Attribute Subspaces for Zero-Shot Learning

Lei Zhou[b,a,1], Yang Liu[c,1], Xiao Bai[a,*], Na Li[d], Xiaohan Yu[e], Jun Zhou[e], Edwin R. Hancock[f]

[a]*School of Computer Science and Engineering, Beihang University, Beijing, China*
[b]*School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China*
[c]*College of Computer Science and Technology, Zhejiang University, Hangzhou, China*
[d]*Institute of Artificial Intelligence, Beihang University, Beijing, China*
[e]*Institute for Integrated and Intelligent Systems, Griffith University, Nathan, Australia*
[f]*Department of Computer Science, University of York, York, U.K.*

Abstract

Zero-shot learning (ZSL) aims to recognize unseen categories without corresponding training samples, which is a practical yet challenging task in computer vision and pattern recognition community. Current state-of-the-art locality-based ZSL methods aim to learn the explicit locality of discriminative attributes, which may suffer from insufficient class-level attribute supervision. In this paper, we introduce an Attribute Subspace learning method for ZSL (AS-ZSL) to learn implicit attribute composition, which is more general than attribute localization with only class-level attribute supervision. AS-ZSL exploits subspace representations that can effectively capture the intrinsic composition of high-dimensional image features and the diversity within attribute appearance. Furthermore, we develop a subspace distance based triplet loss to improve the distinguishability of the attribute subspace representation. Attribute subspace learning module is only needed for the training phase to jointly learn discriminative global features. This leads to a compact inference phase. Furthermore, the proposed AS-ZSL can be naturally extended to adapt to the transductive ZSL setting using a novel self-supervised training strategy. Extensive experimental results on several widely used ZSL datasets, *i.e.*, CUB, AwA2, and SUN, demonstrate the advantage of AS-ZSL compared with the state-of-the-art under different ZSL settings.

*Corresponding author
Email address: baixiao@buaa.edu.cn (Xiao Bai)
[1]Lei Zhou and Yang Liu contribute equally to this work. The main work was done in Beihang University.

Keywords: Zero-shot learning, attribute localization, subspace representation, attribute subspaces, self-supervised learning

1. Introduction

Deep learning based computer vision and pattern recognition tasks have achieved impressive progress due to the abundant human-annotated data, such as image classification, object localization, object detection, and semantic segmentation. Especially object recognition has reached human-like performance benefited from the large-scale labeled dataset, like ImageNet. However, manual labeling of such large-scale training data requires lots of time and expense. Zero-shot learning (ZSL) [1] was presented to classify categories not appearing in the training set. Thus, there is no need for a large number of labeled data. ZSL relies on semantic descriptions of both seen and unseen categories to transfer knowledge from seen to unseen categories. This process is inspired by the human cognitive system that describes a novel object as knowing attribute primitives.

As shown in Fig. 1, depending on the training and test conditions, the ZSL task can be subdivided into four different settings, *i.e.*, a) inductive ZSL (IZSL), b) transductive ZSL (TZSL), c) conventional ZSL, and d) generalized ZSL (GZSL). Semantic descriptions build the bridge from seen categories to unseen categories, which are represented as quantized attribute semantic vectors (semantic space). One ZSL branch aims to learn an embedding between the visual and semantic spaces [2, 3]. However, their performance is often dissatisfactory under the GZSL setting because the embedding model is trained only with images of seen categories, leading to a significant bias towards seen classes. Due to progress in generative models, another ZSL branch [4, 5] proposes to generate samples of unseen categories for training. The generative model-based methods can alleviate the category imbalance problem. Most of these two branches of ZSL approaches utilize pre-trained global image features or extract global features by end-to-end deep neural networks. However, only global features are difficult to capture the fine-grained differences between seen and unseen categories[6], which is crucial for the ZSL task.

