This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2867870, IEEE Transactions on Pattern Analysis and Machine Intelligence

Generative Zero-Shot Learning via Low-Rank Embedded Semantic Dictionary

Zhengming Ding, Ming Shao, *Member, IEEE*, and Yun Fu, *Senior Member, IEEE*

Abstract—Zero-shot learning for visual recognition, which approaches recognizing unseen categories through a shared visual-semantic function learned on the seen categories and is expected to well adapt to unseen categories, has received considerable research attention most recently. However, the semantic gap between discriminant visual features and their underlying semantics is still the biggest obstacle, because there usually exists domain disparity across the seen and unseen classes. To deal with this challenge, we design two-stage generative adversarial networks to enhance the generalizability of semantic dictionary through low-rank embedding for zero-shot learning. In detail, we formulate a novel framework to simultaneously seek a two-stage generative model and a semantic dictionary to connect visual features with their semantic representations under a low-rank embedding. Our first-stage generative model is able to augment more semantic features for the unseen classes, which are then used to generate more discriminant visual features in the second stage, to expand the seen visual feature space. Therefore, we will be able to seek a better semantic dictionary to constitute the latent basis for the unseen classes based on the augmented semantic and visual data. Finally, our approach could capture a variety of visual characteristics from seen classes that are "ready-to-use" for unknown classes. Extensive experiments on four zero-shot benchmarks demonstrate that our proposed algorithm outperforms the state-of-the-art zero-shot algorithms.

Index Terms—Generative Adversarial Network, Low-Rank Embedding, Semantic Dictionary, Zero-Shot Learning

1 Introduction

Recent years have witnessed the tremendous progress of several computer vision and machine learning tasks, for example, object recognition, and fine-grained classification, as well as the recent advance of deep learning techniques. However, the truth behind is popular learning schemes in this line usually require large amount of labeled data for model training, e.g., MS-Celeb-1M with millions of labeled face images. This would be affordable when facing common objects, but the objects "in the wild" follow in a long-tailed distribution, and thus the unusual ones do not occur frequently enough. In reality, new concepts, especially those from the Internet, emerge everyday, for which collecting and labeling sufficiently large training sets could be difficult and expensive [1]. Under this situation, training effective classification systems for the uncommon objects without the labeled data becomes a practically important problem and thus, has drawn considerable research attention from the machine learning and computer vision communities.

Recently, zero-shot learning (ZSL) has attracted significant attention because of its promising performance. ZSL is motivated with the learning mechanism of human cognitive and attempts to identify new objects that are unobserved during the training stage [2], [3], [4], [5], [6], [7], [8], [9], [10], [10], [11], [12]. Humans are able to recognize approximate 30,000 basic objects and many more subordinate ones. Inspiringly, they can recognize new objects given an attribute description. For instance, one can identify a new species of animal after being told what it looks like and how it is similar to or different from other known animals. The reason is simple: humans are able to explore the relationship among different objects via secondary knowledge, and adapt the information from seen classes to unseen ones. Likewise, ZSL attempts to capture the intrinsic semantic relationship between observed and unobserved categories. Generally,

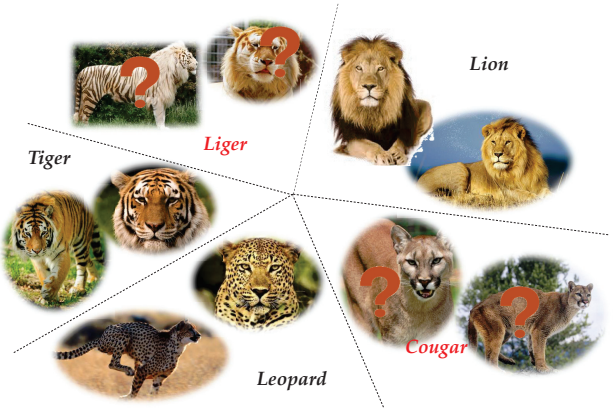


Fig. 1. Illustration of visual features from seen classes and unseen classes. Here we have three seen classes, i.e., Tiger, Lion and Leopard, as well as two unseen class, i.e., cougar and liger. We could notice that the distributions of different categories across seen and unseen data are intertwined.

- *Z. M. Ding is with the Department of Computer, Information and Technology, Indiana University-Purdue University Indianapolis, 420 University Blvd Indianapolis, IN 46202, USA*
  E-mail: zd2@iu.edu
- *M. Shao is with the Department of Computer and Information Science, University of Massachusetts Dartmouth, MA, 02747, USA.*
  E-mail: mshao@umassd.edu
- *Y. Fu is with the Department of Electrical and Computer Engineering and the College of Computer and Information Science, Northeastern University, Boston, MA, 02115 USA.*
  E-mail: yunfu@ece.neu.edu.

three fundamental concepts are required: (1) visual features conveying non-trivial yet informative knowledge; (2) semantic features reflecting the relationship among various categories; (3) learning model properly connecting visual features and their underlying semantics.

ZSL is appealing in simulating the human cognitive process, however, it suffers two degenerating aspects. First of all, the distribution of instances in visual feature space is usually different from that of their underlying semantic space, because visual features in different aspects may convey the same meaning. Such a semantic mismatch ruins the knowledge adaptation from the observed classes to unobserved ones. Secondly, the "hubness" [13] is recently recognized to result in the poor performance, that is exacerbated with a shortage of training samples for unknown categories in visual domain. Thus, the domain mismatch challenge [14] raises an issue for ZSL. The crux of ZSL lies in learning a compatibility function between the visual feature and its semantic representation. On one hand, complex functions are flexible but at risk of over-fitting to the seen categories and adapting poorly to the unseen ones. On the other hand, simple ones would lead to poor classification performance on the seen categories, and will unlikely adapt well on the unseen ones either.

Current ZSL approaches typically assume a shared semantic embedding space, in which both the feature space and the class label spaces of the seen/unseen classes lie in [2], [3], [4], [5], [6], [7], [8], [9]. In general, there are three lines of approaches as follows: (1) direct mapping [15], [16] including semantic coding methods [2], [11], [17]; (2) parameter mapping [1], [18]; and (3) common space learning [3], [19]. However, most ZSL approaches fail to build a better semantic-visual function, since they only have access to the seen data, subject to domain shifts with the unseen data in the test stage.

In this paper, we explore the idea of generative adversarial networks (GANs) to solve zero-shot learning challenge. Generally, seen and unseen classes have different distributions in the view of visual feature space; however, different categories in seen and unseen classes are lying in disjunction distributions (Figure 1). Therefore, we manage to synthesize more semantic features to cover unseen classes, and then further augment visual features from seen classes to expand the visual space for unseen classes. In this way, we could build a more effective visual-semantic relationship to better fight off zero-shot learning. To that end, we design a novel Generative Semantic Dictionary Learning (GSDL) with two-stage generations for zero-shot challenge under a low-rank embedding framework (Figure 2). Our primary assumption is that the latent semantic dictionary for unseen categories should carry the majority shared information with that for seen categories[1], which can be identified in the low-rank embedding space. Additionally, synthesized semantic and visual features generated from seen data will allow to recover a better latent semantic dictionary for the unseen data.

This work is the extension of our previous conference work [17] under the same basic spirit, which is the low-rank

1. Both seen and unseen categories in our work have lots of common semantics.

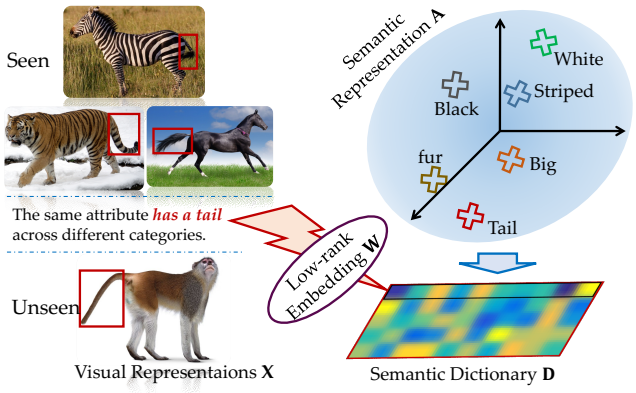


Fig. 2. Illustration of our designed basic semantic-visual framework, in which low-rank embedding W maps visual features X into a new space, and hence similar features, e.g., "has a tail", would gather together. At the same time, a semantic dictionary D is learned through the reconstruction $WX \approx DA$ to link visual features and their semantics. To boost the generalizability of semantic dictionary, we explore two-stage generative adversarial networks (shown in Figure 3) to synthesize more semantic and visual features to cover the feature space of unseen classes. Therefore, adaptive semantic dictionary could constitute the latent basis for the unseen categories.

embedded semantic dictionary to build semantic-visual relationship. In our previous work, we exploited the ensemble strategy to learn multiple semantic dictionaries, targeting at covering the space of unseen classes. In this work, however, we explore the idea of generative adversarial networks to synthesize more semantic and visual features based on seen classes to cover the space for unseen classes. With more augmented data, we could enhance the generalizability of semantic-visual function for unseen classes, which would have a better chance to capture the information for unseen classes. In other word, our current work explores an explicit data augmentation strategy with generative model to improve the generalization ability of semantic dictionary instead of our previous ensemble method. Furthermore, we adopt a weighted rank minimization to replace the direct rank minimization on our previous embedding matrix, so that we could capture more intrinsic structure information from visual and semantic features during the low-rank embedding learning.

Finally, we summarize our contributions in three folds:

- First, a semantic dictionary plays a role to link visual features and semantic representations under a reconstruction scheme. Then, we exploit a low-rank embedding to adapt the intrinsic and shared knowledge from the seen classes. Thus, a better latent semantic dictionary for the unseen classes can be obtained.
- Second, two-stage generative adversarial networks are exploited to synthesize more semantic features and visual features from seen classes. Thus, we are more likely to cover the unseen classes by spanning the semantic and visual space of seen classes. Specifically, an effective semantic dictionary will be optimized to complete the latent semantic dictionary and mitigate the distribution shift between seen and unseen data.

- Finally, we explore a graph-regularized term to involve more intrinsic structure in low-rank embedding learning. In addition, we formulate a novel low-rank re-framing approach to overcome the current sparse singular values issues[2] to secure a better low-rank embedding.

2 Related Work

In this section, we briefly discuss two related works, i.e., zero-shot learning and generative adversarial networks, and highlight the differences across our model and the existing ones.

