IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 27, 2020, accepted September 5, 2020, date of publication September 8, 2020, date of current version September 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022805

Learning Discriminative Projection With Visual Semantic Alignment for Generalized Zero Shot Learning

PENGZHEN DU, HAOFENG ZHANG, AND JIANFENG LU, (Member, IEEE)
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Corresponding author: Pengzhen Du (dupengzhen@njust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61872187, and in part by the 111 Program under Grant B13022.

ABSTRACT Zero Shot Learning (ZSL) aims to solve the classification problem with no training sample, and it is realized by transferring knowledge from source classes to target classes through the semantic embeddings bridging. Generalized ZSL (GZSL) enlarges the search scope of ZSL from only the seen classes to all classes. A large number of methods are proposed for these two settings, and achieve competing performance. However, most of them still suffer from the domain shift problem due to the existence of the domain gap between the seen classes and unseen classes. In this article, we propose a novel method to learn discriminative features with visual-semantic alignment for GZSL. We define a latent space, where the visual features and semantic attributes are aligned, and assume that each prototype is the linear combination of others, where the coefficients are constrained to be the same in all three spaces. To make the latent space more discriminative, a linear discriminative analysis strategy is employed to learn the projection matrix from visual space to latent space. Five popular datasets are exploited to evaluate the proposed method, and the results demonstrate the superiority of our approach compared with the state-of-the-art methods. Beside, extensive ablation studies also show the effectiveness of each module in our method.

INDEX TERMS Generalized zero-shot learning, linear discriminative analysis, visual semantic alignment, prototype synthesis.

I. INTRODUCTION

With the development of deep learning technique, the task of image classification has been transfered to large scale datasets, such as ImageNet [1], and achieved the level of human-beings [2]. Does it mean that we are already to solve large-scale classification problems? Two questions should be answered: 1) Can we collect enough samples of all the classes appeared all over the world for training? 2) Can the trained model with limited classes be transfered to other classes without retraining? The first question cannot be given an affirmative answer because there are 8.7 million classes only in animal species [3] and over 1000 new classes are emerging everyday. Therefore, many researchers moved their focus to the second question by employing transfer learning [4], [5] and Zero-shot Learning (ZSL) [6].

ZSL tries to recognize the classes that have no labeled data available during training, and is usually implemented by employing auxiliary semantic information, such as semantic attributes [7] or word embeddings [8], which is similar to the process of human recognition of new categories. For example, a child who has not seen a ''zebra'' before but knows that a ''zebra'' looks like a ''horse'' and has ''white and black stripes'', will be able to recognize a ''zebra'' very easily when he/she actually sees a zebra.

Since the concept of ZSL was first proposed [9], many ZSL methods have been proposed and most of them try to solve the inherent domain shift problem [10]–[13], which is caused by the domain gap between the seen classes and unseen classes. Although these methods can alleviate the domain shift problem and achieve certain effect, their performance are limited due to their negligence of unseen classes. To fully solve the domain shift problem, Fu *et al.* [14] assumed that the labeled seen samples and the unlabeled unseen samples can be both utilized during training, which is often called transductive learning. This type of method can significantly alleviate the domain shift problem and achieve the state-of-the-art performance [15]–[17], but the unlabeled

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Li.

unseen data usually is inaccessible during training in realistic scenarios.

In addition, conventional inductive learning often assumes that the upcoming test data belongs to the unseen classes, which is also unreasonable in reality because we cannot have the knowledge of the ascription of the future data in advance. Therefore, Chao *et al.* suggested to enlarge the search scope for test data form only the unseen classes to all classes [18], including both seen and unseen categories, which is illustrated in Fig. 1. To better solve the domain shift problem on the more realistic GZSL setting, many synthetic based methods have been proposed [19]–[22]. They often train a deep generative network to synthesize unseen data from its corresponding attribute by applying the frameworks of Generative Adversarial Network (GAN) [23] or Variational Auto-Encoder (VAE) [24], and then the synthesized data and the labeled seen data are combined to train a supervised close-set classification model. The synthetic based methods can also achieve state-of-the-art performance, but there is a serious problem that when a totally new object emerges the trained model will inevitably fail unless new synthetic samples are generated and retrained with previous samples.

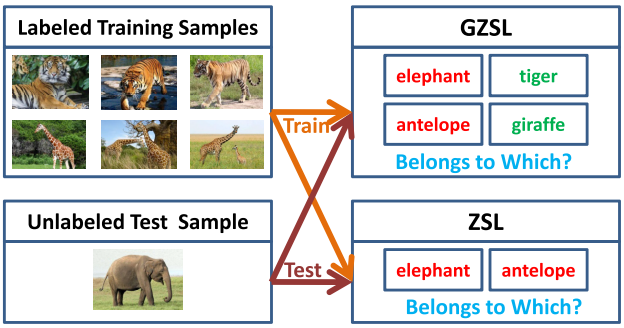


FIGURE 1. An illustration of the difference between ZSL and GZSL.

To solve the above mentioned problems, in this article, we proposed a novel method to learn discriminative projections with visual semantic alignment in a latent space for GZSL, and the proposed framework is illustrated in Fig. 2. In this framework, to solve the domain shift problem, we define a latent space to align the visual and semantic prototypes, which is realized by assuming that each prototype is a linear combination of others, including both seen and unseen ones. With this constraint, the seen and unseen categories are combined together and thus can reduce the domain gap between them. Besides, to make the latent space more discriminative, a Linear Discriminative Analysis (LDA) strategy is employed to learn the projection matrix from visual space to latent space, which can significantly reduce the within class variance and enlarge the between class variance. At last, we conduct experiments on five popular datasets to evaluate the proposed method. The contributions of our method is summarized as follows,

1) We proposed a novel method to solve the domain shift problem by learning discriminative projections with visual semantic alignment in latent space;
2) A linear discriminative analysis strategy is employed to learn the projection from visual space to latent space, which can make the projected features in the latent space more discriminative;
3) We assume that each prototype in all three spaces, including visual, latent and semantic, is a linear sparse combination of other prototypes, and the sparse coefficients for all three spaces are the same. This strategy can establish a link between seen classes and unseen classes, reduce the domain gap between them and eventually solve the domain shift problem;
4) Extensive experiments are conducted on five popular datasets, and the result shows the superiority of our method. Besides, detailed ablation studies also show that the proposed method is reasonable.

The main content of this article is organized as follows: In section II we briefly introduce some related existing methods for GZSL. Section III describes the proposed method in detail, and Section IV gives the experimental results and makes comparison with some existing state-of-the-art methods on several metrics. Finally in section V, we conclude this article.

II. RELATED WORKS

In this section, we will briefly review some related ZSL and GZSL works for the domain shift problem.

A. COMPATIBLE METHODS

Starting from the proposed ZSL concept [9], many ZSL methods have been emerging in recent several years. Due to the existence of the gap between the seen and unseen classes, an inherent problem, called domain shift problem, limits the performance of ZSL. These methods often project a visual sample into semantic space, where Nearest Neighbor Search (NNS) is conducted to find the nearest semantic prototype and its label is assigned to the test sample. Kodirov *et al.* tried to use an autoencoder structure to preserve the semantic meaning from visual features, and thus to solve the domain shift problem [13]. Zhang *et al.* exploited a triple verification, including an orthogonal constraint and two reconstruction constraints, to solve the problem and achieved a significant improvement. Akata *et al.* proposed to view attribute-based image classification as a label-embedding problem that each class is embedded in the space of attribute vectors [25], they employed pair-wise training strategy that the projected positive pair in the attribute space should have shorter distance than that of negative pair. However, the performance of these method are limited due to their negligence of unseen classes during training.