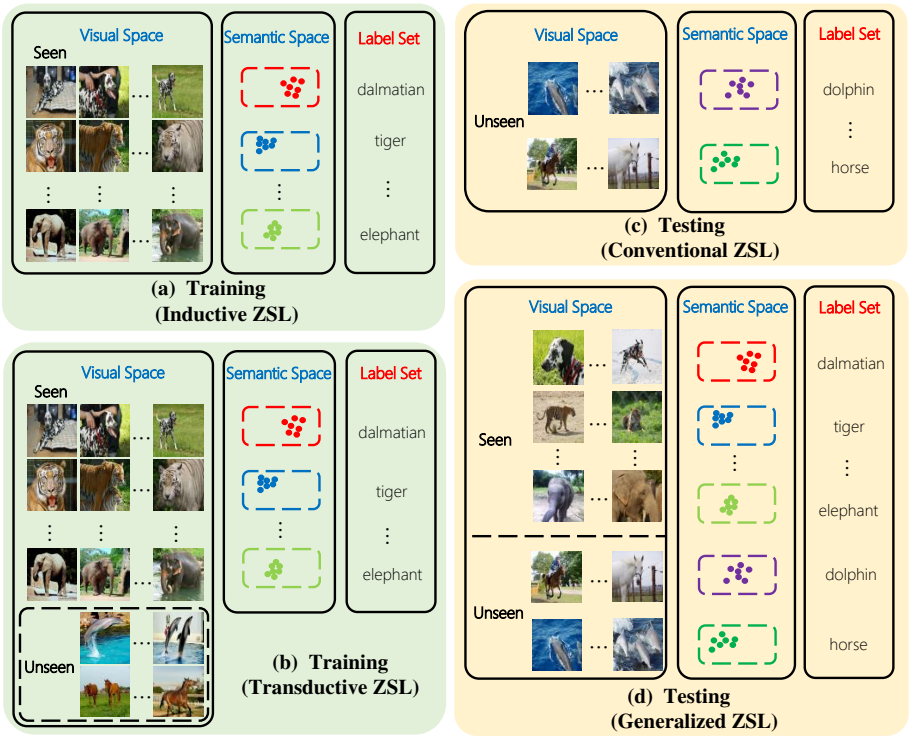
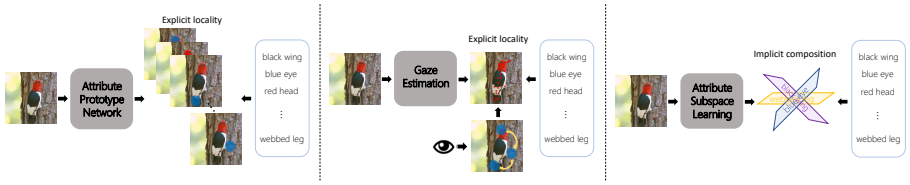


Figure 1: Illustration of different zero-shot learning settings. In the training phase, ZSL includes two settings: (a) **Inductive ZSL**: training with labelled images of only the seen category; (b) **Transductive ZSL**: training with images of both seen and unseen categories and labels of only the seen category. ZSL also has two settings in the test phase: (c) **Conventional ZSL**: test images all belong to the unseen category; (d) **Generalized ZSL**: test images belong to both seen and unseen categories.

Several part-based ZSL models [7, 8, 9] have recently attempted to utilize semantic information as a guide to extract discriminative region features. They learn the region embedding of attribute semantics while neglecting the importance of discriminative attribute localization. Then, locality-based models, APN [10] and GEM-ZSL [11], were proposed as a solution that learn the explicit locality of discriminative attributes with class-level attributes or by human gaze supervision. This significantly improved the GZSL performance. However, the generalization of these methods still suffers from the unsatisfactory part or locality learning due to the lack of a supervision signal to guide the network to focus on the correct local regions. Additionally, the spatial locality



(a) APN (b) GEM-ZSL (c) AS-ZSL

Figure 2: Comparison of state-of-the-art locality-based ZSL methods (a) APN [10], (b) GEM-ZSL [11], and our proposed (c) AS-ZSL. APN and GEM-ZSL both aim to learn the explicit locality of discriminative attributes. The performance is not satisfying for two reasons: 1) only class-level attributes or human gaze (attention) supervision is not enough to learn the accurate locality of attributes; 2) the spatial locality and visual appearance of attributes in different images have numerous diversity. Alternatively, the AS-ZSL learns implicit attributes composition by subspace representation with class-level attributes supervision.

and visual appearance of attributes in different images have significant diversity. It is therefor difficult to forcefully learn their locality.

As shown in Fig. 2, in a manner different from attribute localization, we propose a novel Attribute Subspace learning method for ZSL (AS-ZSL) that explicitly learns the attribute composition using class-level attributes supervision. Motivated by the success of subspace representation learning in many visual tasks [12, 13, 14], we adopt it to capture the intrinsic composition of high-dimensional image features and the diversity within attributes appearance.