2.1 Zero-shot learning

The goal of Zero-shot learning (ZSL) is to learn models from visual concepts without any test information. As visual information from those test categories is not available during learning process, ZSL needs external information to compensate for the missing visual knowledge. Attribute-based descriptions are the most widely-used characteristics shared among different categories [11], [15], [21], [22], [23], [24], [25], which can be treated as a special case of multi-view learning [19], [26]. Provided the low-level visual image features and their correspondent high-level semantics, the core issue for ZSL turns to adapt knowledge from the observed categories to unobserved ones [2], [3], [4], [5], [6], [7]. According to the strategies for semantic gaps, there are three lines of ZSL models in general.

First, direct mapping is proposed to build a mapping function from visual features to their semantics [2], [15]. In this fashion, Direct/Indirect Attribute Prediction were first developed to explore the hidden layer of attributes as variables to decouple the images from the layer of labels [16]. Later on, Gan et al. presented a representation transformation in visual space to enhance the attribute-level discriminability for attribute annotation [27].

Second, shared space learning attempts to seek new spaces in which visual features and semantic representations enjoy the maximum similarities for within-class samples. The learned common space is either interpretable or latent. For example, Zhang et al. exploited a probabilistic framework to seek joint similarity latent embedding in which both visual and semantic embedding along with a class-independent similarity measure were obtained simultaneously [3].

Finally, parameter mapping manages to estimate model parameters for unseen classes by "tuning" model parameters learned from observed classes. Essentially, it discovers the inter-class relationship across observed and unseen classes in semantic space [1], [18]. Following this, Mensink et al. explored co-occurrences statistics of visual concepts in images for knowledge adaptation and proposed a novel classifier with the co-occurrences [18].

2. Differently, Hu et al. [20] proposed a truncated trace norm via minimizing the sum of r-smallest singular values, in order to reduce the effect of large singular values. However, it is an l_1-minimization problem by minimizing the sum of r-smallest singular values, which leads to a sparse solution, i.e., several r-smallest singular values tend to be zero, but the left may remain large values.

However, all these algorithms pay limited attention to discriminative knowledge in the unseen categories given high intra-class variability, and may easily ignore the shared semantics among different domains. Our designed model belongs to the direct mapping scheme and share certain spirits with the pioneering work of "dictionary learning + sparse coding" [2], [28]. Notably, a complex semantic-visual function would hurt the generalization performance given the domain shift. Differently, we explore the generative model in zero-shot learning to synthesize more data to enhance the semantic-visual function. Specifically, we build a two-stage generative adversarial network by jointly optimizing low-rank embedding and semantic dictionary to uncover common discriminative features between seen and unseen categories.

2.2 Generative Adversarial Networks

Generative adversarial networks (GANs) [29] consist of two players trained in opposition to one another. The generator $G(\cdot)$ inputs with a random noise vector z and outputs a fake image $G(z)$; while the discriminator $D(\cdot)$ inputs either a real image or a fake image from the generator, then outputs a probability distribution over possible image domains. The discriminator is learned by maximizing the log-likelihood it assigns to the right domain.

The vanilla GANs could be enhanced with side information. Conditional GANs are to provide the generator and discriminator with class labels or latent information [30], [31]. Class conditional synthesis can significantly enhance the quality of generated data [32]. Richer auxiliary knowledge could further improve sample quality, e.g, image captions or bounding box localizations [33]. Different from feeding side information to the discriminator, one can learn the discriminator with reconstructing external information. This is done by equipping the discriminator with an auxiliary decoder that outputs the class label for the training data [34], [35] or a subset of the latent variables from which the samples are synthesized [36]. It is well-known to improve performance on the original task by forcing a model to perform additional tasks [37]. In addition, an auxiliary decoder could leverage pre-trained discriminators (e.g., image classifiers) to further improve the synthesized images' quality [38]. Chen et al. designed a new structure-aware convolutional network to implicitly consider such priors during the deep network training [39].

3 The Proposed Algorithm

In this section, we will propose our novel low-rank embedded semantic dictionary learning stacked with two-stage generative adversarial nets.

3.1 Preliminary & Motivation

In ZSL scenario, we are able to access to many "seen" classes with visual features and semantic representations for training, while evaluating on "unseen" classes which have no identity overlap with seen classes. Assume there exist C_s seen classes with n labeled samples $\mathcal{S} = \{X, A, y\}$ and C_u unseen classes with n_u unlabeled samples

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2867870, IEEE Transactions on Pattern Analysis and Machine Intelligence

$\mathcal{U} = \{X_u, A_u, y_u\}$. Each instance is represented as a d-dimensional visual feature. Thus, we have $X \in \mathbb{R}^{d\times n}$ and $X_u \in \mathbb{R}^{d\times n_u}$. $y \in \mathbb{R}^n$ and $y_u \in \mathbb{R}^{n_u}$ are class label vectors for the seen and unseen data, respectively. In ZSL, the seen and unseen classes are disjoint, i.e., $y \cap y_u = \emptyset$. $A \in \mathbb{R}^{m\times n}$ and $A_u \in \mathbb{R}^{m\times n_u}$ are the m-dimensional semantic representations of instances in the seen and unseen datasets, respectively. For the seen classes, *a* is provided since each visual sample *x* for seen classes is annotated by either a binary attribute vector or a continuous semantic vector denoting its corresponding class label *y*.

ZSL is challenging due to the unseen test samples and labels during the model learning. Hence, it is essential to establish the visual-semantic relationship based on the seen classes. However, seen classes and unseen classes always have different distributions in the view of visual features, and thus cause the domain shift between seen and unseen classes. The key for ZSL is how to well address the domain mismatch issue in the visual feature space. Recently, semantic representation gradually plays a bridge role to link seen and unseen classes, and one of typical methods is *visual attribute features*. Generally, different combinations of attributes could generate different semantic representation for various classes. The same attribute would link to different visual features in the same class due to the intra-class variance. In this sense, if we could synthesize more semantic features as well as visual features of the seen classes, we could expand the space of seen classes and maximize the overlap between seen and unseen classes. So the domain gap between them tends to be alleviated.

3.2 Low-Rank Embedded Semantic Dictionary

Known from previous works [8], [12], [25], [40], the seen data X and unseen data X_u are drawn from different feature spaces, but A and A_u may have more similar semantics. For instance, in attribute-based description, both seen and unseen data could be represented with pre-defined attributes with various weights, for example, binary and continuous values. The intuition behind ZSL is that the classifier can capture the relationship across the visual-input space and the individual semantic feature space [21]. Inspired by Kodirov et al. recent work [2] that considers semantic representation A as the encoded coefficients of X based on a semantic dictionary, we present a novel embedded semantic dictionary learning formula that explores the merits from semantic dictionary and discriminative embedding:

$$\min_{W,D} \|WX - DA\|_{\mathrm{F}}^2 + \alpha f(W) \ \text{ s.t. } \ \|d_j\|_2^2 \le 1, \forall j, \qquad (1)$$

in which α is the trade-off parameter, $\|\cdot\|_{\mathrm{F}}$ means the Frobenius norm, and $d_j \in \mathbb{R}^d$ is the j-th atom of semantic dictionary $D \in \mathbb{R}^{d\times m}$. $f(W)$ denotes a regularizer on W.

We elaborate the key difference and motivations in the following paragraphs. First of all, considering the scenario in which samples from a class are sampled over a complex manifold, e.g., crescent manifold, using mean vector as the prototype or exemplar of each class [40] is less reasonable. Secondly, in the test phase, the unseen categories do not match any classes from the training data, and will not be able to obtain corresponding discriminant information from the learned feature space. To that end, we instead concentrate on the semantic representations, and explore manifold learning by assuming that if the semantic representations of some instances after embedding lie on the same local manifold structure, they most likely have the same class label.

For that purpose, a graph regularizer is utilized to preserve the locality information for the embedded visual data as follows:

$$\begin{aligned} \min_{W,D} \|WX - DA\|_{\mathrm{F}}^2 + \alpha \mathrm{tr}(WXLX^\top W^\top) \\ \text{s.t. } \ \|d_j\|_2^2 \le 1, \forall j, \end{aligned} \qquad (2)$$

in which $\mathrm{tr}(\cdot)$ represents the trace operator of a matrix. Specifically, we adopt a spectral Dual-Graph approach [12]: a concatenation of two supervised graphs to jointly preserve the data structures from X and A, and induce a novel graph Laplacian. We build k-nn graphs for both visual and semantic features, and adopt the cosine similarity to define the weights (S_x and S_a). Thus, we achieve weights for dual-graph as $S = \frac{S_x + S_a}{2}$. L is the Laplacian graph calculated as $L = S_d - S$ (S_d is the degree matrix of S with its i-th diagonal element as $\sum_j S_{ij}$.

We further decompose $L = U\Sigma U^\top = U\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}U^\top = U\Sigma^{\frac{1}{2}}(U\Sigma^{\frac{1}{2}})^\top = U_\Sigma U_\Sigma^\top$ by eigen-decomposition followed by a few matrix operations. Afterwards, we have $\mathrm{tr}(WXLX^\top W^\top) = \|WXU_\Sigma\|_{\mathrm{F}}^2$. To promote the structure consistency of dual-graph and suppress the impacts of outlier samples, we adopt the rank minimization instead of the Frobenius norm. This gives the proposed graph weighted low-rank regularization term and we have a robust graph regularized dictionary learning model:

$$\min_{W,D} \|WX - DA\|_{\mathrm{F}}^2 + \alpha \mathrm{rank}(WXU_\Sigma) \ \text{ s.t. } \ \|d_j\|_2^2 \le 1, \forall j, \qquad (3)$$

in which $\mathrm{rank}(\cdot)$ means the rank operator of a matrix. Briefly, the rank constraint on WXU_Σ enforces the learned representation for seen classes to highlight common semantics. For example, attribute "it has a tail" would be annotated to various categories, e.g., loin, dog and cat. Rank constraint on WXU_Σ will assist collecting such visual features underlying the embedding space. Mathematically, the low-rankness can be propagated to DA in Eq. (1), further yielding a low-rank semantic dictionary D, which includes common semantics among seen categories.

Since rank minimization is an NP-hard challenge, majority approaches focus on searching a surrogate instead. One widely-used term is to trace norm $\|WXU_\Sigma\|_*$ [41]. Specifically, trace norm has been corroborated to be able to encourage low-rank matrix structure in the matrix completion literature, which summarize all singular values of WXU_Σ. Unfortunately, it does not allow an explicit control on the rank of WXU_Σ, which indicates that the non-zero singular values of WXU_Σ may change along with $\|WXU_\Sigma\|_*$, while the rank of WXU_Σ can keep the same. In this sense, trace norm is not a suitable surrogate to rank minimization.