In addition, conventional ZSL assumes that the upcoming test sample belongs to the target classes, which is often unreasonable in realistic scenarios. Therefore, Chao *et al.* extended the search scope from only unseen classes to all classes, including both seen and unseen categories [18]. Furthermore, Xian *et al.* re-segmented the five popular benchmark datasets to avoid the unseen classes from overlapping with the

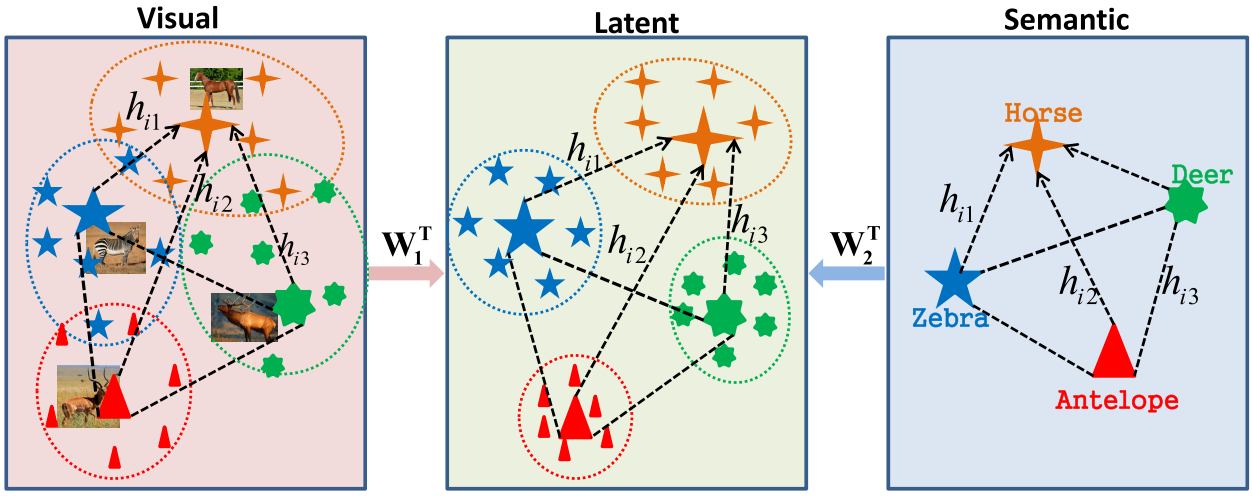


FIGURE 2. Illustration of the proposed method. The visual features and the semantic attributes are both projected into the latent space, where the projected vectors are scattered between classes and clustered within classes. Besides, the combinational coefficients are kept same in all three spaces to make the visual-semantic alignment.

categories in ImageNet [26]. Beside, they also proposed a new harmonic metric to evaluate the performance of GZSL, and release the performance of some state-of-the-art method on the new metric and datasets. From then on, many methods have been proposed on this more realistic setting. For example, Zhang *et al.* proposed a probabilistic approach to solve the problem within the NNS strategy [27]. Liu *et al.* designed a Deep Calibration Network (DCN) to enable simultaneous calibration of deep networks on the confidence of source classes and uncertainty of target classes [28]. Pseudo distribution of seen samples on unseen classes is also employed to solve the domain shift problem on GZSL [29]. Besides, there are many other methods developed for this more realistic setting [30], [31].

B. SYNTHETIC BASED METHODS

To solve the domain shift problem, synthetic based methods have attracted wide interest among researchers since they can obtain very significant improvement compared with traditional compatible methods.

Long *et al.* [32] firstly tried to utilize the unseen attribute to synthesize its corresponding visual features, and then train a fully supervised model by combining both the seen data and the synthesized unseen features. Since then, more and more synthetic based methods have being proposed [8], [30], [31], [33], and most of them are based on GAN [23] or VAE [24] because adversarial learning and VAE can facilitate the networks to generate more realistic samples [34], [35]. CVAE-ZSL [36] exploits a conditional VAE (cVAE) to realize the generation of unseen samples. Xian *et al.* proposed a f-CLSWGAN method to generate sufficiently discriminative CNN features by training a Wasserstein GAN with a classification loss [19]. Huang *et al.* [37] tried to learn a visual generative network for unseen classes by training three component to evaluate the closeness of an image feature and a class embedding, under the combination of cyclic consistency loss and dual adversarial loss. Dual Adversarial Semantics-Consistent Network (DASCN) [20] learns two GANs, namely primal GAN and dual GAN, in a unified framework, where the primal GAN learns to synthesize semantics-preserving and inter-class discriminative visual features and the dual GAN enforces the synthesized visual features to represent prior semantic knowledge via semantics-consistent adversarial learning.

Although these synthetic based methods can achieve excellent performance, they all suffer from a common serious problem that when an object of a new category emerges, the model should be retrained with the new synthesized samples of the new category. Different from these GAN or VAE based synthetic methods, our approach is a compatible one, which does not have the previous mentioned problem, and it can still accept new category without retraining even though there will be a little performance degradation.

C. TRANSDUCTIVE METHODS

Fu *et al.* tried to include the unlabeled unseen data in training, which is often called transductive learning, to solve the domain shift problem and achieved a surprising improvement [14]. Unsupervised Domain Adaptation (UDA) [38] formulates a regularized sparse coding framework, which utilizes the unseen class labels' projections in the semantic space, to regularize the learned unseen classes projection thus effectively overcoming the projection domain shift problem. QFSL [15] maps the labeled source images to several fixed points specified by the source categories in the semantic embedding space, and the unlabeled target images are forced to be mapped to other points specified by the target categories. Zhang *et al.*proposed a explainable Deep Transductive Network (DTN) by training on both labeled seen data and unlabeled unseen data, the proposed network exploits a KL Divergence constraint to iteratively refine the probability of classifying unlabeled instances by learning from their high

confidence assignments with the assistance of an auxiliary target distribution [17]. Although these transductive methods can achieve significant performance and outperform most of conventional inductive ZSL methods, the target unseen samples are usually inaccessible in realistic scenarios.

III. METHODOLOGY

A. PROBLEM DEFINITION

Let $\boldsymbol{Y} = \{y_1, \cdots, y_s\}$ and $\boldsymbol{Z} = \{z_1, \cdots, z_u\}$ denote a set of s seen and u unseen class labels, and they are disjoint $\boldsymbol{Y} \cap \boldsymbol{Z} = \emptyset$. Similarly, let $\boldsymbol{A_Y} = \{\boldsymbol{a_{y1}}, \cdots, \boldsymbol{a_{ys}}\} \in \mathbb{R}^{l \times s}$ and $\boldsymbol{A_Z} = \{\boldsymbol{a_{z1}}, \cdots, \boldsymbol{a_{zu}}\} \in \mathbb{R}^{l \times u}$ denote the corresponding s seen and u unseen attributes respectively. Given the training data in 3-tuple of N seen samples: $(\boldsymbol{x}_1, \boldsymbol{a}_1, \boldsymbol{y}_1), \cdots, (\boldsymbol{x}_N, \boldsymbol{a}_N, \boldsymbol{y}_N) \subseteq \boldsymbol{X}_s \times \boldsymbol{A_Y} \times \boldsymbol{Y}$, where $\boldsymbol{X}_s$ is d-dimensional features extracted from N seen images. When testing, the preliminary knowledge is u pairs of attributes and labels:$(\widehat{\boldsymbol{a}}_1, \widehat{\boldsymbol{z}}_1), \cdots, (\widehat{\boldsymbol{a}}_u, \widehat{\boldsymbol{z}}_u) \subseteq \boldsymbol{A_Z} \times \boldsymbol{Z}$. Zero-shot Learning aims to learn a classification function $f\colon \boldsymbol{X}_u \rightarrow \boldsymbol{Z}$ to predict the label of the input image from unseen classes, where $\boldsymbol{x}_i \in \boldsymbol{X}_u$ is totally unavailable during training.

B. OBJECTIVE

In this subsection, we try to propose an novel idea to learn discriminative projection with visual semantic alignment for generalized zero shot learning, the whole architecture is illustrated in Fig. 2.

1) SAMPLING FROM PROTOTYPES

Suppose we have already know the prototypes of seen classes, the seen features should be sampled from these prototypes, so we can have the following constraint,

$$\mathcal{L}_{basic} = \|\boldsymbol{X}_s - \boldsymbol{P}_s \boldsymbol{Y}_s\|_F^2, \tag{1}$$

where, $\boldsymbol{P}_s$ is the prototypes of seen categories, $\boldsymbol{Y}_s$ is the one-hot labels of seen samples, and $\|\cdot\|_F^2$ denotes for the Frobenius norm.