Avoiding learning explicit attribute locality with only class-level supervision, the proposed AS-ZSL learns implicit attribute composition by representing an attribute as a subspace extracted from its local CNN features. The framework of the AS-ZSL is illustrated in Fig. 3. The entire model consists of three modules: a) Image Encoder (IE), b) Attribute Subspace Learning (ASL), and c) Cosine Metric Learning (CML). Firstly, global image features are extracted by the IE module. Then, the ASL module learns the attribute subspace by singular value decomposition (SVD). We utilize a subspace distance to measure the similarity between attribute subspaces. Further based on the subspace distance, we design a triplet loss to enhance the discriminability of the subspace representation. Finally, a cosine metric based feature space is learned for the nearest neighbor classification in the CML module. The AS-ZSL supports an end-to-end training manner. Specifically, the ASL module is only used in the training phase

to jointly learn discriminative global features. In the inference phase, we obtain the global features of the test image through the IE and then search for its nearest neighbor in the cosine metric space to obtain its category label. Since the subspace representation learning process can be accomplished under unsupervised conditions, AS-ZSL can be naturally extended to adapt to the transductive ZSL setting (AS-TZSL). Under the TZSL setting, the unlabeled images are available for unseen classes. Their subspace representation can be learned in the same way without labels. For the triplet loss construction, we utilize a self-supervised training strategy. To do so, we adopt the data augmentation method to a specific unseen image to gain its positive sample. Then images from seen classes are negative samples that generate triplets for TZSL training.

We summarize the main contributions of this paper:

1. We propose a novel attribute subspace method for discriminative attribute representation learning. Compared with the existing explicit attribute locality learning methods, our AS-ZSL learns the intrinsic attribute composition with only class-level attribute supervision. To the best of our knowledge, AS-ZSL is the first work to introduce subspace representation learning to investigate attribute composition for the ZSL task.
2. We design a subspace distance based triplet loss, which enhances the discriminability of the attribute subspace representation. In addition, the attribute subspace learning module is only used for training to jointly learn discriminative global features. Therefore, the inference phase of AS-ZSL is very compact.
3. The AS-ZSL is naturally extended to further adapt to the transductive ZSL setting since the subspace representation learning is an unsupervised procedure. Moreover, we propose a novel data augmentation based self-supervised training strategy for triplet loss learning of the TZSL.
4. We conduct extensive experiments on widely used ZSL datasets. Results validate the advantage of our proposed AS-ZSL compared with the state-of-the-art under different ZSL settings.

2. Related Works

2.1. Embedding and Generative Models for IZSL

Early ZSL approaches build embedding from the visual space to the semantic space. Typically, ALE [2] learns the images and the attribute descriptions into a semantic space where the compatibility between them can be measured. DeViSE [15] proposes a deep visual-semantic embedding model to map images into a semantic embedding space. Thus it can recognize unannotated images using the semantic relationships between labels. However, since the embedding is from the high-dimensional visual space to the low-dimensional semantic space, hubness problem is inevitable [16]. To mitigate the issue, embedding from the semantic space to the visual space or embedding both the semantic and visual features into an intermediate space have been proposed [17]. Recently, an abundance of generative model based ZSL methods [4, 18, 19] have been proposed to improve the performance of generalized ZSL (GZSL) using synthesized unseen class features to alleviate the training data imbalance problem. For example, f-CLSWGAN [4] utilizes the Wasserstein GAN (WGAN) to generate unseen class features with a classification loss. Cycle-CLSWGAN [20] exploits the idea of CycleGAN to map the generated visual features back to their original semantic features with a cycle consistency loss. Thus it can learn more robust and authentic features for unseen classes. Since the training of GAN-based models is unstable and difficult, VAE based methods achieve more robust performance. CADA-VAE [21] adopts a cross-aligned VAE to align both the visual and semantic distributions generated from VAE in a common latent space. More recently, IZF [22] proposes to generate samples of unseen classes by a generative flow network, which obtains superior performance on ZSL task.

2.2. Part and Locality-based Models for IZSL

Although generative model based methods have achieved encouraging performance for GZSL, learning global image features alone cannot effectively represent the fine-grained differences between seen and unseen classes. To tackle this issue, several part-based ZSL models [23, 24] try to use semantic information to guide the learning of

more discriminative local features. However, they learn the region embedding of attribute semantics but neglect the important discriminative attribute localization [25]. More recently, the locality-based method APN [10] jointly learns discriminative global and local features using class-level attributes for supervision by an attribute prototype network. Another locality-based method GEM-ZSL [11] proposes a novel gaze estimation module to mimic human attention when recognizing an unseen class. This is supervised by both class-level attributes and human gaze. However, these methods have attempted to learn explicit locality of discriminative attributes with only class-level supervision that cannot obtain satisfy locality information. One major reason is that a class-level attribute vector ignores the spatial structure and diversity of the attributes. There is insufficient supervision signal to effectively guide the network to focus on the correct local regions.