Instead, we propose to enforce the rank of the updated WXU_Σ to be smaller than a target rank r [17], which skillfully transforms the problem to minimizing the square

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2867870, IEEE Transactions on Pattern Analysis and Machine Intelligence

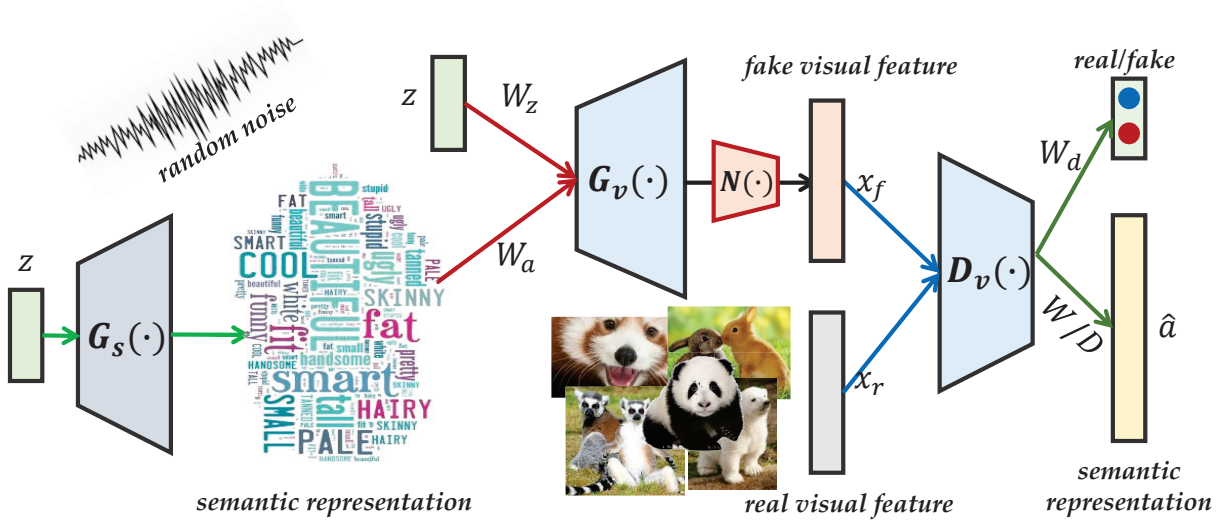


Fig. 3. Illustration of two-stage generative adversarial networks (GANs) boosted semantic dictionary for zero-shot learning. The first-stage generator $\mathcal{G}_s(\cdot)$ aims to synthesize more semantic features to cover unseen classes based on random noise z. Then, the second-stage generator $\mathcal{G}_v(\cdot)$ attempts to generate more visual features based on the real and synthesized semantic representations (a and $\mathcal{G}_s(z)$). Therefore, we could enhance the semantic dictionary and low-rank embedding learning.

sum of the $(d\text{-}r)$-smallest singular value for WXU_Σ, i.e., $\sum_{i=r+1}^{d}\sigma_i^2(WXU_\Sigma)$. While the non-zero singular values increase largely, however, they are excluded with our designed term since the norm value keeps constant. According to Ky Fan's Theorem [42], the new formula with fixed rank constraint can be reformulated as:

$$\begin{aligned}\min_{W,D,\Theta}\ &\|WX-DA\|_{\mathrm{F}}^2+\alpha\mathrm{tr}(\Theta^\top WXLX^\top W^\top\Theta)\\ &\text{s.t.}\ \ \|d_j\|_2^2\le 1,\forall j,\ \ \Theta^\top\Theta=\mathrm{I}_{d-r},\end{aligned}\tag{4}$$

where $\Theta\in\mathbb{R}^{d\times(d-r)}$ and $\mathrm{I}_{d-r}\in\mathbb{R}^{(d-r)\times(d-r)}$ is an identity matrix.

Discussion: *Non-linearity*. Robust graph regularization introduces non-linearity into the objective function, i.e. L is non-linear w.r.t. WX. *Low-Rankness*. It is well-known that rank minimization has promoted to uncover the intrinsic global structure of the data. Additionally, such locality-constrained reconstruction would generate latent semantic representation in the semantic manifold space. *Denoising*. In the presence of noise and outliers, the magnitude of $\|WXU_\Sigma\|_{\mathrm{F}}$ of the regularization term becomes very large, and so as to the whole objective function. In contrast, the fixed rank constraint on $\mathrm{rank}(WXU_\Sigma)$ will be able to suppress the impacts of outliers and noises, and thus optimize the objective function.

Remark: Compared with our previous direct rank constraint on W [17], in this paper, we adopt a weighted rank minimization regularizer by incorporating the locality and graph similarity. Thus, we can leverage additional knowledge to seek a better low-rank embedding. Compared with a plain graph regularizer on WX, the precise rank control by Θ will generate better low-rank manifold embedding.

3.3 Two-Stage Generative Adversarial Networks Boosted Semantic Dictionary Learning

If we could synthesize more data to cover the space of unseen classes as much as possible, we are more likely to build an effective visual-semantic function. Thus, we propose a two-stage generative learning scheme. The semantic feature space across seen and unseen classes is very close, e.g., the attributes of unseen classes could be represented as a combination of attributes from seen classes. While the visual feature distributions across seen and unseen have larger discrepancy. Therefore, our two-stage generative model could build a smooth path to span the semantic feature and visual feature space, attempting to cover that of the unseen classes.

In the first stage, we aim to expand the semantic space by generating features and thus are more likely to cover the semantic space of unseen classes. We follow the conventional generative model $\mathcal{G}_s(\cdot)$ and use the prior input noise $p_z(z)$ to synthesize more semantic representation. In the discriminator $\mathcal{D}_s(\cdot)$, a and synthesized semantic features $\mathcal{G}_s(z)$ are presented as inputs to a discriminative function. They formulate an objective function of a two-player minimax game as follows:

$$\mathcal{L}_1^f=\mathbb{E}[\log(1-\mathcal{D}_s(\mathcal{G}_s(z)))],\tag{5}$$

$$\mathcal{L}_1^r=\mathbb{E}[\log(\mathcal{D}_s(a))],\tag{6}$$

where the generator manages to synthesize features similar to real features and maximize $\mathcal{L}_1^f$; the discriminator attempts to differentiate the real and synthesized features by maximizing $\mathcal{L}_1^r+\mathcal{L}_1^f$. Specifically, we explore two fully-connected layers for $\mathcal{G}_s(\cdot)$ and one fully-connected layer $\mathcal{D}_s(\cdot)$.

For the second stage, we explore to generate more visual features based on synthesized and real semantic features to cover more unseen class knowledge. Given random noise $z\in\mathbb{R}^{d_z}$, real visual feature $x\in\mathbb{R}^{d_x}$ with its semantic representation $a\in\mathbb{R}^{d_a}$, we follow the conditional generative model $\mathcal{G}_v(\cdot)$ by using the prior input noise $p_z(z)$ and semantic representation to synthesize more visual representation. In the discriminator $\mathcal{D}_v(\cdot)$, x and synthesized semantic features $\mathcal{G}_v(\cdot)$ are presented as inputs to a discriminative

function. The formulation of a two-player minimax game is as follows:

$$\begin{aligned}\mathcal{L}_2^f = \quad & \mathbb{E}[\log(1-\mathcal{D}_v(\mathcal{G}_v(z|\mathcal{G}_s(z))))] \\ & +\mathbb{E}[\log(1-\mathcal{D}_v(\mathcal{G}_v(z|a)))]\end{aligned} \tag{7}$$

$$\mathcal{L}_2^r = \mathbb{E}[\log(\mathcal{D}_v(x))], \tag{8}$$

in which the generator attempts to enforce the generated features similar to real visual features and to maximize $\mathcal{L}_2^f$; the discriminator manages to classify the real and fake features by maximizing $\mathcal{L}_2^r+\mathcal{L}_2^f$. Specifically, we explore one fully-connected layer for both $\mathcal{G}_v(\cdot)$ and $\mathcal{D}_v(\cdot)$. In detail, the generator attempts to synthesize fake feature x_f with the input of random noise $z \in \mathbb{R}^{d_z}$ and semantic representation $a \in \mathbb{R}^{d_a}$. Therefore, we design $\mathcal{G}_v(\cdot)$ by taking z and a as input for example:

$$\begin{aligned}\mathcal{G}_v(\bar{z}|a) \quad &= f_1(W_g \begin{bmatrix}\bar{z}\\ a\end{bmatrix}) = f_1([W_z, W_a]\begin{bmatrix}z\\ a\end{bmatrix}) \\ &= f_1(W_z z + W_a a),\end{aligned} \tag{9}$$

where $W_g = [W_z, W_a]$ while $W_z \in \mathbb{R}^{d\times d_z}$ and $W_a \in \mathbb{R}^{d\times d_a}$. And $f_1(\cdot)$ denotes the element-wise activation function, e.g., leaky-ReLU function or Sigmoid function. To constrain the scale of the synthesized semantic and visual features, we follow a normalization process to constrain the fake feature to be the same scale of the real feature, i.e., $\mathcal{N}(x_f) = x_f/\|x_f\|_2 * \mathcal{M}$, where $\mathcal{M}$ is the mean norm of real visual feature.

Recall our goal is to synthesize more data to facilitate our low-rank embedded semantic dictionary learning, which plays a key role in revealing the visual-semantic relationship. In our two-stage generative model, we first synthesize visual features based on real semantic features A and fake semantic features $\mathcal{G}_s(z)$. Thus, we are allowed to use the generated data to improve the generalization ability of our semantic dictionary by introducing the new visual features as $\bar{X} = [X, \lambda_1\mathcal{G}_v(z|A), \lambda_2\mathcal{G}_v(z|\mathcal{G}_s(z))]$ and the new semantic features $\bar{A} = [A, \lambda_1 A, \lambda_2\mathcal{G}_s(z)]$. Note the two parameters λ_1 and λ_2 are leveraged to balance the influence of semantic dictionary learning across the real data and synthesized data with the expectation that the real data could still dominate the semantic-visual relationship learning, while the synthesized data complement the learning and expand the feature space.