2) PROTOTYPE SYNTHESIS

Here we think each class prototype can be described as the linear combination of other ones with corresponding reconstruction coefficients. The reconstruction coefficients are sparse because the class is only related with certain classes. Moreover, to make the combination more flexible, we define another latent space, and construct a sparse graph in all three space as,

$$\begin{aligned} \mathcal{L}_{syn} &= \|\boldsymbol{P} - \boldsymbol{P}\boldsymbol{H}\|_F^2 + \alpha\|\boldsymbol{C} - \boldsymbol{C}\boldsymbol{H}\|_F^2 \\ &\quad + \beta\|\boldsymbol{A} - \boldsymbol{A}\boldsymbol{H}\|_F^2, \\ s.t.\ & diag(\boldsymbol{H}) = 0, \end{aligned} \tag{2}$$

where, $\boldsymbol{H}$ is the coefficient matrix; $\boldsymbol{P} = [\boldsymbol{P}_s, \boldsymbol{P}_u]$, $\boldsymbol{P}_s$ and $\boldsymbol{P}_u$ are visual prototypes of seen classes and unseen classes respectively; $\boldsymbol{C} = [\boldsymbol{C}_s, \boldsymbol{C}_u]$, $\boldsymbol{C}_s$ and $\boldsymbol{C}_u$ are the prototypes of seen classes and unseen classes respectively in latent space; $\boldsymbol{A} = [\boldsymbol{A}_s, \boldsymbol{A}_u]$, α and β are the balancing parameters. We apply $diag(\boldsymbol{H}) = 0$ to avoid the trivial solution.

3) VISUAL-SEMANTIC ALIGNMENT

In the latent space, the prototypes are the projections from both visual space and semantic space, so the alignment can be represented as,

$$\mathcal{L}_{eqnarray} = \|\boldsymbol{W}_1^T\boldsymbol{P} - \boldsymbol{C}\|_F^2 + \|\boldsymbol{W}_2^T\boldsymbol{A} - \boldsymbol{C}\|_F^2, \tag{3}$$

where, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the projection matrices from visual space and semantic space respectively.

4) LINEAR DISCRIMINATIVE PROJECTION

In visual space, the features might not be discriminative, which is illustrated in Fig. 2, so the direct strategy is to cluster them within class and scatter them between classes. Linear Discriminative Analysis is the proper choice and we can maximize the following function to achieve such purpose,

$$\mathcal{L}_{LDA} = \frac{\boldsymbol{W}_1^T \boldsymbol{S}_B \boldsymbol{W}_1}{\boldsymbol{W}_1^T \boldsymbol{S}_W \boldsymbol{W}_1}, \tag{4}$$

where, $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$ are the between-class scatter matrix and within-class scatter matrix respectively.

C. SOLUTION

Since we have already defined the loss function for each constraint, we can combine them and obtain the final objective as follows,

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{syn} + \gamma\mathcal{L}_{eqnarray} - \kappa\mathcal{L}_{LDA} + \lambda\mathcal{L}_{basic} \\ + \theta(\|\boldsymbol{W}_1\|_F^2 + \|\boldsymbol{W}_2\|_F^2 + \|\boldsymbol{H}\|_F^2 + \|\boldsymbol{P}\|_F^2 + \|\boldsymbol{C}\|_F^2), \end{aligned} \tag{5}$$

where γ, κ and λ are the balancing coefficients.

1) INITIALIZATION

Since Eq. 5 is not joint convex over all variables, there is no close-form solution simultaneously. Thus, we propose an iterative optimization strategy to update a single unresolved variable each time. Because proper initialization parameters can not only improve the model performance but also increase the convergence speed, we further split the solution into two sub problems, *i.e.*, initializing the parameters with reduced constraints, and iterative optimizing them with the full constraints.

Initializing $\boldsymbol{H}$: Since $\boldsymbol{A}$ is known in advance, we initialize $\boldsymbol{H}$ first with the last term of Eq. 2. We exploit the following formulation as the loss function for $\boldsymbol{H}$,

$$\begin{aligned} &\mathcal{L}_{\boldsymbol{H}} = \beta\|\boldsymbol{A} - \boldsymbol{A}\boldsymbol{H}\|_F^2 + \theta\|\boldsymbol{H}\|_F^2, \\ &subject\ to\ diag(\boldsymbol{H}) = 0. \end{aligned} \tag{6}$$

To solve the constraint $diag(\boldsymbol{H}) = 0$, we calculate $\boldsymbol{H}$ once per column,

$$\begin{aligned} \boldsymbol{H}_i &= \underset{\boldsymbol{H}_i}{argmin}\beta\|\boldsymbol{A}_i - \boldsymbol{A}_{\setminus i}\boldsymbol{H}_i\|_F^2 + \theta\|\boldsymbol{H}_i\|_F^2 \\ &= (\boldsymbol{A}_{\setminus i}^T\boldsymbol{A}_{\setminus i} + \frac{\theta}{\beta}\boldsymbol{I})^{-1}\boldsymbol{A}_{\setminus i}^T\boldsymbol{A}_i, \end{aligned} \tag{7}$$

where $\boldsymbol{H}_i$ is the i_{th} column of $\boldsymbol{H}$ and the i_{th} entry of $\boldsymbol{H}_i$ is also removed, $\boldsymbol{A}_{\setminus i}$ is the matrix of $\boldsymbol{A}$ excluding the *ith* column.

Initializing $\boldsymbol{P}_s$*:* We use Eq. 1 to initialize $\boldsymbol{P}_s$, and the closed-form solution can be obtained as follows,

$$\boldsymbol{P}_s = \boldsymbol{X}_s\boldsymbol{Y}_s^T(\boldsymbol{Y}_s\boldsymbol{Y}_s^T + \theta\boldsymbol{I})^{-1}, \tag{8}$$

Initializing $\boldsymbol{P}_u$*:* Since there is no training data for unseen classes, we cannot use the similar initialization strategy as $\boldsymbol{P}_s$ to initialize $\boldsymbol{P}_u$. However, we have already get $\boldsymbol{H}$ with Eq. 7 in advance, it is easy to utilize $\boldsymbol{P}_s$ and $\boldsymbol{H}$ to calculate $\boldsymbol{P}_u$. The simplified loss function can be formulated as follows,

$$\begin{aligned}\mathcal{L}_{\boldsymbol{P}_u} &= \|[\boldsymbol{P}_s, \boldsymbol{P}_u] - [\boldsymbol{P}_s, \boldsymbol{P}_u]\boldsymbol{H}\|_F^2 + \theta\|\boldsymbol{P}_u\|_F^2 \\ &= \|[\boldsymbol{P}_s, \boldsymbol{P}_u] - [\boldsymbol{P}_s, \boldsymbol{P}_u]\begin{bmatrix}\boldsymbol{H}_{s1} & \boldsymbol{H}_{u1}\\ \boldsymbol{H}_{s2} & \boldsymbol{H}_{u2}\end{bmatrix}\|_F^2 + \theta\|\boldsymbol{P}_u\|_F^2 \\ &= \|[\boldsymbol{P}_s - \boldsymbol{P}_s\boldsymbol{H}_{s1} - \boldsymbol{P}_u\boldsymbol{H}_{s2}, \boldsymbol{P}_u - \boldsymbol{P}_s\boldsymbol{H}_{u1} - \boldsymbol{P}_u\boldsymbol{H}_{u2}]\|_F^2 \\ &\quad + \theta\|\boldsymbol{P}_u\|_F^2 \\ &= \|\boldsymbol{P}_u\boldsymbol{H}_{s2} + \boldsymbol{P}_s\boldsymbol{H}_{s1} - \boldsymbol{P}_s\|_F^2 \\ &\quad + \|\boldsymbol{P}_u(\boldsymbol{I} - \boldsymbol{H}_{u2}) - \boldsymbol{P}_s\boldsymbol{H}_{u1}\|_F^2 + \theta\|\boldsymbol{P}_u\|_F^2. \end{aligned} \tag{9}$$