We propose an implicit attribute subspace learning method to tackle this problem. Since the subspace representation can capture the intrinsic structure of the global features, which could learn the attribute composition with only class-level attributes supervision. Although we have not learned the explicit locality of attributes, accurate attribute composition information can also significantly improve the ZSL task.

2.3. Transductive ZSL

Under the TZSL setting, the unlabeled images are available during training while the labels of these images are still unavailable. With unseen samples for training, the domain shift problem can be alleviated. Thus, the core challenge of TZSL is how to train the image of unseen classes without label. SABR-T [26] learns generative adversarial networks to generate latent space representations of both the seen and unseen class images, then a conditional probability distribution of latent representations of the semantic labels are transferred from seen classes to unseen classes. GXE [27] utilizes a self-training strategy which generates pseudo labels for unseen images and then alternately updatas the classifier and generator. SDGN [28] proposes a self-supervised learning method and designs a cross-domain triplet mining mechanism to connect the seen and unseen classes. Recently, Zero-VAE-GAN [29] presents a joint generative model for feature generation with two self-training strategies. This enables the

model to further mitigate the strong bias towards seen classes. VMAN [30] introduces a weighted encoder-decoder framework for virtual mainstay sample generation. An instance-category matching regularization strategy is proposed to exploit the unseen data for training the required weights.

In this paper, we extend our AS-ZSL to solve the TZSL task by a self-supervised triplet loss. Firstly, we can learn the subspace representation for unseen images under unsupervised conditions. Then, we adopt the data augmentation method to the case of unseen images to obtain positive samples. The negative samples can be selected from seen classes to generate triplets for training.

2.4. Subspace Representation Learning

Subspace representation learning has been widely used on various computer vision tasks. It has demonstrated excellent capabilities in representing the intrinsic structural information present in high-dimensional data, especially for specific types of data, such as face images [31], different identities [32], classes of similar objects [33], and video clips [34]. JFSSL [35] projects multimodal data into a common subspace, then the similarity of the different modalities of the subspace representation can be measured to address the cross-modal retrieval task. KRP-FS [34] introduces a kernelized low-rank feature subspace to represent the sequences of human action videos to solve the action recognition problem. HSS-SMM [32] develops a matrix classifier based binary code learning framework to transform the subspace representation into a hash code for efficient subspace search, and which can be applied to face recognition, gesture recognition, video retrieval, and action recognition. DSN [36] represents classes by subspace bases and learns dynamic subspace classifiers to improve few-shot classification.

Since the attribute composition of classes is an alternative to the attribute localization for ZSL, additionally, the attribute composition could be learned by subspace representation learning with only class-level attribute supervision. In this work, we propose a novel attribute subspace method to learn the attribute composition for the ZSL task.

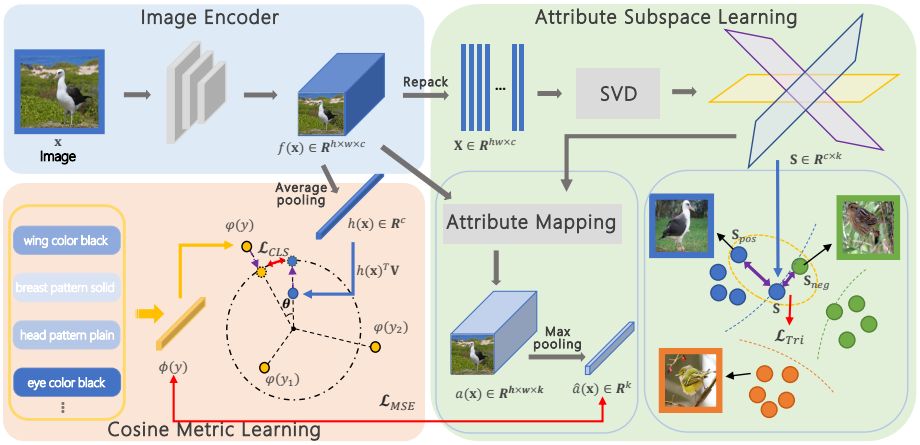


Figure 3: The pipeline of our AS-ZSL. The whole model consists of three modules: the Image Encoder (IE), the Attribute Subspace Learning (ASL), and the Cosine Metric Learning (CML). Firstly, global image features are extracted by the IE module. Then, the ASL module learns the attribute composition by singular value decomposition (SVD) and triplet loss constraint, which is used to jointly train the global features. Finally, in the CML module, a cosine metric based feature space is learned for the nearest neighbor classification. Red arrows indicate the losses.