To sum up, we design a new loss for low-rank embedded dictionary learning (with constraints $\|d_j\|_2^2 \leq 1, \forall j$ and $\Theta^\top\Theta = \mathrm{I}_{n-r}$) by involving the synthesized semantic and visual features as follows:

$$\mathcal{L}_c = \|W\bar{X} - D\bar{A}\|_{\mathrm{F}}^2 + \alpha\mathrm{tr}(\Theta^\top W X L X^\top W^\top \Theta). \tag{10}$$

Finally, we could iteratively optimize discriminator and generator in a minimax strategy as the following objective:

$$\begin{aligned}&\min_{\substack{W,D,\Theta,\\ \mathcal{D}_s(\cdot),\mathcal{D}_v(\cdot)}} \ \max_{\substack{\mathcal{G}_s(\cdot),\\ \mathcal{G}_v(\cdot)}} \ \mathcal{L}_c + \beta\sum_{i=1}^{2}(\mathcal{L}_i^f + \mathcal{L}_i^r), \\ &\text{s.t.} \ \ \|d_j\|_2^2 \leq 1, \forall j, \ \ \Theta^\top\Theta = \mathrm{I}_{d-r}.\end{aligned} \tag{11}$$

where β is a trade-off parameter to balance square-loss and cross-entropy loss.

Discussion: With objective function (11), we manage to synthesize more semantic and visual features from seen classes to augment the feature space to cover the unseen classes as much as possible. In this way, we could build a more effective visual-semantic function by better solving the domain shift issue across seen classes and unseen classes. And hence, we are able to handle the zero-shot learning more efficiently.

Moreover, Zero-shot learning can a special case of transfer learning (TL) [14], [26]. We could relate ZSL and TL in different lines. First of all, seen classes in ZSL can be treated as source domain in TL; while unseen classes in ZSL could be considered as target domain in TL. However, the general setting of ZSL is different from TL, since we cannot have access to any unseen data in the training stage for ZSL, while target domain is available in the training stage of TL. Secondly, there are two views (visual data and semantic data) in seen classes for ZSL. We could treat two views as two domains, while the key is to adapt the relationship between two views to unseen classes. However, for TL, the key is to adapt the label knowledge from one domain (source) to another (target). Generally, the ideas of GAN in TL are in two lines: one is to seek domain-invariant features [43], [44] and the other is to transform one domain to another in image space [45]. The key of TL is to solve the domain shift across source and target domains. However, in ZSL, the key is to learn knowledge from seen classes and generalize to unseen classes. There exists domain shift between seen classes and unseen classes. Thus, we propose a two-stage generative model, where we aim to synthesize more semantic features in the first stage, then we could augment the visual feature space to have better chance to cover the unseen classes. The motivation is also different. For TL, GAN loss is used as a domain alignment strategy. Our goal is to do data augmentation to boost the low-rank embedded dictionary learning, which is used in the test stage for unseen classes.

Zero-shot learning is also much related to "emerging new classes learning". The major difference is that emerging new classes learning does not assume knowledge about the object, while lots of zero-shot learning research exploits the knowledge (such as object parts) about the object. Along this line, Mu et al. proposed to dynamically maintain two low-dimensional matrices to detect emerging new classes and update the model in the data stream [46]. Following this, Mu et al. further explored completely-random trees to classify streaming emerging new classes [47]. Similarly, Zhu et al. proposed a model to recognize samples on currently observed labels, detect the emergence of a new label in new samples, and build a new classifier for every new label which works collaboratively with the classifier for observed labels [48].

3.4 Optimization

As the optimization in Eq. (11) is not joint convex over all variables, there is no closed-form solution. Thus, we resort to an iterative strategy to optimize a single variable each time. We further group into two sub-problems, i.e., learning semantic dictionary D and low-rank embedding W, Θ by fixing generators and discriminators; and learning

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2867870, IEEE Transactions on Pattern Analysis and Machine Intelligence

$\mathcal{G}_{s/v}(\cdot)$ and $\mathcal{D}_{s/v}(\cdot)$ with semantic dictionary and low-rank embedding fixed. Those two sub-problems will be updated iteratively.

3.4.1 Learning Semantic Dictionary and Low-rank Embedding

Semantic Dictionary Refinement: When W, $\mathcal{G}_{s/v}(\cdot)$ and $\mathcal{D}_{s/v}(\cdot)$ are fixed, we could update the semantic dictionaries D as:

$$D = \arg\min_{D} \|W\bar{X} - D\bar{A}\|_{\mathrm{F}}^2 \ \text{ s.t. } \ \|d^j\|_2^2 \leq 1, \forall j. \tag{12}$$

By applying projected gradient descent, we optimize the j-th dictionary atom d^j as follows:

$$\begin{cases} s^j = d^j - \frac{1}{\mu}\nabla_{d^j}\mathcal{F}(W, D), \\ d^j = \arg\min_{\|d^j\|_2^2} \|d^j - s^j\|_2 = \frac{s^j}{\|s^j\|_2}, \end{cases} \tag{13}$$

in which μ is the step size, $\mathcal{F}(W, D) = \|W\bar{X} - D\bar{A}\|_{\mathrm{F}}^2$.

Learning Low-Rank Embedding: When D is fixed, we could update W, Θ.

Update W:

$$\begin{aligned} W = \ & \arg\min_{W} \|W\bar{X} - D\bar{A}\|_{\mathrm{F}}^2 \\ & + \alpha \mathrm{tr}(\Theta^\top W X L X^\top W^\top \Theta). \end{aligned} \tag{14}$$

We then calculate the deviation to W and set it to zero:

$$\begin{aligned} & (W\bar{X} - D\bar{A})\bar{X}^\top + \alpha\Theta\Theta^\top W = 0, \\ \Rightarrow \ & W\bar{X}\bar{X}^\top + \alpha\Theta\Theta^\top W = D\bar{A}\bar{X}^\top, \end{aligned} \tag{15}$$

which is a standard Sylvester formulation and could be effectively solved with the BartelsStewart algorithm [49].

Update Θ:

When W is updated, we could optimize Θ with the following eigen-decomposition problem as:

$$\Theta = \arg\min_{\Theta^\top\Theta = \mathrm{I}_{d-r}} \mathrm{tr}(\Theta^\top W X L X^\top W^\top \Theta), \tag{16}$$

where the optimal solution of Θ is formed by the $(d-r)$ eigenvectors of $WXLX^\top W^\top$ corresponding to the $(d-r)$ smallest eigenvalues.

For clarity, we present the details of the optimization in **Algorithm 1**. For μ, we set them to be 10^{-3} as default. α and r are two hyper parameters, which need to be tuned during the experiments. For initializations of W, D, Θ, we first use the real visual and semantic features for seen classes to optimize Eq. (4). Then, we use the values for W, D, Θ to further optimize Eq. (10) with the synthesized data.

3.4.2 Learning Two-Stage Generative Adversarial Networks

Given W, D, we could update generators $\mathcal{G}_{s/v}(\cdot)$ and discriminators $\mathcal{D}_{s/v}(\cdot)$ iteratively. Because GANs could be optimized as a *minimax* optimization problem, we first fix the generator to optimize the discriminator, then fix the discriminator to optimize the generator by following the literature [29], [33], [50].

Algorithm 1 Solving First Sub-Problem

Input: $X, A, \alpha, \beta, \lambda_{1/2}, \mathcal{G}_{s/v}(\cdot)$

Initialize: $W, D, \Theta, \mu = 10^{-3}, \epsilon = 10^{-5}, t = 0$.

while not converged **do**

1. Optimize D via Eq. (12) by fixing others.
2. Optimize W via Eq. (15) by fixing others.
3. Optimize Θ via Eq. (16) by fixing others.
4. Check the convergence conditions: $|\mathcal{L}_c^{t+1} - \mathcal{L}_c^t| < \epsilon$.
5. $t = t + 1$.

end while

output: W, D, Θ.

Since we have two-stage GANs, we first train them separately for a better initialization. In detail, we first train the first-stage GAN by only using $\mathcal{G}_s(\cdot)$ and $\mathcal{D}_s(\cdot)$, where the parameters are initialized randomly. Then we train the second-stage GAN with the fixed first-stage generator $\mathcal{G}_s(\cdot)$ and the semantic dictionary loss $\mathcal{L}_c$. Specifically, we hope $W_a a$ to preserve the class center, while $W_z\bar{z}$ to mimic the class variation. Thus, we initialize W_a via $W_a = \arg\min_{W_a} \|X - W_a A\|_{\mathrm{F}}^2$, and then we could obtain its class center by multiplying W_a with its semantic representation a. For random noise part, we attempt to simulate the class variations to span the space. Hence, we randomly initialize W_z and z. To this end, we could model the data variation from seen classes and synthesize more meaningful data for unseen classes.

After initialization separately, we joint train them as whole by considering W/D are fixed. To make the training more stable, we adopt the Wasserstein GAN strategy [51], and implement our model through TensorFlow[3]. We set the learning rate to be 10^{-4} and the optimizer with Adam optimizer. We exploit leaky-relu and sigmoid activation functions for $\mathcal{G}_{s/v}(\cdot)$ and $\mathcal{D}_{s/v}(\cdot)$, respectively.

3.5 Zero-Shot Learning

Given a test visual data x_u^i, we explore a reconstruction scheme over semantic dictionary D and low-rank embedding W to predict its semantic representation $\tilde{a}_u$ as:

$$\tilde{a}_u = \arg\min_{\tilde{a}_u} \|Wx_u^i - D\tilde{a}_u\|_2^2 + \xi\|\tilde{a}_u\|_2^2, \tag{17}$$

where ξ is simply set as 10^{-2}.

Generally, zero-shot learning includes many sub-tasks, e.g., zero-shot recognition, zero-shot annotation or zero-shot image retrieval, and here we explain the basic settings in test phase one by one. For zero-shot recognition, the A_u^c is known in the test phase as reference samples. Each reference sample has been represented by the attributes through our model and the average of each class will be used as the representative of class c_u. Thus, we would have a label-attribute table for unseen classes (i.e., A_u with C_u classes). When we get the predicted semantic feature for unseen visual features, we will search from label-attribute and the most similar one will tell us the predicted label. Based on this, we can calculate the recognition accuracy. Similarly, for zero-shot annotation, we should verify if the predicted annotation is correct or not by comparing with A_u^c. For zero-shot image retrieval, we need to search the top similar visual images for unseen classes by given any semantic representation.

3. https://www.tensorflow.org/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2867870, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1
Statistics of four benchmark datasets.

Dataset	aP&aY	AwA	CUB	SUN
#Training classes	20	40	150	707
#Test classes	12	10	50	10
#Instances	15,339	30,475	11,788	14,340
#Semantic Space	64	85	312	102

4 EXPERIMENT

In this part, we evaluate on four standard zero-shot learning benchmarks in terms of different ZSL tasks and properties analysis.