By computing the derivative of Eq. 10 with respected to $\boldsymbol{P}_u$ and setting it to zero, we can obtain the following solution,

$$\begin{aligned}\boldsymbol{P}_u = &(\boldsymbol{P}_s\boldsymbol{H}_{s2}^T - \boldsymbol{P}_s\boldsymbol{H}_{s1}\boldsymbol{H}_{s2}^T + \boldsymbol{P}_s\boldsymbol{H}_{u1}(\boldsymbol{I} - \boldsymbol{H}_{u2}^T)) \\ &\times(\boldsymbol{H}_{s2}\boldsymbol{H}_{s2}^T + (\boldsymbol{I} - \boldsymbol{H}_{u2})(\boldsymbol{I} - \boldsymbol{H}_{u2}^T) + \theta\boldsymbol{I})^{-1}. \end{aligned} \tag{10}$$

Initializing $\boldsymbol{W}_1$*:* Since $\boldsymbol{C}$ is unknown till now, we cannot calculate $\boldsymbol{W}_1$ withe Eq. 3. The only way for $\boldsymbol{W}_1$ is to optimize Eq. 4, from which we can deduce the following formulation,

$$(\boldsymbol{W}_1^T\boldsymbol{S}_w\boldsymbol{W}_1)\boldsymbol{S}_B\boldsymbol{W}_1 = \boldsymbol{S}_w\boldsymbol{W}_1(\boldsymbol{W}_1^T\boldsymbol{S}_B\boldsymbol{W}_1). \tag{11}$$

If we define $\frac{\boldsymbol{W}_1^T\boldsymbol{S}_B\boldsymbol{W}_1}{\boldsymbol{W}_1^T\boldsymbol{S}_W\boldsymbol{W}_1} = \tau$, then $\boldsymbol{W}_1$ can be solved by obtaining the eigenvector of $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B$.

Initializing $\boldsymbol{C}$*:* Since $\boldsymbol{W}_1$ and $\boldsymbol{P}$ are already known, it is easy to initialize $\boldsymbol{C}$ with the first item of Eq. 3, and the solution is,

$$\boldsymbol{C} = \boldsymbol{W}_1^T\boldsymbol{P}. \tag{12}$$

Initializing $\boldsymbol{W}_2$*:* By employing the second item of Eq. 3, $\boldsymbol{W}_2$ can be solved with following formulation,

$$\boldsymbol{W}_2 = (\boldsymbol{A}\boldsymbol{A}^T + \theta\boldsymbol{I})^{-1}\boldsymbol{A}\boldsymbol{C}^T. \tag{13}$$

2) OPTIMIZATION

Since the initialized value of each variable has already been obtained, the optimization of them can be executed iteratively by fixing others.

Updating $\boldsymbol{H}$*:* Similar as that for initializing $\boldsymbol{H}$, we can obtain $\boldsymbol{H}$ once per column with the following loss function,

$$\begin{aligned}\mathcal{L}_{\boldsymbol{H}_i} = &\|\boldsymbol{P}_i - \boldsymbol{P}_{\setminus i}\boldsymbol{H}_i\|_F^2 + \alpha\|\boldsymbol{C}_i - \boldsymbol{C}_{\setminus i}\boldsymbol{H}_i\|_F^2 \\ &+ \beta\|\boldsymbol{A}_i - \boldsymbol{A}_{\setminus i}\boldsymbol{H}_i\|_F^2 + \theta\|\boldsymbol{H}_i\|_F^2. \end{aligned} \tag{14}$$

By taking the derivative of $\mathcal{L}_{\boldsymbol{H}_i}$ with respect to $\boldsymbol{H}_i$, and setting the result to 0, we can obtain the solution of $\boldsymbol{H}_i$ as follows,

$$\begin{aligned}\boldsymbol{H}_i = &(\boldsymbol{P}_{\setminus i}^T\boldsymbol{P}_{\setminus i} + \alpha\boldsymbol{C}_{\setminus i}^T\boldsymbol{C}_{\setminus i} + \beta\boldsymbol{A}_{\setminus i}^T\boldsymbol{A}_{\setminus i} + \theta\boldsymbol{I})^{-1} \\ &\times(\boldsymbol{P}_{\setminus i}^T\boldsymbol{P}_i + \alpha\boldsymbol{C}_{\setminus i}^T\boldsymbol{C}_i + \beta\boldsymbol{A}_{\setminus i}^T\boldsymbol{A}_i). \end{aligned} \tag{15}$$

Updating $\boldsymbol{P}_s$*:* By fixing other variables except $\boldsymbol{P}_s$, we can obtain the following loss function from Eq. 5,

$$\begin{aligned}\mathcal{L}_{\boldsymbol{P}_s} = &\|\boldsymbol{P}_s\boldsymbol{H}_{u1} + \boldsymbol{P}_u\boldsymbol{H}_{u2} - \boldsymbol{P}_u\|_F^2 \\ &+ \|\boldsymbol{P}_s(\boldsymbol{I} - \boldsymbol{H}_{s1}) - \boldsymbol{P}_u\boldsymbol{H}_{s2}\|_F^2 \\ &+ \gamma\|\boldsymbol{W}_1^T\boldsymbol{P}_s - \boldsymbol{C}_s\|_F^2 + \lambda\|\boldsymbol{X}_s - \boldsymbol{P}_s\boldsymbol{Y}_s\|_F^2 + \theta\|\boldsymbol{P}_s\|_F^2, \end{aligned} \tag{16}$$

which can be expanded as,

$$\begin{aligned}&(\gamma\boldsymbol{W}_1\boldsymbol{W}_1^T + \theta\boldsymbol{I})\boldsymbol{P}_s \\ &\quad + \boldsymbol{P}_s((\boldsymbol{I} - \boldsymbol{H}_{s1})(\boldsymbol{I} - \boldsymbol{H}_{s1}^T) + \boldsymbol{H}_{u1}\boldsymbol{H}_{u1}^T + \lambda\boldsymbol{Y}_s\boldsymbol{Y}_s^T) \\ &= \boldsymbol{P}_u\boldsymbol{H}_{s2}(\boldsymbol{I} - \boldsymbol{H}_{s1}^T) - \boldsymbol{P}_u\boldsymbol{H}_{u2}\boldsymbol{H}_{u1}^T + \boldsymbol{P}_u\boldsymbol{H}_{u1}^T \\ &\quad + \gamma\boldsymbol{W}_1\boldsymbol{C}_s + \lambda\boldsymbol{X}_s\boldsymbol{Y}_s^T. \end{aligned} \tag{17}$$

If we set $\widehat{\boldsymbol{A}} = \gamma\boldsymbol{W}_1\boldsymbol{W}_1^T + \theta\boldsymbol{I}$, $\widehat{\boldsymbol{B}} = (\boldsymbol{I} - \boldsymbol{H}_{s1})(\boldsymbol{I} - \boldsymbol{H}_{s1}^T) + \boldsymbol{H}_{u1}\boldsymbol{H}_{u1}^T + \lambda\boldsymbol{Y}_s\boldsymbol{Y}_s^T$, and $\widehat{\boldsymbol{C}} = \boldsymbol{P}_u\boldsymbol{H}_{s2}(\boldsymbol{I} - \boldsymbol{H}_{s1}^T) - \boldsymbol{P}_u\boldsymbol{H}_{u2}\boldsymbol{H}_{u1}^T + \boldsymbol{P}_u\boldsymbol{H}_{u1}^T + \gamma\boldsymbol{W}_1\boldsymbol{C}_s + \lambda\boldsymbol{X}_s\boldsymbol{Y}_s^T$, then Eq. 17 can be simplified to $\widehat{\boldsymbol{A}}\boldsymbol{P}_s + \boldsymbol{P}_s\widehat{\boldsymbol{B}} = \widehat{\boldsymbol{C}}$, which is a well-known Sylvester equation and can be solved efficiently by the Bartels-Stewart algorithm [39]. Therefore, Eq. 17 can be implemented with a single line of code $\boldsymbol{P}_s = sylvester(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{C}})$ in MATLAB.