3. Proposed Method

3.1. Problem Setting and Notations

We define the image space as $\mathcal{X}$, which contains both seen classes, $\mathcal{X}^S$, and unseen classes, $\mathcal{X}^U$, such that the image space is the union of the seen and unseen classes: $\mathcal{X} = \mathcal{X}^S \cup \mathcal{X}^U$. Let $\mathcal{S} = \{(\mathbf{x}, y, \varphi(y))|\mathbf{x} \in \mathcal{X}^S, y \in \mathcal{Y}^S, \varphi(y) \in \phi^S\}$ denote the seen class set, where $\mathcal{S}$ consists of triplets $(\mathbf{x}, y, \varphi(y))$. In each triplet, $\mathbf{x}$ is an image from the image space $\mathcal{X}^S$, y is its class label from the label space $\mathcal{Y}^S$, and $\varphi(y) \in \mathbb{R}^k$ is the attribute vector. Let $\mathcal{U} = \{(\mathbf{x}^u, u, \varphi(u))|\mathbf{x}^u \in \mathcal{X}^U, u \in \mathcal{Y}^U, \varphi(u) \in \phi^U\}$ denote the unseen test set, where $\mathcal{U}$ consists of triplets $(\mathbf{x}^u, u, \varphi(u))$. In each triplet, $\mathbf{x}^u$ is an image from the unseen image space $\mathcal{X}^U$, u is its unseen class label from the unseen label space $\mathcal{Y}^U$. The seen classes $\mathcal{Y}^S$ and unseen classes $\mathcal{Y}^U$ are disjoint. The combined attribute space is defined as $\phi = \phi^S \cup \phi^U$. The aim of conventional ZSL is to recognize images only belong to unseen classes, *i.e.*, $\mathcal{X}^U \rightarrow \mathcal{Y}^U$. The goal of GZSL is to recognize images from both seen and unseen classes, *i.e.*, $\mathcal{X} \rightarrow \mathcal{Y}^U \cup \mathcal{Y}^S$.

Existing locality-based ZSL approaches invariably suffer from insufficient class-level attribute supervision when learning the explicit locality of attributes. In this work, from a new perspective, we propose a novel attribute subspace learning method, AS-ZSL, to mitigate this problem. Our AS-ZSL method is the first work that learns the implicit attribute composition by subspace representation learning to solve the ZSL task. The overall architecture of AS-ZSL is illustrated in Fig. 3. AS-ZSL firstly learns the attribute subspace representation, which is obtained by the reconstruction of CNN features for this image. Then, the similarity between attributes can be calculated using a subspace distance. We design a triplet loss based on the subspace distance to learn a more discriminative attribute subspace. A ground-truth attribute vector supervised mean square loss is simultaneously used to enhance the subspace learning. Finally, the global features are projected to the semantic space for the nearest neighbor search to realize zero-shot recognition.

In the remainder of this section, we first introduce the concept of subspace representation. Then, we present the subspace distance based triplet loss underpinning our method and together with the end-to-end training process in the context of zero-shot recognition. Finally, we extend the TZSL setting using a self-supervised training strategy, which can effectively learn the attribute subspace from unseen images.

3.2. Attribute Subspace Learning

3.2.1. Subspace Representation Learning

Since learning explicit attribute locality with class-level supervision is difficult, we focus on investigating the implicit attribute representation learning from the global image features. As subspace representation learning has the ability that extracts the intrinsic composition of high-dimensional features [32], we utilize the subspace representation technique to learn the attribute composition, *i.e.*, local discriminative features from the global image features. Specifically, given an image $\mathbf{x}$, the feature map extracted by backbone ResNet-101 from the entire image is denoted as $f(\mathbf{x}) \in \mathbb{R}^{h \times w \times c}$, where h, w and c are the height, width and channel of the feature, respectively. We repack the c-dimensional feature vectors at all the spatial location of the tensor $f(\mathbf{x})$ to form a matrix $\mathbf{X} \in \mathbb{R}^{hw \times c}$. Then, the subspace basis representation can be learned by