4.1 Dataset & Experimental Setting

Four standard benchmarks are evaluated with their statistics provided in Table 1.

aPascal-aYahoo (aP&aY) contains 20 object classes from the PASCAL VOC 2008 dataset and 12 object categories collected with the Yahoo image search engine. Following previous work [3], [5], [52], [53], we train on image samples from PASCAL VOC 2008 and evaluate on images from Yahoo image engine. In addition, 64 attributes that characterize shape, material and the presence of important parts are provided.

Animal with Attribute (AwA) has 50 animals in total, each with over 92 samples, paired with a human provided 85-attribute inventory and corresponding class-attribute associations.

Caltech-UCSD Birds 2011 (CUB) is a fine-grained bird benchmark with attributes, which includes 200 different birds, with 11,788 images in total. The class level attribute annotations are given with 312 visual attributes (e.g., color, part pattern).

SUN scene attribute dataset is a fine-grained dataset, in which the difference between classes is quite marginal. There exist 717 scenes, each with 20 samples. There are 102 attributes to annotate each image.

More specifically, each image of the aP&aY, CUB and SUN datasets has its own attribute description, meaning that two images of the same class can have different attributes. Differently, all the images of a given class in AwA share the same attributes. We exploit the continuous attributes as the semantic representation, as it performs better than the binary one [1]. Moreover, we follow previous zero-shot learning works [3], [10], [12], [25], [40], [53] and adopt the same training and test protocol shown in Table 1.

Regarding the representation of visual images, we utilized the recently proposed deep features, i.e., AlexNet CNN FC_7 [54], VGG-VeryDeep-19 [55], and GoogLeNet [56]. For AlexNet CNN FC_7, we adopt the 7-th layer (FC_7) as visual features with dimensions 4,096. For VGG-VeryDeep-19, we utilize the top layer hidden unit activations of the network as a 4,096-dimensional CNN feature. For GoogLeNet, we adopt the 1,024-dimensional activations of the pooling units.

For β, λ_1 and λ_2, we set them as $0.1, 0.5, 0.1$ throughout the experiments for simplicity. For k-nn graph, we fix $k = 10$ across different datasets. The dictionary size for AwA is set as 300 and those for the other three datasets are chosen to 700. Following previous ZSL works [3], [5], [52], [53], we adopt cross-validation to tune the parameter α. To be specific, we conduct parameter selection over the training set with 3-fold cross-validation which splits the seen class data into a training set and a validation set. The trade-off parameter is selected from the range $\{10^\tau | \tau = -3, -2, \cdots, 2, 3\}$ based on the test performance on the labeled instances from the observed classes in the validation set. After parameter selection, we used the selected parameters to perform ZSL with the original seen and unseen classes. Since different initializations could lead to different optimal solutions, thus, we run five times to average our results.

TABLE 2
Zero-shot classification accuracy (%) of the comparisons on the four datasets. There are two kinds of features: AlexNet CNN features **[ALEX]**, and GoogLeNet features **[GGL]**.

Features	Methods	aP&aY	AwA	CUB	SUN
[ALEX]	DAP [57]	-	53.2	31.4	-
	UDA [2]	-	73.2	39.5	-
	COSTA [18]	-	55.2	36.9	-
	SJE [15]	-	61.9	40.3	-
	ESZSL [52]	-	53.2	37.2	-
	ISEC [5]	46.1	-	42.0	75.5
	SynC [1]	-	64.8	47.1	-
	LESD [17]	48.9	71.4	43.9	77.1
	Ours	50.2	73.2	45.3	79.6
[GGL]	DAP [57]	-	60.5	39.1	-
	COSTA [18]	-	61.8	40.8	47.9
	SJE [15]	-	66.7	50.1	87.0
	ESZSL [52]	-	59.6	44.0	82.1
	SynC [1]	-	72.9	54.5	90.0
	LESD [17]	58.8	76.6	56.2	88.3
	MFMR [11]	46.4	76.6	46.2	81.5
	UVDS [12]	38.7	80.3	57.2	60.8
	Ours	60.2	79.5	58.4	89.4

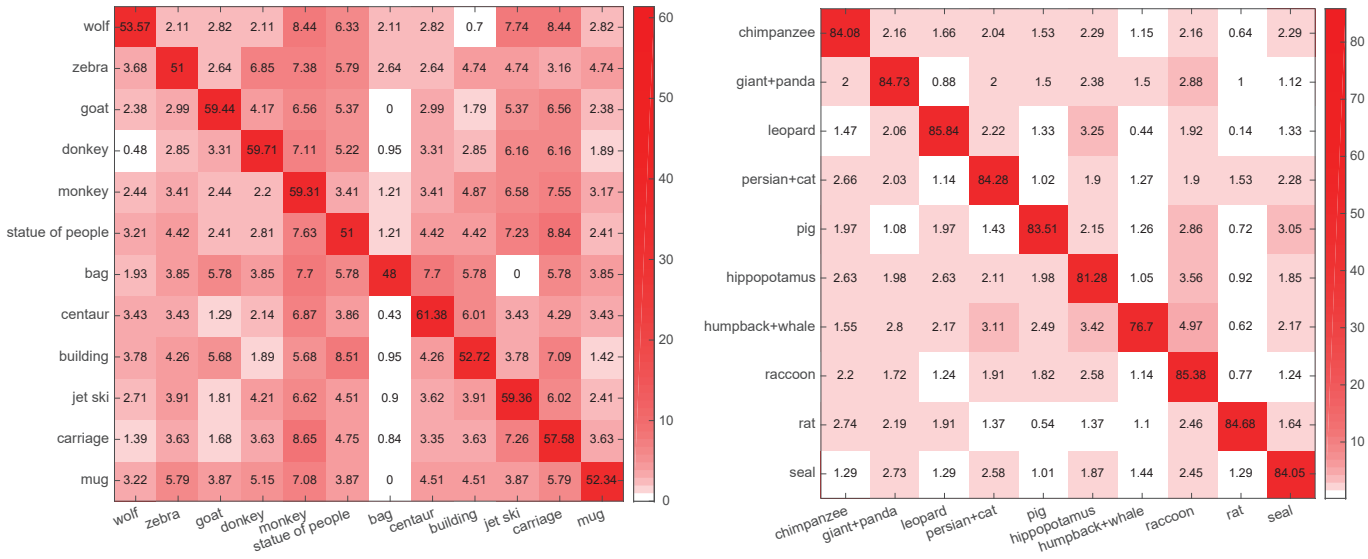
TABLE 3
Zero-shot classification accuracy (%) of the comparisons on the four datasets using VGG-VeryDeep-19 CNN features.

Methods	aP&aY	AwA	CUB	SUN
DAP [57]	38.2	57.2	39.8	72.0
ESZSL [52]	24.2±2.9	75.3±2.3	-	82.1±0.3
SSE [53]	46.2±0.5	76.3±0.8	30.4±0.2	82.5±1.3
JSLE [3]	50.4±3.0	79.1±0.5	41.8±0.5	83.8±0.3
ISEC [5]	53.2±0.9	77.3±1.0	43.3±0.4	84.4±0.7
KDICA [27]	-	73.8	43.7	-
LESD [17]	55.2±1.1	82.8±0.9	45.2±0.3	86.0±0.3
MFMR [11]	48.2	79.8	47.7	84.0
UVDS [12]	42.3±0.5	82.1±0.1	44.9±0.9	80.5±0.8
JDZSL [58]	-	88.2	47.1	85.9
LDLA [10]	51.1	80.5	56.4	83.5
DSRL [25]	56.3±0.4	77.4±0.1	50.3±0.1	82.0±0.0
DSRL-LP [25]	51.3±1.4	87.2±0.3	57.1±0.1	85.4±0.2
Ours	56.3±0.2	84.4±0.4	48.5±0.3	88.2±0.7

4.2 Zero-Shot Classification

In this part, we mainly evaluate on the zero-shot recognition task by comparing to DAP [57], ESZSL [52], SSE [53], JSLE [3], ISEC [5], KDICA [27], UDA [2], COSTA [18], SJE [15], SynC [1], MFMR [11], LESD [17], UVDS [12], JDZSL [58], LDLA [10], and DSRL (DSRL-LP) [25]. The recognition

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2867870, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) aP&aY (b) AwA

Fig. 4. Confusion matrices of the test results on unseen classes for the proposed method on (a) aP&aY and (b) AwA. Diagonal numbers indicate the correct prediction accuracy. Column means the ground truth and row denotes the predictions.

performance in term of top1 accuracy is reported in Tables 2 and 3. Note that we directly copy partial results from the published works.

As wen can see, our proposed approach works better than other competitors in most cases across four benchmarks with remarkable margins. Note that JDZSL and DSRL do not strictly follow ZSL setting, since they involve either visual features or semantic features from unseen classes during model training, thus, they could have better performance in some cases. However, our model can still beat them in some cases, although we only train on seen classes. For the comparison of three different visual features, we find that VGG-VeryDeep-19 and GoogLeNet features outperforms AlexNet CNN FC_7, which indicates that these two features are more discriminative in representing visual images. Comparing VGG-VeryDeep-19 with GoogLeNet features, we notice that VGG-VeryDeep-19 presents superiority for AwA, while GoogLeNet features work better for aP&aY, CUB and SUN. Furthermore, we also observe that all comparisons achieve better performance on AwA and SUN over on aP&aY and CUB. The reason we consider is that the class association in AwA and SUN is much significant than in aP&aY, since AwA only contains animals while SUN only has scene classes, whereas aP&aY includes random object categories. Hence, it is more efficient to learn the common information across different categories in AwA/SUN than in aP&aY. Besides, the provided attributes of AwA show special descriptions tailored for animals; however, the provided semantic attributes of aP&aY on shape, part, and material are insufficient to describe an object comprehensively. Thus, richer and more valuable knowledge could be transferred from the seen classes to the unseen ones for AwA than for aP&aY. Differently, there exist 200 bird species in CUB and some birds look very similar from the visual appearance, which leads CUB to a very challenging ZSL benchmark.

Furthermore, we provide the zero-shot classification accuracy in term of the confusion matrices for our model (Figure 4), in which we experiment on aP&aY and AwA with VGG-VeryDeep-19 features. In each confusion matrix, the column indicates the ground truth and the row means the predictions. From the confusion matrix of AwA, we observe that our algorithm obtains over 85% accuracy for some animal categories, e.g., leopard (85.84%) and raccoon (85.38%). Given the fact that our approach is trained with no labeled data from these classes at all, these results are quite impressive. The confusion matrix for aP&aY also provides promising results on some categories, e.g.,, donkey (59.71%) and centaur (61.38%). All these results verify the effectiveness of our model in solving zero-shot classification.

