*. [Online]. Available: https://arxiv.org/abs/1701.07875

[31] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2642–2651.

[32] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong, "Transductive multi-view embedding for zero-shot recognition and annotation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 584–599.

[33] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Unsupervised domain adaptation for zero-shot learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2452–2460.

[34] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4447–4456.

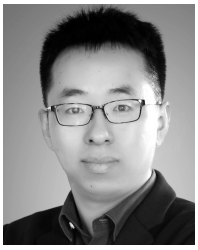
[35] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4166–4174.

[36] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 69–77.

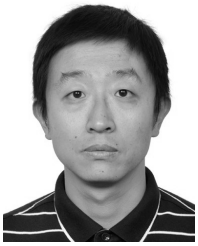
[37] Y. Annadani and S. Biswas, "Preserving semantic relations for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7603–7612.

[38] V. K. Verma, G. Arora, A. Mishra, and P. Rai, "Generalized zero-shot learning via synthesized examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4281–4289.

[39] I. Jolliffe, *Principal Component Analysis*. Berlin, Germany: Springer, 2011.



Zhen Jia received the B.Eng. degree in electronic engineering from Yunnan University, Kunming, China, in 2013. He is currently pursuing the Ph.D. degree with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include pattern recognition and computer vision. He currently focuseing on zero-shot and few-shot learning.



Zhang Zhang received the B.Sc. degree in computer science and technology from the Hebei University of Technology, Tianjin, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2008. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include activity recognition, video surveillance, and time series analysis. He has published a number of articles at top venues, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, CVPR, and ECCV.

Liang Wang (F'19) received the B.Eng. and M.Eng. degrees from Anhui University, in 1997 and 2000, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he was a Research Assistant at Imperial College London, U.K., and also at Monash University, Australia, a Research Fellow at the University of Melbourne, Australia, and a Lecturer at the University of Bath, U.K., respectively. He is currently a Full Professor of the Hundred Talents Program with the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He is a fellow of the IAPR. He has widely published in highly ranked international journals such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the IEEE TRANSACTIONS ON IMAGE PROCESSING, and leading international conferences such as CVPR, ICCV, and ICDM.

Caifeng Shan (SM'13) received the Ph.D. degree in computer vision from Queen Mary, University of London. His research interests include computer vision, pattern recognition, image and video analysis, machine learning, bio-medical imaging, and related applications. He has authored over 150 scientific articles and patent applications. He has been an Associate Editor and the Guest Editor of many scientific journals, including the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (T-CSVT), the IEEE TRANSACTIONS ON MULTIMEDIA (T-MM), and the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS (J-BHI).

Tieniu Tan (F'04) received the B.Sc. degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and the M.Sc. and Ph.D. degrees in electronic engineering from Imperial College London, U.K., in 1986 and 1989, respectively. He is currently a Professor with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China. His research interests include biometrics, image and video understanding, information hiding, and information forensics. He is a fellow of the CAS, TWAS, BAS, IAPR, and the U.K. Royal Academy of Engineering, and the Past President of the IEEE Biometrics Council.