minimizing

$$\min_{\mathbf{S}} \|\mathbf{X} - \mathbf{X}\mathbf{S}\mathbf{S}^T\|_F \quad s.t.\ \mathbf{S}^T\mathbf{S} = \mathbf{I} \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm, the columns of $\mathbf{S} \in \mathbb{R}^{c\times k}$ are a set of orthonormal basis vectors, where k is the dimensionality of the annotated attribute vector. The objective appearing in (1) is a convex optimization problem from which we can calculate the optimal solution by singular value decomposition (SVD) of the matrix $\mathbf{X}$. The detailed solution process is shown in Algorithm. 1. We can then learn the attribute subspace representation $a(\mathbf{x})$ by embedding the feature map into the subspace, *i.e.*, $a(\mathbf{x}) = f(\mathbf{x})\mathbf{S} \in \mathbb{R}^{h\times w\times k}$.

Algorithm 1: Subspace representation learning.

Input: $f(\mathbf{x}) \in \mathbb{R}^{h\times w\times c}$

Output: $\mathbf{S} \in \mathbb{R}^{c\times k}$

Steps:

1. Repack the feature map tensor $f(\mathbf{x})$ into matrix $\mathbf{X} \in \mathbb{R}^{hw\times c}$.
2. Calculate the eigen decomposition of $\mathbf{X}\mathbf{X}^T$ to obtain the eigenvalues λ_i and the right-singular vectors v_i, $i = 1, 2, \ldots, hw$.
3. Calculate the singular values ϵ_i by the square root of λ_i, $\epsilon_i = \sqrt{\lambda_i}$.
4. Choose the top-k largest singular values from ϵ_i, then the optimal $\mathbf{S}$ consists of their corresponding right-singular vectors.

3.2.2. Subspace Distance based Triplet Loss

With the label of seen classes to hand, we can further utilize a triplet loss to enhance the subspace representation. The triplet loss is known as constraining the features of the same class aggregating and for the different classes which are decentralized. Here we introduce a widely used subspace distance [32] to calculate the similarity of different subspaces. The subspace distance is measured by the relative spatial position of two subspaces, which is revealed by the principal angles [37]. Let $\mathbf{S}_i = [\mathbf{p}_1, ..., \mathbf{p}_k]$ and $\mathbf{S}_j = [\mathbf{q}_1, ..., \mathbf{q}_k]$ be two subspaces learned from different image

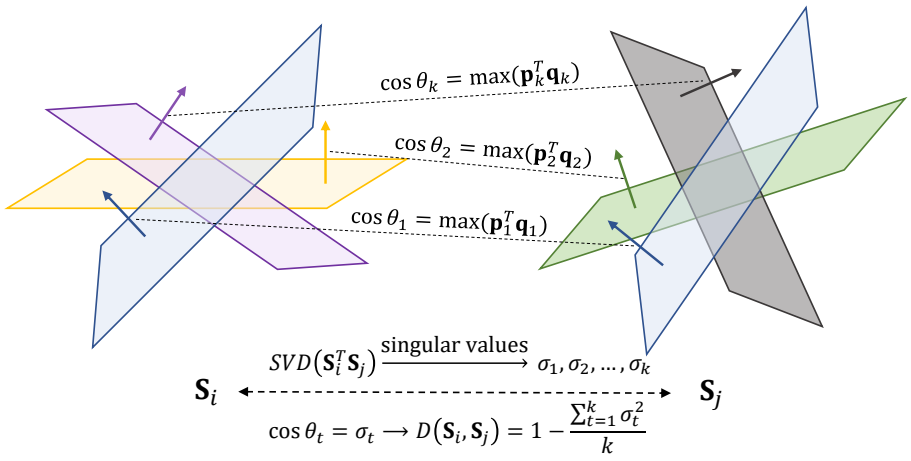


Figure 4: Illustration of the subspace distance. The principal angles are defined recursively by the largest dot product of the orthonormal basis from the two subspaces.