4.3 Zero-shot Retrieval

In the retrieval task, a semantic representation of an unseen class is provided to retrieve top matched image instances. mean average precision (mAP) is usually adopted to measure the performance. As the retrieval evaluation is not widely explored in previous ZSL studies, we present a comprehensive comparison to several state-of-the-art approaches, i.e., SSE [53], JSLE [3], SynC [1], ISEC [5], MFMR [11] and our conference version, i.e., LESD [17].

Table 4 illustrates comparative results in terms of mAP for four benchmarks using VGG-VeryDeep-19 features. We could observe that our proposed model achieves average mAP score of 58.0%, compared to the best counterpart of MFMR (its performance is 56.2%). Especially for three datasets out of AwA, our model achieves consistent superiority over the state-of-the-arts. It thus again validates the effectiveness of our GSDL on conducting more effective knowledge transfer for unseen classes. For AwA, we consider it only has class-level attributes, which hinders to synthesize more diverse semantic and visual features.

Furthermore, we provide qualitative results for the proposed model. Given only semantic representations (no image samples), we attempt to present what visual information our model is able to capture for unseen classes. Figure 5 presents 10 unseen class labels of CUB and Figure 6 lists 12 unseen class labels of aP&aY, where we show

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2867870, IEEE Transactions on Pattern Analysis and Machine Intelligence

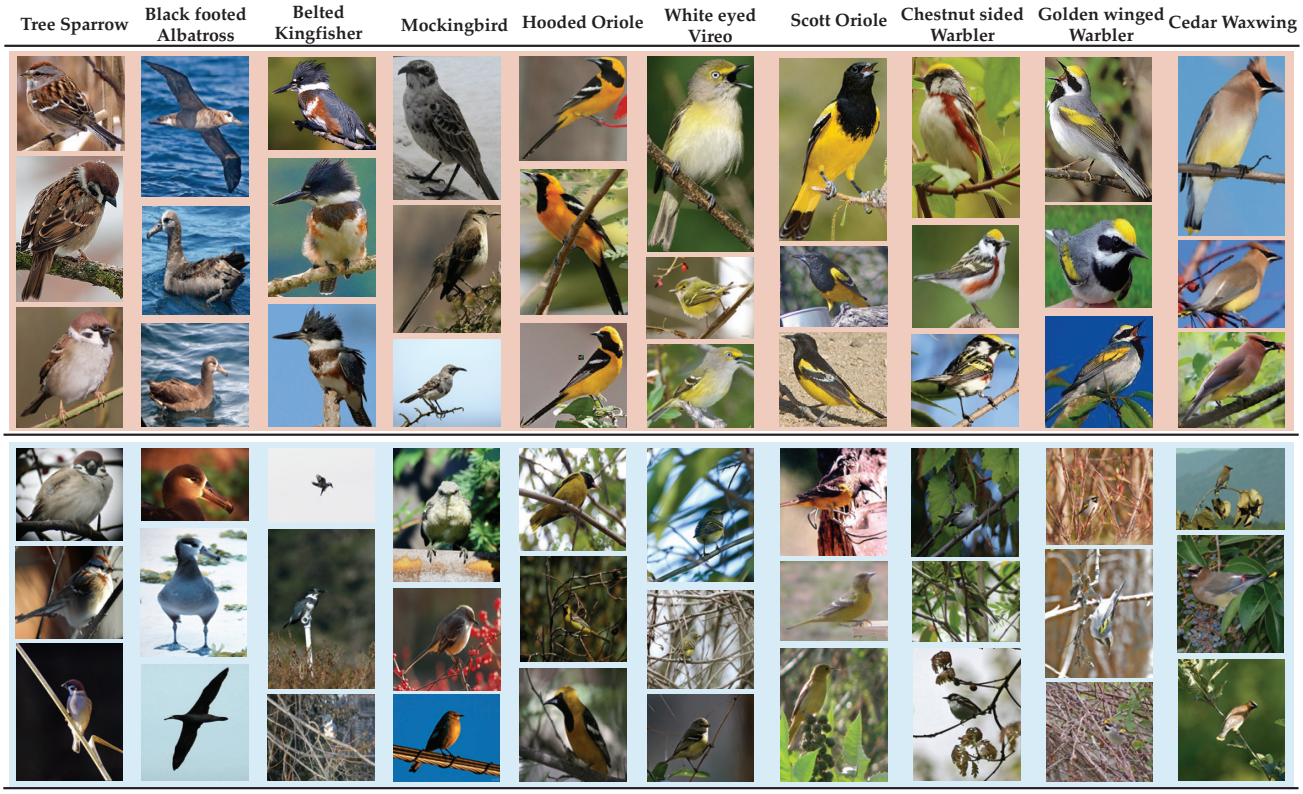


Fig. 5. Qualitative results of our approach on CUB, where 10 unseen class labels are shown on the top. Then we list the top-3 instances recognized in each class in the middle. The third row represents the top-3 misclassified samples.

TABLE 4
Retrieval performance comparison (%) in terms of mAP. The best results on each dataset are highlighted in bold font.

Method	aP&aY	AwA	CUB	SUN	Avg.
SSE [53]	15.4	46.25	4.7	58.9	31.3
JSLE [3]	32.7	66.5	23.9	76.5	49.9
SynC [1]	30.4	65.4	34.3	74.3	51.1
ISEC [5]	36.9	68.1	25.3	52.7	45.8
MFMR [11]	**45.6**	70.8	30.6	77.4	56.2
LESD [17]	40.3	71.2	31.3	76.6	54.9
Ours	44.8	**73.6**	**34.2**	**79.2**	**58.0**

TABLE 5
Generalized ZSL recognition (%) in terms of accuracy, where $\mathcal{U}$: Unseen classes; $\mathcal{S}$: Seen classes; $\mathcal{T} = \mathcal{S} + \mathcal{U}$.

Datasets	Methods	$\mathcal{U}$-$\mathcal{U}$	$\mathcal{S}$-$\mathcal{S}$	$\mathcal{U}$-$\mathcal{T}$	$\mathcal{S}$-$\mathcal{T}$
AWA	DAP [57]	51.1	78.5	2.4	77.9
	SynC [1]	73.4	81.0	0.4	**81.0**
	MFMR [11]	79.9	76.1	13.4	75.6
	UVDS [12]	80.3	86.7	15.3	79.5
	LESD [17]	80.9	87.2	16.4	78.2
	Ours	**82.1**	**88.6**	**17.8**	80.4
CUB	DAP [57]	38.8	56.0	4.0	55.1
	SynC [1]	54.4	73.0	13.2	72.0
	MFMR [11]	56.9	74.1	23.4	73.2
	UVDS [12]	57.5	75.4	23.8	76.5
	LESD [17]	56.7	76.4	23.6	75.6
	Ours	**58.6**	**77.8**	**25.4**	**78.2**

the top-3 correct images classified (second row in red) and top-3 wrong images misclassified (third row in blue) into each category with VGG-VeryDeep-19 features. Seeing from the top images, our approach reasonably captures discriminative visual information for every unseen category solely based on its semantic representation. We could also witness that the misclassified samples have very similar visual appearance to that of predicted class. Thus, even humans cannot distinguish them easily.

4.4 Generalized Zero-shot Recognition

Most recently, researchers manage to explore the generalized setting of ZSL (GZSL), where the evaluation set from both seen and unseen categories, as it is much more practical in the real-world applications. Thus, we evaluate our model on GZSL strictly following the experimental setting in [59]. Specifically, we reserve 20% data samples of the seen categories and merge them with the unseen samples as the evaluation set. For consistency, we adopt the GoogLeNet feature and the continuous attributes. Table 5 shows the four cases for GZSL, in which $\mathcal{U}$-$\mathcal{U}$ means the traditional unseen-to-unseen ZSL, $\mathcal{S}$-$\mathcal{S}$ denotes the conventional supervised learning task, $\mathcal{U}$-$\mathcal{T}$ and $\mathcal{S}$-$\mathcal{T}$ represent two kinds of GZSL which evaluate whether the learned unseen/seen models can be confused by one another. From the results, we notice that the proposed approach could consistently achieve better performance than other comparisons in most cases. In AwA, our approach performs worse in one out of the four scenarios, that is, our performance is slightly lower than that of SynC [1] for $\mathcal{S}$-$\mathcal{T}$. The seen/unseen balance could be treated as an over-fitting issue: when

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2867870, IEEE Transactions on Pattern Analysis and Machine Intelligence

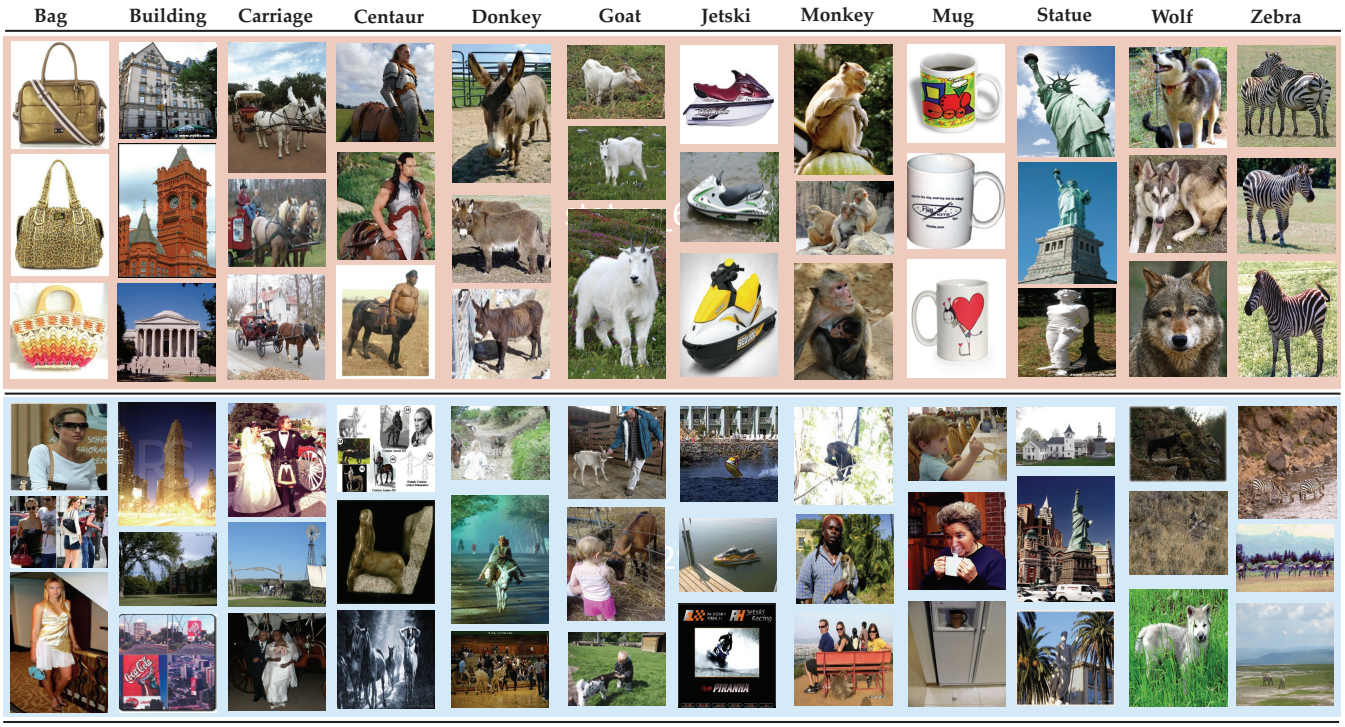


Fig. 6. Qualitative results of our approach on aP&aY, where 12 unseen class labels are shown on the top. Then we list the top-3 instances recognized in each class in the middle. The third row represents the top-3 misclassified samples.