Updating $\boldsymbol{P}_u$*:* Similar as that for $\boldsymbol{P}_s$, we fix other variables except $\boldsymbol{P}_u$, and obtain,

$$\begin{aligned}\mathcal{L}_{\boldsymbol{P}_u} = &\|\boldsymbol{P}_u\boldsymbol{H}_{s2} + \boldsymbol{P}_s\boldsymbol{H}_{s1} - \boldsymbol{P}_s\|_F^2 \\ &+ \|\boldsymbol{P}_u(\boldsymbol{I} - \boldsymbol{H}_{u2}) - \boldsymbol{P}_s\boldsymbol{H}_{u1}\|_F^2 \\ &+ \gamma\|\boldsymbol{W}_1^T\boldsymbol{P}_u - \boldsymbol{C}_u\|_F^2 + \theta\|\boldsymbol{P}_u\|_F^2. \end{aligned} \tag{18}$$

By taking the derivative of $\mathcal{L}_{\boldsymbol{P}_u}$ with respect to $\boldsymbol{P}_u$, and set the result to 0, we can obtain the following equation,

$$\begin{aligned}&(\gamma\boldsymbol{W}_1\boldsymbol{W}_1^T + \theta\boldsymbol{I})\boldsymbol{P}_u \\ &\quad + \boldsymbol{P}_u((\boldsymbol{I} - \boldsymbol{H}_{u2})(\boldsymbol{I} - \boldsymbol{H}_{u2}^T) + \boldsymbol{H}_{s2}\boldsymbol{H}_{s2}^T) \\ &= \boldsymbol{P}_s\boldsymbol{H}_{u1}(\boldsymbol{I} - \boldsymbol{H}_{u2}^T) - \boldsymbol{P}_s\boldsymbol{H}_{s1}\boldsymbol{H}_{s2}^T \\ &\quad + \boldsymbol{P}_s\boldsymbol{H}_{s2}^T + \gamma\boldsymbol{W}_1\boldsymbol{C}_u. \end{aligned} \tag{19}$$

Similarly, if we set $\widetilde{\boldsymbol{A}} = \gamma\boldsymbol{W}_1\boldsymbol{W}_1^T + \theta\boldsymbol{I}$, $\widetilde{\boldsymbol{B}} = (\boldsymbol{I} - \boldsymbol{H}_{u2})(\boldsymbol{I} - \boldsymbol{H}_{u2}^T) + \boldsymbol{H}_{s2}\boldsymbol{H}_{s2}^T$, and $\widetilde{\boldsymbol{C}} = \boldsymbol{P}_s\boldsymbol{H}_{u1}(\boldsymbol{I} - \boldsymbol{H}_{u2}^T) - \boldsymbol{P}_s\boldsymbol{H}_{s1}\boldsymbol{H}_{s2}^T + \boldsymbol{P}_s\boldsymbol{H}_{s2}^T + \gamma\boldsymbol{W}_1\boldsymbol{C}_u$, then Eq. 19 can be simplified to $\widetilde{\boldsymbol{A}}\boldsymbol{P}_u + \boldsymbol{P}_u\widetilde{\boldsymbol{B}} = \widetilde{\boldsymbol{C}}$, which can also be solved efficiently with $\boldsymbol{P}_u = sylvester(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}}, \widetilde{\boldsymbol{C}})$ in MATLAB.

Updating $\boldsymbol{C}$*:* If we only let $\boldsymbol{C}$ variable and make others fixed, Eq. 5 can be reduced as,

$$\begin{aligned}\mathcal{L}_{\boldsymbol{C}} = &\alpha\|\boldsymbol{C} - \boldsymbol{C}\boldsymbol{H}\|_F^2 + \gamma(\|\boldsymbol{W}_2^T\boldsymbol{A} - \boldsymbol{C}\|_F^2 \\ &+ \|\boldsymbol{W}_1^T\boldsymbol{P} - \boldsymbol{C}\|_F^2) + \theta\|\boldsymbol{C}\|_F^2. \end{aligned} \tag{20}$$

By taking the derivative of $\mathcal{L}_{\boldsymbol{C}}$ with respect to $\boldsymbol{C}$, and set the result to 0, we can obtain the solution of $\boldsymbol{C}$ as follows,

$$\begin{aligned}\boldsymbol{C} = &(\gamma\boldsymbol{W}_1^T\boldsymbol{P} + \gamma\boldsymbol{W}_2^T\boldsymbol{A})((2\gamma + \theta)\boldsymbol{I} \\ &+ \alpha(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}^T))^{-1}. \end{aligned} \tag{21}$$

Updating $\boldsymbol{W}_2$*:* As for $\boldsymbol{W}_2$, Eq. 5 can be simplified as,

$$\mathcal{L}_{\boldsymbol{W}_2} = \gamma\|\boldsymbol{W}_2^T\boldsymbol{A} - \boldsymbol{C}\|_F^2 + \theta\|\boldsymbol{W}_2\|_F^2. \tag{22}$$

By taking the derivative of $\mathcal{L}_{W_2}$ with respect to W_2, and set the result to 0, we can obtain the solution of W_2 as follows,

$$W_2 = (AA^T + \frac{\theta}{\gamma}I)^{-1}AC^T. \quad (23)$$

Updating W_1: Similar as W_2 for W_1, Eq. 5 can be reduced as,

$$\mathcal{L}_{W_1} = \gamma\|W_1^TP - C\|_F^2 - \kappa\frac{W_1^TS_BW_1}{W_1^TS_WW_1} + \theta\|W_2\|_F^2. \quad (24)$$

Due to the direct derivative of Eq. 22 will cause the negative order of W_1, we rewrite it as follows,

$$\mathcal{L}_{W_1} = \gamma\|W_1^TP - C\|_F^2 - \kappa(W_1^TS_BW_1 - \eta W_1^TS_WW_1) + \theta\|W_1\|_F^2, \quad (25)$$

where, η is a coefficient and set as the maximum eigenvalue of $S_W^{-1}S_B$ here.

By taking the derivative of $\mathcal{L}_{W_1}$ with respect to W_1, and set the result to 0, we can obtain the solution of W_1 as follows,

$$W_1 = (PP^T - \frac{\kappa}{\gamma}S_B + \frac{\kappa\eta}{\gamma}S_W + \theta I)^{-1}PC^T. \quad (26)$$

After these steps, the test sample can be classified by projecting it into the latent space and finding the nearest neighbor of it from C. The algorithm of the proposed method is described in Alg. 1.

Algorithm 1 The Training Framework of the Proposed Method

Input:
The set of visual features of seen classes: X_s;
The set of one-hot labels of X_s: Y_s;
the set of semantic attributes, including both seen and unseen classes: A; The number of iterative time for optimization: *iter*;
The hyper-parameters: α, β, γ, λ, κ and θ;

Output:
The projection matrices: W_1 and W_2;
The visual prototypes of both seen and unseen classes: P_s and P_u;
The latent prototypes of both seen and unseen classes: C_s and C_u;

Initializing H with Eq. 7 once per clolumn;
Initializing P_s and P_u with Eq. 8 and Eq. 10 respectively;
Initializing W_1 with the eigenvectors of $S_W^{-1}S_B$ from Eq. 11;
Initializing C with Eq. 12;
for $k = 1 \rightarrow iter$ **do**
  Update H with Eq. 15 once per column;
  Update P_s with Eq. 17 by applying $P_s = sylvester(\widehat{A}, \widehat{B}, \widehat{C})$;
  Update P_u with Eq. 19 by applying $P_u = sylvester(\widetilde{A}, \widetilde{B}, \widetilde{C})$;
  Update C with Eq. 21;
  Update W_2 with Eq. 23;
  Update W_1 with Eq. 26;
end for
return W_1, W_2, P_s, P_u and C.