$\mathbf{x}_i$ and $\mathbf{x}_j$, $\mathbf{p}_1, ..., \mathbf{p}_k, \mathbf{q}_1, ..., \mathbf{q}_k \in \mathbb{R}^c$ are orthonormal bases. The principal angles $\{\theta_t\}_{t=1}^k \in [0, \pi/2]$ are defined recursively by the largest dot product of orthonormal bases of the two subspaces $\mathbf{S}_i$ and $\mathbf{S}_j$

$$\cos\theta_t = \max_{\substack{\mathbf{p}_t \in \mathbf{S}_i, \|\mathbf{p}_t\|_2 = 1 \\ \mathbf{p}_t^T [\mathbf{p}_1, ..., \mathbf{p}_{t-1}] = 0}} \max_{\substack{\mathbf{q}_t \in \mathbf{S}_j, \|\mathbf{q}_t\|_2 = 1 \\ \mathbf{q}_t^T [\mathbf{q}_1, ..., \mathbf{q}_{t-1}] = 0}} \mathbf{p}_t^T \mathbf{q}_t \tag{2}$$

As illustrated in Fig. 4, the first principal angle θ_1 is the smallest angle between a pair of orthonormal bases from the two subspaces. The t-th principal angle is then defined recursively. Since $\mathbf{S}_i$ and $\mathbf{S}_j$ are $c \times k$ matrices whose columns are orthonormal bases, the cosine of each principal angle can be obtained by a singular value of $\mathbf{S}_i^T \mathbf{S}_j$ [37]. We then calculate the singular values of $\mathbf{S}_i^T \mathbf{S}_j$ represented as $\sigma_1, \sigma_2, ..., \sigma_k$, where $1 \geq \sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k \geq 0$. Assuming that the principal angles satisfy the ordering $0 \leq \theta_1 \leq \theta_2 \leq ... \leq \theta_k \leq \pi/2$, we obtain

$$\cos\theta_t = \sigma_t, \quad t = 1, 2, ..., k \tag{3}$$

As a result the distance between subspaces $\mathbf{S}_i$ and $\mathbf{S}_j$ is given by

$$D(\mathbf{S}_i, \mathbf{S}_j) = 1 - \frac{\sum_{t=1}^{k} \cos^2\theta_t}{k} = 1 - \frac{\sum_{t=1}^{k} \sigma_t^2}{k} \tag{4}$$

Since the singular values have the property that $\sum_{t=1}^{k} \sigma_t^2 = \|\mathbf{S}_i^T \mathbf{S}_j\|_F^2$, we can rewrite the subspace distance as

$$D(\mathbf{S}_i, \mathbf{S}_j) = 1 - \frac{\|\mathbf{S}_i^T \mathbf{S}_j\|_F^2}{k} \tag{5}$$

where $D(\mathbf{S}_i, \mathbf{S}_j) \in [0, 1]$, the smaller value of $D(\mathbf{S}_i, \mathbf{S}_j)$ means more similarity between the two subspaces.

Triplet loss. Based on the subspace distance, we aim to make the attribute subspace of a specific class closer to those of the same class (positive sample) than the remaining different classes (negative sample). The triplet loss is defined as

$$\mathcal{L}_{Tri} = \sum_{i|y_i \in \mathcal{Y}^S} [D(\mathbf{S}_i, \mathbf{S}_{pos}) - D(\mathbf{S}_i, \mathbf{S}_{neg}) + \alpha]_+ \tag{6}$$

where $\mathbf{S}_{pos}$ and $\mathbf{S}_{neg}$ are the positive and negative samples of $\mathbf{S}_i$ respectively, *i.e.*, $y_{pos} = y_i$ and $y_{nag} \neq y_i$. α is a margin that is enforced between positive and negative sample pairs, and $[x]_+ = \max\{x, 0\}$.

3.2.3. Attribute Subspace Alignment

After mapping the feature map to the learned attribute subspace, we obtain the attribute subspace feature representation $a(\mathbf{x}) \in \mathbb{R}^{h \times w \times k}$. We then utilize a global max pooling operation across the dimensionality of h and w on $a(\mathbf{x})$ to predict the attribute response value $\hat{a}(\mathbf{x})$. To align the learned subspace representation with the discriminative attributes, a mean square error (MSE) loss is calculated with the ground-truth attribute vector $\varphi(y)$ as supervision

$$\mathcal{L}_{MSE} = \|\hat{a}(\mathbf{x}) - \varphi(y)\|_2^2 \tag{7}$$

where y is the ground-truth class of $\mathbf{x}$. The attribute subspace representation can be further improved by minimizing the MSE loss.

3.3. Cosine Metric Learning

Following the other end-to-end ZSL methods [10, 11], we choose pre-trained ResNet-101 as the Image Encoder to extract the feature of image $\mathbf{x}$, represented as $f(\mathbf{x}) \in$

$\mathbb{R}^{h\times w\times c}$. We then apply a global average pooling operation over the dimensionality of h and w to learn a global discriminative feature $h(\mathbf{x}) \in \mathbb{R}^c$ for classification.