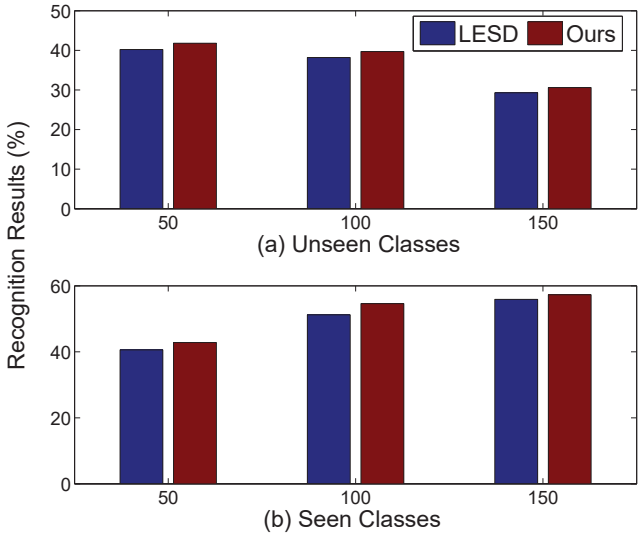


Fig. 7. Performance of our conference model and the proposed method under different number of seen/unseen classes on CUB.

we sacrifice the performance on seen categories ($\mathcal{S}$-$\mathcal{T}$), we tend to significantly improve the performance on unseen classes $\mathcal{U}$-$\mathcal{T}$. The evaluation on CUB dataset also shows such evidence.

4.5 Evaluation on the Seen/Unseen Class Scale

In this section, we testify the performance with various scales for the seen/unseen categories using the CUB dataset (GoogLeNet features).

First, we evaluate the performance of zero-shot recognition on different sizes of unseen classes by fixing the scale of seen classes to 50. We conduct 10 random selections for the seen/unseen classes and report the average results in Figure 7 (a). From the results, we could see that the classification accuracy drops when the scale of unseen classes increases. That means a fixed seen knowledge set has a limited adaptation ability.

Second, we experiment on the different class scale for seen classes by fixing the size of unseen classes as 50. We also do 10 random selections of the seen/unseen classes and report the average results in Figure 7 (b). Interestingly, we find that we can obtain higher recognition rate by involving more seen classes, which indicates that we can enhance the generalization ability by seeing more objects.

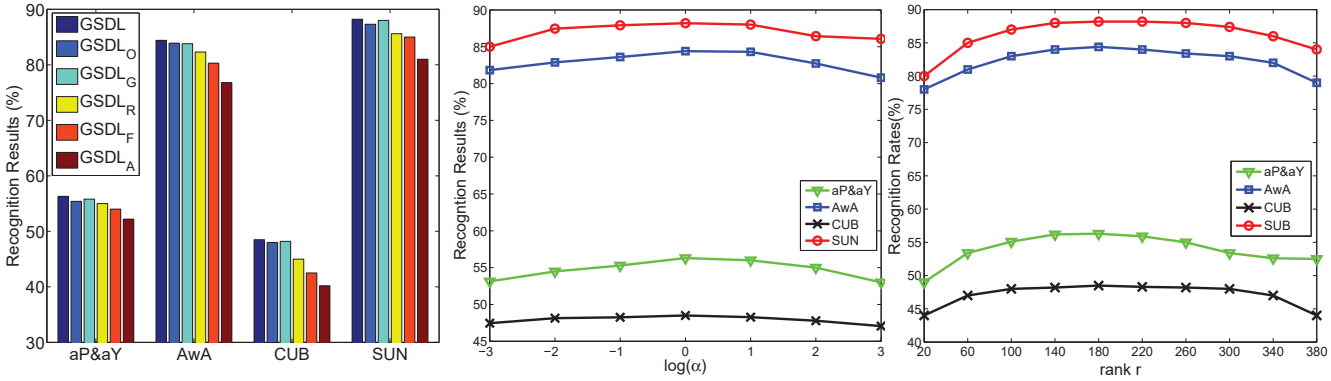
4.6 Empirical Analysis

In this section, we testify more properties of our proposed framework on four benchmarks.

First of all, to dive deeper into the efficacy of our two-stage generative zero-shot model, we evaluate several variants: GSDL$_A$ means that we remove the two-stage GANs and only learn W, D on real data as a baseline; GSDL$_O$ means that we preserve the second-stage GAN boosted low-rank embedded semantic dictionary model; GSDL$_F$ means that we only explore Frobenius norm on W by removing rank constraint on WXU_Σ; GSDL$_R$ means that we only explore low-rank constraint on W instead of WXU_Σ; GSDL$_G$ means that we only explore graph regularizer on WX instead of rank constraint on WXU_Σ.

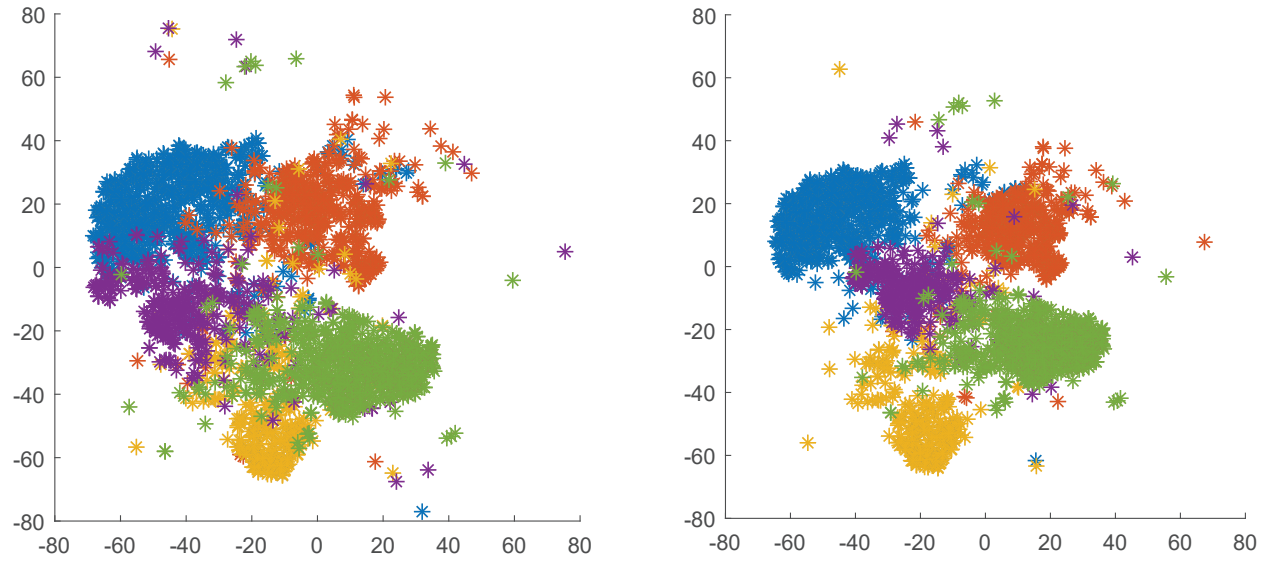
From the results in Figure 8(a), we notice that two-stage generative model contributes a lot, which could synthesize more data to mitigate the domain shift across seen and unseen classes. This indicates the effectiveness of our two-stage generative model. Compared with GSDL$_A$ and GSDL$_O$, we observe that two-stage GANs model helps to enhance the generalization ability of our low-rank embedding W and semantic dictionary D, which further denotes that the first-stage GAN can also span the semantic feature space

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2867870, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Variants (b) parameter analysis (c) rank r

Fig. 8. (a) Comparison of different variants to our model, (b) parameter analysis on α, (c) rank analysis of r on four benchmarks with VGG-VeryDeep-19.



(a) GSDL$_A$ (b) Ours

Fig. 9. Visualization of 5 hard unseen class samples from AwA using their learned semantic representations, where (a) is from GSDL$_A$ and (b) is from GSDL (ours). Each color represents a class and each point represents an image.

while further span the visual feature space. Compared with GSDL$_F$, we notice that rank constraint also helps to improve the performance, since rank constraint could help seek a more effective embedding space to gather similar visual features. By involving the manifold information to the rank constraint, we also achieve better performance. Furthermore, we observe that our proposed weighted low-rank constraint would improve limited performance based on the traditional manifold regularizer,, i.e., GSDL$_G$.

To further validate whether our two-stage generative model enhances the generalization ability of semantic dictionary, we visualize 5 hard unseen AwA classes of their predicted latent semantics with the VGG visual features as the input. Specifically, we adopt t-SNE[4] to map the learned latent semantic representations of each unseen class instance to a 2-D plane. We compare our two-stage generative model with GSDL$_A$ and show the results in Figure 9. Interestingly, we could observe that the intra-class images are grouped much closer and those from different classes are more separated for our model compared with GSDL$_A$, which indicates that the two-stage generative model could improve discriminability and generalizability of semantic dictionary for the unseen classes recognition.