IV. EXPERIMENTS

In this section, we first briefly review some datasets applied in our experiments, then some settings for the experiments are given, and at last we show the experiment results and ablation study to demonstrate the performance of the proposed method.

A. DATASETS

In this experiment, we utilize five popular datasets to evaluate our method, *i.e.*, SUN (SUN attribute) [40], CUB (Caltech-UCSD-Birds 200-2011) [41], AWA1 (Animals with Attributes) [42], AWA2 [42] and aPY (attribute Pascal and Yahoo) [43]. Among them, SUN and CUB are fine-grained datasets while AWA1/2 and aPY are coarse-grained ones. The detailed information of the datasets is summarized in Tab. 1, where "SS" denotes the number of Seen Samples for training, "TS" and "TR" refer to the numbers of unseen class samples and seen class samples respectively for testing.

TABLE 1. Summary of the five employed datasets. "SS" denotes the number of Seen Samples for training, "TS" and "TR" refer to the numbers of unseen class samples and seen class samples respectively for testing.

Datasets	Dimension		Class Number		Samples Number		
	Feat.	*Att.*	*Seen*	*Unseen*	*SS*	*TS*	*TR*
SUN	2048	102	645	72	10320	1440	2580
CUB	2048	312	150	50	7057	2967	1764
AWA1	2048	85	40	10	19832	4958	5685
AWA2	2048	85	40	10	23527	5882	7913
aPY	2048	64	20	12	5932	7924	1483

Moreover, we use the same split setting, which is proposed by Xian *et al.* in [26], for all the comparisons with the state-of-the-art methods listed in Tab. 2.

B. EXPERIMENTAL SETTING

We exploit the extracted features with ResNet [2] as our training and testing samples, which are released by Xian *et al.* [26], and all the the settings, including both attributes and classes split, are also the same as those in [26]. In addition, there are six hyper-parameters α, β, γ, λ, κ and θ. Since θ is only used to control the regularization terms, we set it with a small value 1×10^{-4}. As for other five parameters, due to the fact that different dataset usually performs well with different parameters, thus we choose our hyper-parameters from the set of $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ by adopting a cross validation strategy. To be specific, we hereby compare the difference of ZSL cross-validation to conventional cross-validation for machine learning approaches. Compared to inner-splits of training samples within each class, ZSL problem requires inter splits by in turn regarding part of seen classes as unseen, for example, 20% of the seen classes are selected

TABLE 2. Comparison of our method and state-of-the-art methods under GZSL setting. Bold font stands for the best result of the corresponding column and '-' means not reported.

	SUN			CUB			AWA1			AWA2			APY		
Method	*ts*	*tr*	*H*	*ts*	*tr*	*H*	*ts*	*tr*	*H*	*ts*	*tr*	*H*	*ts*	*tr*	*H*
DAP [9]	4.2	25.1	7.5	1.7	67.9	3.3	0.0	**88.7**	0.0	0.0	84.7	0.0	4.8	78.3	9.0
CONSE [10]	6.8	**39.9**	11.6	1.6	**72.2**	3.1	0.4	88.6	0.8	0.5	**90.6**	1.0	0.0	**91.2**	0.0
CMT [44]	8.1	21.8	11.8	7.2	49.8	12.6	0.9	87.6	1.8	0.5	90.0	1.0	0.0	91.2	0.0
LATEM [45]	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
SSE [46]	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	8.1	82.5	14.8	0.2	78.9	0.4
ALE [25]	21.8	33.1	26.3	23.7	62.8	34.4	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
DEVISE [47]	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
SJE [11]	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
ESZSL [48]	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [49]	7.0	43.4	13.4	11.5	70.9	19.8	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
SAE [13]	8.8	18.0	11.8	7.8	54.0	13.6	1.8	77.1	3.5	1.1	82.2	2.2	1.1	82.2	2.2
GFZSL [50]	0.0	39.6	0.0	0.0	45.7	0.0	1.8	80.3	3.5	2.5	80.1	4.8	0.0	83.3	0.0
LAGO [51]	18.8	33.1	23.9	21.8	73.6	33.7	23.8	67.0	35.1	-	-	-	-	-	-
PSEUDO [52]	19.0	32.7	24.0	23.0	51.6	31.8	22.4	80.6	35.1	-	-	-	15.4	71.3	25.4
KERNEL [53]	21.0	31.0	25.1	24.2	63.9	35.1	18.3	79.3	29.8	18.9	82.7	30.8	11.9	76.3	20.5
TRIPLE [54]	18.2	28.9	22.3	26.5	62.3	37.2	27.0	67.9	38.6	28.5	66.7	39.9	16.1	66.9	25.9
VZSL [22]	15.2	23.8	18.6	17.1	37.1	23.8	22.3	77.5	34.6	21.7	78.6	34.0	8.4	75.5	15.1
LESAE [55]	21.9	34.7	26.9	24.3	53.0	33.3	19.1	70.2	30.0	21.8	70.6	33.3	12.7	56.1	20.1
LESD [56]	15.2	19.8	17.2	14.6	38.5	21.2	12.6	71.0	21.4	15.3	71.5	25.2	11.8	49.3	19.0
Ours	**22.7**	34.9	**27.5**	**30.5**	48.4	**37.4**	**33.5**	87.4	**48.4**	**34.2**	78.2	**47.6**	**21.3**	83.5	**33.9**

as the validational unseen classes in our experiments, and the parameters of best average performance of 5 executions are selected as the final optimal parameters for each dataset. It should be noted that the parameters may not be the most suitable for the test set, because the labels of test data are strictly inaccessible during training.

C. COMPARISON WITH BASELINES

In this subsection, we conduct experiments to compare our method with some baselines methods. In addition to the methods evaluated in [26], we also compare our method with some newly proposed frameworks, such as GFZSL [50], LAGO [51], PSEUDO [52], KERNEL [53], TRIPLE [54], LESAE [55], LESD [56] and VZSL [22]. To be specific, we directly cite the results from [26] or from their own papers if it is feasible, otherwise we re-implement them according to the methods described in their own papers. We exploit the harmonic mean H to evaluate our model under the GZSL setting, and it is defined as,

$$H = \frac{2 \times acc_{tr} \times acc_{ts}}{acc_{tr} + acc_{ts}}, \tag{27}$$

where, acc_{tr} and acc_{ts} are the accuracies of test samples from seen classes and unseen categories respectively, and we adopt the average per-class top-1 accuracy as the final result.

Since our method utilizes both seen and unseen semantic attributes and focuses on the more realistic GZSL setting, we do not report the result on conventional ZSL setting. The results of our method and the compared method are recorded in Tab. 2, and the best result of each column is highlighted with bold font. From this table, we can clearly discover that our method can outperform the state-of-the-art methods on both *ts* and H. Concretely, our method can improve *ts* by 0.8% on SUN, 4.0% on CUB, 6.5% on AWA1, 5.7% on AWA2 and 5.2% on APY, and enhance H by 0.6% on SUN, 0.2% on CUB, 9.8% on AWA1, 7.7% on AWA2 and 8.0% on APY respectively compared with the best methods LESAE and TRIPLE. Besides, compared to those existing methods that have high *tr* but low *ts* and H, such as DAP and CONSE, our method can achieve more balanced performance on *ts* and *tr* and eventually obtain a significant improvement on H. We ascribe this improvement to the discriminative projection with LDA and the prototype synthesis with both seen and unseen classes, because the first one can make the projected features from same class cluster and from different classes disperse, and the second one combines both seen and unseen classes into a unified framework to alleviate the domain shift problem.