Secondly, we evaluate the parameter α in our model for the weighted rank constraint. Through the analysis on parameter α (Figure 8 (b)), we notice that our model can obtain promising performance around $[1, 10]$ on four datasets. This shows the weighted rank constraint could help improve the zero-shot learning ability of our model by exploring more intrinsic structure within seen data, i.e., semantic and visual features.

Finally, we evaluate the rank of WXU_Σ to see its influence during our visual-semantic function training. From the evaluation on rank r, we witness that the classification accuracy tends to be better when r is around 140 to 220 (Figure 8 (c)). Specifically, the performance would degrade when r is either too small or too large. This indicates that a low-rank constraint would benefit grouping similar features across different categories for zero-shot learning purpose.

5 Conclusion

In this paper, we proposed a novel two-stage generative model boosted semantic dictionary learning framework under low-rank embedding for zero-shot tasks. Specifically, we

4. https://lvdmaaten.github.io/tsne/

designed an effective knowledge transfer model by jointing semantic dictionary and low-rank embedding learning into a unified framework. Thus, the semantic gap between visual features and their semantics could be mitigated with the learned visual-semantic function. To enhance the generalizability of our model, we explored two-stage generative adversarial networks to synthesize more semantic and visual features to cover the space of unseen classes in the training stage. Hence, it was more likely to build a powerful semantic dictionary and low-rank embedding. Extensive experiments on four zero-shot benchmarks indicated the superiority of our proposed model through comparing with the state-of-the-art ZSL algorithms.

ACKNOWLEDGMENT

This work is supported in part by the NSF IIS award 1651902, NIJ Graduate Research Fellowship 2016-R2-CX-0013, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Award W911NF-17-1-0367.

REFERENCES

[1] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 5327–5336.

[2] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Unsupervised domain adaptation for zero-shot learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2452–2460.

[3] Z. Zhang and V. Saligrama, "Zero-shot learning via joint latent similarity embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 6034–6042.

[4] X. Li, Y. Guo, and D. Schuurmans, "Semi-supervised zero-shot classification with label representation learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4211–4219.

[5] M. Bucher, S. Herbin, and F. Jurie, "Improving semantic embedding consistency by metric learning for zero-shot classiffication," in *European Conference on Computer Vision*. Springer, 2016, pp. 730–746.

[6] X. Xu, T. M. Hospedales, and S. Gong, "Multi-task zero-shot action recognition with prioritised data augmentation," in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 343–359.

[7] R. Qiao, L. Liu, C. Shen, and A. v. d. Hengel, "Less is more: zero-shot learning from online textual documents with noise suppression," *Proceedings of the IEEE International Conference on Computer Vision*, 2016.

[8] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, "Zero-shot object recognition by semantic manifold distance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2635–2644.

[9] G.-J. Qi, W. Liu, 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*, 2016.

[10] H. Jiang, R. Wang, S. Shan, Y. Yang, and X. Chen, "Learning discriminative latent attributes for zero-shot classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4223–4232.

[11] X. Xu, F. Shen, Y. Yang, D. Zhang, H. T. Shen, and J. Song, "Matrix tri-factorization with manifold regularizations for zero-shot learning," in *Proceeding of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7140–7148.

[12] Y. Long, L. Liu, F. Shen, L. Shao, and X. Li, "Zero-shot learning using synthesised unseen visual data with diffusion regularisation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[13] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *Journal of Machine Learning Research*, vol. 11, no. Sep, pp. 2487–2531, 2010.

[14] Z. Ding, M. Shao, and Y. Fu, "Incomplete multisource transfer learning," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 2, pp. 310–323, 2018.

[15] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2927–2936.

[16] 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*. IEEE, 2009, pp. 951–958.

[17] Z. Ding, M. Shao, and Y. Fu, "Low-rank embedded ensemble semantic dictionary for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2050–2058.

[18] T. Mensink, E. Gavves, and C. G. Snoek, "Costa: Co-occurrence statistics for zero-shot classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2441–2448.

[19] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 11, pp. 2332–2345, 2015.

[20] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 9, pp. 2117–2130, 2013.

[21] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Advances in neural information processing systems*, 2009, pp. 1410–1418.

[22] D. Parikh and K. Grauman, "Relative attributes," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 503–510.

[23] X. Yu and Y. Aloimonos, "Attribute-based transfer learning for object categorization with zero/one training example," in *European conference on computer vision*. Springer, 2010, pp. 127–140.

[24] P. Peng, Y. Tian, T. Xiang, Y. Wang, and T. Huang, "Joint learning of semantic and latent attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 336–353.

[25] M. Ye and Y. Guo, "Zero-shot classification with discriminative semantic representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 5103–5111.

[26] Z. Ding, M. Shao, and Y. Fu, "Robust multi-view representation: A unified perspective from multi-view learning to domain adaption." in *International Joint Conference on Artificial Intelligence*, 2018, pp. 5434–5440.

[27] C. Gan, T. Yang, and B. Gong, "Learning attributes equals multi-source domain generalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 87–97.

[28] Z. Ding, S. Ming, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, pp. 1192–1198.

[29] 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*, 2014, pp. 2672–2680.

[30] Z. Ding, Y. Guo, L. Zhang, and Y. Fu, "One-shot face recognition via generative learning," in *13th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2018, pp. 1–7.

[31] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[32] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with pixelcnn decoders," *arXiv preprint arXiv:1606.05328*, 2016.

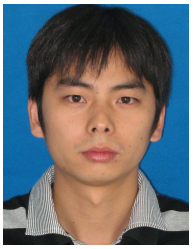
[33] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[34] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

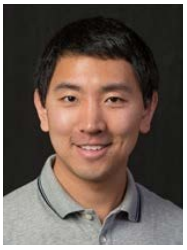
[35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[36] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2016, pp. 2172–2180.

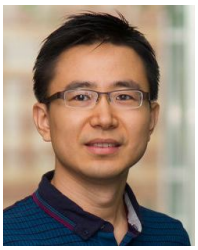
[37] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
[38] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 3387–3395.
[39] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1212–1221.
[40] Y. Li, D. Wang, H. Hu, Y. Lin, and Y. Zhuang, "Zero-shot recognition using dual visual-semantic mapping paths," pp. 3279–3287, 2017.
[41] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.
[42] K. Fan, "On a theorem of weyl concerning eigenvalues of linear transformations i," *Proceedings of the National Academy of Sciences*, vol. 35, no. 11, pp. 652–655, 1949.
[43] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 4068–4076.
[44] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, 2015, pp. 1180–1189.
[45] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *The IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 2, 2017, p. 7.
[46] X. Mu, F. Zhu, J. Du, E.-P. Lim, and Z.-H. Zhou, "Streaming classification with emerging new class by class matrix sketching." 2017, pp. 2373–2379.
[47] X. Mu, K. M. Ting, and Z.-H. Zhou, "Classification under streaming emerging new classes: A solution using completely-random trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1605–1618, 2017.
[48] Y. Zhu, K. M. Ting, and Z.-H. Zhou, "Multi-label learning with emerging new labels," *IEEE Transactions on Knowledge and Data Engineering*, 2018.
[49] R. H. Bartels and G. Stewart, "Solution of the matrix equation ax+ xb= c [f4]," *Communications of the ACM*, vol. 15, no. 9, pp. 820–826, 1972.
[50] A. Mehrotra and A. Dukkipati, "Generative adversarial residual pairwise networks for one shot learning," *arXiv preprint arXiv:1703.08033*, 2017.
[51] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5769–5779.
[52] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proceedings of The 32nd International Conference on Machine Learning*, 2015, pp. 2152–2161.
[53] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4166–4174.
[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
[57] 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*, vol. 36, no. 3, pp. 453–465, 2014.
[58] S. Kolouri, M. Rostami, Y. Owechko, and K. Kim, "Joint dictionaries for zero-shot learning," in *The Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 3431–3439.
[59] 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 *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 52–68.



Zhengming Ding (S'14) received the B.Eng. degree in information security and the M.Eng. degree in computer software and theory from University of Electronic Science and Technology of China (UESTC), China, in 2010 and 2013, respectively. He received the Ph.D. degree from the Department of Electrical and Computer Engineering, Northeastern University, USA in 2018. He is a faculty member affiliated with Department of Computer, Information and Technology, Indiana University-Purdue University Indianapolis since 2018. His research interests include machine learning and computer vision. Specifically, he devotes himself to develop scalable algorithms for challenging problems in transfer learning and deep learning scenario. He received the National Institute of Justice Fellowship during 2016-2018. He was the recipients of the best paper award (SPIE 2016) and best paper candidate (ACM MM 2017).



Ming Shao (S'11-M'16) received the B.E. degree in computer science, the B.S. degree in applied mathematics, and the M.E. degree in computer science from Beihang University, Beijing, China, in 2006, 2007, and 2010, respectively. He received the Ph.D. degree in computer engineering from Northeastern University, Boston MA, 2016. He is a tenure-track Assistant Professor affiliated with College of Engineering at the University of Massachusetts Dartmouth since 2016 Fall. His current research interests include sparse modeling, low-rank matrix analysis, deep learning, and applied machine learning on social media analytics. He was the recipient of the Presidential Fellowship of State University of New York at Buffalo from 2010 to 2012, and the best paper award winner/candidate of IEEE ICDM 2011 Workshop on Large Scale Visual Analytics, and ICME 2014. He has served as the reviewers for many IEEE Transactions journals including TPAMI, TKDE, TNNLS, TIP, and TMM. He has also served on the program committee for the conferences including AAAI, IJCAI, and FG. He is a member of IEEE.



Yun Fu (S'07-M'08-SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xian Jiaotong University, China, respectively, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, respectively. He is an interdisciplinary faculty member affiliated with College of Engineering and the College of Computer and Information Science at Northeastern University since 2012. His research interests are Machine Learning, Computational Intelligence, Big Data Mining, Computer Vision, Pattern Recognition, and Cyber-Physical Systems. He has extensive publications in leading journals, books/book chapters and international conferences/workshops. He serves as associate editor, chairs, PC member and reviewer of many top journals and international conferences/workshops. He received seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; nine Best Paper Awards from IEEE, IAPR, SPIE, SIAM; many major Industrial Research Awards from Google, Samsung, and Adobe, etc. He is currently an Associate Editor of the IEEE Transactions on Neural Networks and Leaning Systems (TNNLS) and IEEE Transactions on Image Processing (TIP). He is fellow of IAPR and SPIE, a Lifetime Senior Member of ACM, Lifetime Member of AAAI, OSA, and Institute of Mathematical Statistics, member of AAAS, Global Young Academy (GYA), INNS and Beckman Graduate Fellow during 2007-2008.