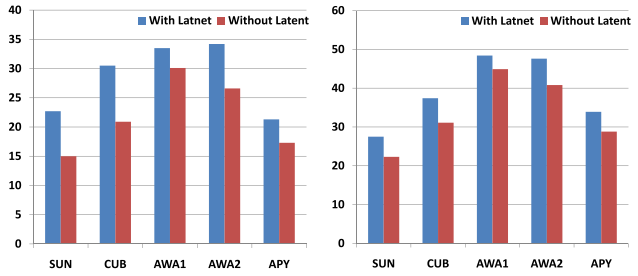
D. ABLATION STUDY

1) EFFECT OF LATENT SPACE

In our method, we utilize the latent space as the intermediate space for both visual and semantic features and we have claimed that this space can obtain more discriminative projection and alleviate the domain shift problem. Therefore, it is necessary to verify whether this space can achieve such statement. In this subsection, we remove the latent prototypes from Eq. 2 and Eq. 3, modify the discriminative projection item with LDA, and redefine the three loss functions as follows,

$$\begin{cases} \mathcal{L}_{syn} = \|\boldsymbol{P} - \boldsymbol{PH}\|_F^2 + \beta\|\boldsymbol{A} - \boldsymbol{AH}\|_F^2 \\ \mathcal{L}_{eqnarray} = \|\boldsymbol{W}^T\boldsymbol{P} - \boldsymbol{A}\|_F^2 \\ \mathcal{L}_{LDA} = \dfrac{\boldsymbol{W}^T\boldsymbol{S}_B\boldsymbol{W}}{\boldsymbol{W}^T\boldsymbol{S}_W\boldsymbol{W}}. \end{cases} \tag{28}$$

We replace the three items $\mathcal{L}_{syn}$, $\mathcal{L}_{eqnarray}$ and $\mathcal{L}_{LDA}$ in Eq. 5 and re-optimize it, the performance with the new loss function is illustrated in Fig. 3, form which it can be clearly seen that the accuracies with the latent space are higher than those



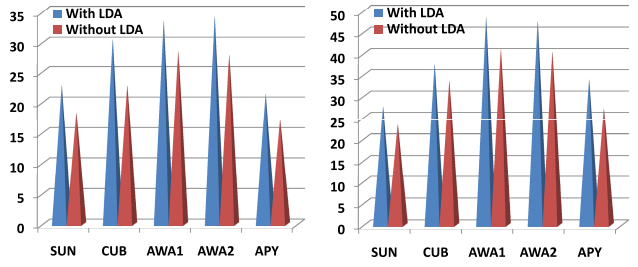
(a) ts (b) H

FIGURE 3. Illustration of performance with and without the latent space in our proposed method.

without the latent space on all five datasets. To be specific, we can obtain more improvement on SUN and CUB than on AWA and APY, especially on the metric *ts*. We attribute this phenomenon to that the learned vectors in latent space can preserve more discriminative characteristic, and the employment of unified synthesis framework on both seen and unseen classes can well alleviate the domain shift problem.

2) EFFECT OF LDA

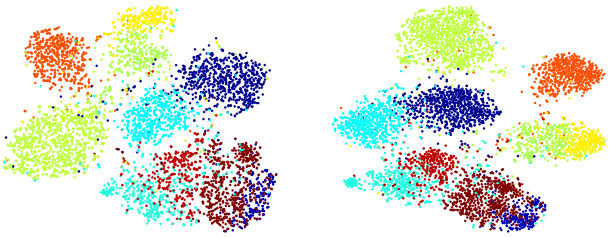
In our method, we utilize the LDA strategy to project visual features into latent space to make them more discriminative, so it is necessary to find how much this mechanism can improve the final performance. In this experiment, we remove the loss item $\mathcal{L}_{LDA}$ from Eq. 5, and conduct the evaluation on the five popular datasets. The experimental results are illustrated in Fig. 4, from which it can be clearly observed that the method with LDA can significantly outperform that without LDA constraint. This phenomenon reveals that the LDA constraint plays a very important role in improving the performance due to its powerful ability of learning discriminative features in latent space.



(a) ts (b) H

FIGURE 4. Illustration of performance with and without LDA on AWA1 in our proposed method.

Moreover, to more intuitively display the improvement of our method, we also show the distributions of unseen samples on AWA1 with and without LDA in latent space with t-SNE [57]. The results are illustrated in Fig. 5, from which it can be discovered that the distribution with LDA is more compact than that without LDA in each class, especially those classes at the bottom of the figure. This situation further prove that LDA is necessary for our method to learn discriminative features in latent space.



(a) Without LDA (b) With LDA

FIGURE 5. Illustration of distributions of unseen samples with and without LDA on AWA1 in our proposed method.

3) DIFFERENT DIMENSION OF LATENT SPACE

Since we apply latent space in our method, It is necessary to discuss the effect of the dimension of the latent space on the final performance. In our experiment, we take AWA1 as an example and change the dimension of the latent space from 5 to 60 to show the performance change. The performance curves are recorded in Fig. 6, from which it can be clearly seen that the curves monotonically increase for both *ts* and *H*, and nearly stop increase when the dimension is larger than 50. This phenomenon reveals that we can obtain better performance when we have larger dimension in latent space, but this increasing will stop when it reaches the number of classes.

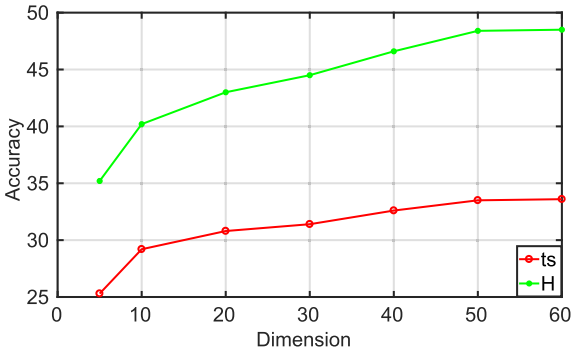


FIGURE 6. Illustration of performance with different dimension of the latent space in our proposed method.

E. ZERO SHOT IMAGE RETRIEVAL

In this subsection, we conduct experiments to show zero shot retrieval performance of our proposed method. In this task, we apply the semantic attributes of each unseen category as the query vector, and compute the mean Average Precision (mAP) of the returned images. MAP is a popular metric for evaluating the retrieval performance, it comprehensively evaluates the accuracy and ranking of returned results, and defined as,

$$mAP = \frac{1}{u}\sum_{i=1}^{u}\left(\frac{1}{r_i}\sum_{j=1}^{r_i}\frac{j}{p_i(j)}\right), \tag{29}$$

where, r_i is the number of returned correct images from the dataset corresponding to the ith query attribute, $p_i(j)$ represents the position of the jth retrieved correct image among

all the returned images according to the *i*th query attribute. In this experiment, the number of returned images equals the number of the samples in unseen classes.

For the convenience of comparison, we employ the standard split of the four datasets, including SUN, CUB, AWA1 and aPY, which can be found in [26], and the results are shown in Tab. 3. The values of the baseline methods listed in Tab. 3 are directly cited from [58]. The results show that our method can outperform the baselines on all four datasets, especially on the coarse-grained dataset AWA1, which reveals that our method can make the prototypes in latent space more discriminative.

TABLE 3. The mean Average Precision (mAP) for zero shot image retrieval.

Methods	SUN	CUB	AWA1	aPY	Average
SSE [46]	58.9	4.7	46.25	15.4	31.3
JSLE [59]	76.5	23.9	66.5	32.7	49.9
SynC [49]	74.3	34.3	65.4	30.4	51.1
ISEC [60]	52.7	25.3	68.1	36.9	45.8
MFMR [61]	77.4	30.6	70.8	45.6	56.2
LESD [56]	76.6	31.3	71.2	40.3	54.9
Ours	**77.2**	**36.5**	**74.8**	**45.3**	**58.5**

V. CONCLUSION

In this article, we have proposed a novel method to learn discriminative features with visual-semantic alignment for generalized zero shot learning. in this method, we defined a latent space, where the visual features and semantic attributes are aligned. We assumed that each prototype is the linear combination of others and the coefficients are the same in all three spaces, including visual, latent and semantic. To make the latent space more discriminative, a linear discriminative analysis strategy was employed to learn the projection matrix from visual space to latent space. Five popular datasets were exploited to evaluate the proposed method, and the results demonstrated the superiority compared with the state-of-the-art methods. Beside, extensive ablation studies also showed the effectiveness of each module of the proposed method.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[3] A. Zhao, M. Ding, J. Guan, Z. Lu, T. Xiang, and J.-R. Wen, "Domain-invariant projection learning for zero-shot recognition," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1027–1038.

[4] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Transfer independently together: A generalized framework for domain adaptation," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 2144–2155, Jun. 2019.

[5] J. Li, Y. Wu, and K. Lu, "Structured domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1700–1713, Aug. 2017.

[6] H. Zhang, Y. Long, W. Yang, and L. Shao, "Dual-verification network for zero-shot learning," *Inf. Sci.*, vol. 470, pp. 43–57, Jan. 2019.

[7] Y. Long and L. Shao, "Describing unseen classes by exemplars: Zero-shot learning using grouped simile ensemble," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 907–915.

[8] H. Zhang, Y. Long, L. Liu, and L. Shao, "Adversarial unseen visual feature synthesis for zero-shot learning," *Neurocomputing*, vol. 329, pp. 12–20, Feb. 2019.

[9] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[10] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–10.

[11] 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. (CVPR)*, Jun. 2015, pp. 2927–2936.

[12] H. Zhang, Y. Long, and L. Shao, "Zero-shot leaning and hashing with binary visual similes," *Multimedia Tools Appl.*, vol. 78, no. 17, pp. 24147–24165, Sep. 2019.

[13] 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. 3174–3183.

[14] 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.*, 2014, pp. 584–599.

[15] J. Song, C. Shen, Y. Yang, Y. Liu, and M. Song, "Transductive unbiased embedding for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1024–1033.

[16] Y. Yu, Z. Ji, X. Li, J. Guo, Z. Zhang, H. Ling, and F. Wu, "Transductive zero-shot learning with a self-training dictionary approach," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2908–2919, Oct. 2018.

[17] H. Zhang, L. Liu, Y. Long, Z. Zhang, and L. Shao, "Deep transductive network for generalized zero shot learning," *Pattern Recognit.*, vol. 105, Sep. 2020, Art. no. 107370.

[18] W.-L. Chao, C. Soravit, 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.*, 2016, pp. 52–68.

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

[20] J. Ni, S. Zhang, and H. Xie, "Dual adversarial semantics-consistent network for generalized zero-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6146–6157.

[21] F. Zhang and G. Shi, "Co-representation network for generalized zero-shot learning," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 7434–7443.

[22] W. Wang, Y. Pu, V. K. Verma, K. Fan, Y. Zhang, C. Chen, P. Rai, and L. Carin, "Zero-shot learning via class-conditioned deep generative models," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4211–4218.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[24] P. D. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–14.

[25] 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.

[26] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning—The good, the bad and the ugly," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4582–4591.

[27] H. Zhang, H. Mao, Y. Long, W. Yang, and L. Shao, "A probabilistic zero-shot learning method via latent nonnegative prototype synthesis of unseen classes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2361–2375, Jul. 2020.

[28] S. Liu, M. Long, J. Wang, and M. I. Jordan, "Generalized zero-shot learning with deep calibration network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2005–2015.

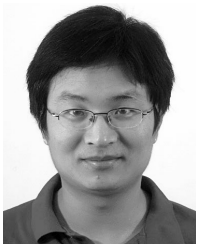
[29] H. Zhang, J. Liu, Y. Yao, and Y. Long, "Pseudo distribution on unseen classes for generalized zero shot learning," *Pattern Recognit. Lett.*, vol. 135, pp. 451–458, Jul. 2020.

[30] 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.*, Jun. 2018, pp. 4281–4289.

[31] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero- and few-shot learning via aligned variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8247–8255.

[32] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han, "From zero-shot learning to conventional supervised classification: Unseen visual data synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1627–1636.
[33] L. Chen, H. Zhang, J. Xiao, W. Liu, and S.-F. Chang, "Zero-shot visual recognition using semantics-preserving adversarial embedding networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1043–1052.
[34] W. Zhang, Z. Zuo, Y. Wang, and Z. Zhang, "Double-integrator dynamics for multiagent systems with antagonistic reciprocity," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 4110–4120, Sep. 2020.
[35] Y. Zhang and Y. Liu, "Nonlinear second-order multi-agent systems subject to antagonistic interactions without velocity constraints," *Appl. Math. Comput.*, vol. 364, Jan. 2020, Art. no. 124667.
[36] A. Mishra, S. K. Reddy, A. Mittal, and H. A. Murthy, "A generative model for zero shot learning using conditional variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2188–2196.
[37] H. Huang, C. Wang, P. S. Yu, and C.-D. Wang, "Generative dual adversarial network for generalized zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 801–810.
[38] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Unsupervised domain adaptation for zero-shot learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2452–2460.
[39] R. H. Bartels and G. W. Stewart, "Solution of the matrix equation AX + XB = c [F4]," *Commun. ACM*, vol. 15, no. 9, pp. 820–826, Sep. 1972.
[40] G. Patterson, C. Xu, H. Su, and J. Hays, "The SUN attribute database: Beyond categories for deeper scene understanding," *Int. J. Comput. Vis.*, vol. 108, nos. 1–2, pp. 59–81, May 2014.
[41] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
[42] 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.*, Jun. 2009, pp. 951–958.
[43] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1778–1785.
[44] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2013, pp. 935–943.
[45] 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. (CVPR)*, Jun. 2016, pp. 69–77.
[46] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4166–4174.
[47] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.
[48] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. ICML*, 2015, pp. 2152–2161.
[49] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5327–5336.
[50] V. K. Verma and P. Rai, "A simple exponential family framework for zero-shot learning," in *Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discovery Databases*, Skopje, Macedonia. Springer, 2017, pp. 792–808.
[51] Y. Atzmon and G. Chechik, "Probabilistic AND-OR attribute grouping for zero-shot learning," 2018, *arXiv:1806.02664*. [Online]. Available: http://arxiv.org/abs/1806.02664
[52] T. Long, X. Xu, Y. Li, F. Shen, J. Song, and H. T. Shen, "Pseudo transfer with marginalized corrupted attribute for zero-shot learning," in *Proc. ACM Multimedia Conf. Multimedia Conf. MM*, 2018, pp. 1802–1810.
[53] H. Zhang and P. Koniusz, "Zero-shot kernel learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7670–7679.
[54] H. Zhang, Y. Long, Y. Guan, and L. Shao, "Triple verification network for generalized zero-shot learning," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 506–517, Jan. 2019.
[55] Y. Liu, Q. Gao, J. Li, J. Han, and L. Shao, "Zero shot learning via low-rank embedded semantic AutoEncoder," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2490–2496.
[56] Z. Ding, M. Shao, and Y. Fu, "Low-rank embedded ensemble semantic dictionary for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2050–2058.
[57] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[58] Z. Ding, M. Shao, and Y. Fu, "Generative zero-shot learning via low-rank embedded semantic dictionary," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2861–2874, Dec. 2019.
[59] Z. Zhang and V. Saligrama, "Zero-shot learning via joint latent similarity embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 6034–6042.
[60] M. Bucher, S. Herbin, and F. Jurie, "Improving semantic embedding consistency by metric learning for zero-shot classification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 6034–6042.
[61] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3798–3807.

PENGZHEN DU received the Ph.D. degree from the Nanjing University of Science and Technology, in 2015. He is currently an Assistant Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include computer vision, evolutionary computation, robotics, and deep learning.



HAOFENG ZHANG received the B.Eng. and Ph.D. degrees from the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2003 and 2007, respectively. From December 2016 to December 2017, he was an Academic Visitor with the University of East Anglia, Norwich, U.K. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include computer vision and robotics.

JIANFENG LU (Member, IEEE) received the B.S. degree in computer software and the M.S. and Ph.D. degrees in pattern recognition and intelligent system from the Nanjing University of Science and Technology, Nanjing, China, in 1991, 1994, and 2000, respectively. He is currently a Professor and the Vice Dean of the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include image processing, pattern recognition, and data mining.