In our method, the nearest neighbour search is performed in the semantic space. Thus, the global feature $h(\mathbf{x})$ is mapped into the semantic space by a linear layer $\mathbf{V} \in \mathbb{R}^{c\times k}$, where k is the dimensionality of the attribute vector.

Unlike previous work [10] which computes class logits by taking the dot product of the projected visual features and attribute vectors, our method calculates cosine similarity between the visual features and attribute vectors to constrain and reduce the variance of the neuron activations.

Using cosine similarity rather than dot products enables our models to generalize better, as noted in prior work [27]. First, we measure the cosine distance between the projected visual feature $h(\mathbf{x})^T\mathbf{V}$ and the y-th attribute vector $\varphi(y)$. Then the score function can be defined as

$$p(y|\mathbf{x}) = \frac{\exp(\sigma \cos(h(\mathbf{x})^T\mathbf{V}, \varphi(y)))}{\sum_{\hat{y}\in\mathcal{Y}^S} \exp(\sigma \cos(h(\mathbf{x})^T\mathbf{V}, \varphi(\hat{y})))} \tag{8}$$

where $\sigma = 20.0$. The classification loss L_{CLS} then can be written as

$$\mathcal{L}_{CLS} = -\log p(y|\mathbf{x}) \tag{9}$$

3.4. Zero-Shot Recognition

The full model is optimized in an end-to-end training manner. The overall loss is defined as

$$\mathcal{L} = \mathcal{L}_{CLS} + \beta_1\mathcal{L}_{Tri} + \beta_2\mathcal{L}_{MSE} \tag{10}$$

where β_1 and β_2 are hyper-parameters for the triplet loss and the MSE loss, respectively.

For the conventional ZSL setting, after training the entire model, the inference phase is accomplished in the learned cosine metric space. Given a test image $\mathbf{x}$, we first extract its feature through the trained Image Encoder and then map the feature into the semantic space, *i.e.*, the cosine metric space. The nearest attribute vector $\varphi(\hat{u})$ of the test image is then searched in the cosine metric space via

$$\hat{u} = \arg\max_{u\in\mathcal{Y}^U} \cos(h(\mathbf{x})^T\mathbf{V}, \varphi(u)) \tag{11}$$

For the GZSL setting, test images are the mixture of seen and unseen classes while there are only images of seen classes used for model training. As aforementioned, there is a significant bias towards predicting the seen classes. To alleviate this bias, we utilize a widely used strategy in GZSL to directly reduce the seen class scores by a calibration factor γ. The nearest neighbor search for GZSL can be realized by

$$\hat{y} = \underset{\tilde{y} \in \mathcal{Y}^U \cup \mathcal{Y}^S}{\arg\max} \left(\sigma \cos(h(\mathbf{x})^T \mathbf{V}, \varphi(\tilde{y})) - \gamma \mathbb{I}[\tilde{y} \in \mathcal{Y}^S] \right) \tag{12}$$

where $\mathbb{I} = 1$ if $\tilde{y}$ is a seen class and 0 for unseen classes.

4. Transductive ZSL

Since the subspace representation can be learned in an unsupervised manner by Eq. (1), we can easily extend our method to the specific TZSL setting. The critical challenge for TZSL is how to exploit the unlabeled images of the unseen classes to train the zero-shot learning classifier. In this paper, we propose a novel subspace based self-supervised strategy to cope with this problem. Firstly, we learn the subspace of unseen images $\mathbf{x}^u \in \mathcal{X}^U$ using the same manner with inductive setting

$$\begin{aligned} &\min_{\bar{\mathbf{S}}} \|\bar{\mathbf{X}} - \bar{\mathbf{X}}\bar{\mathbf{S}}\bar{\mathbf{S}}^T\|_F \\ &s.t.\ \bar{\mathbf{S}}^T\bar{\mathbf{S}} = \mathbf{I} \end{aligned} \tag{13}$$

here we use $\bar{\mathbf{S}} \in \mathbb{R}^{c \times k}$ to denote the subspace of the unseen images. $\bar{\mathbf{X}} \in \mathbb{R}^{hw \times c}$ obtained from the feature map tensor $f(\mathbf{x}^u)$. With the learned subspaces to hand, we then design a self-supervised triplet loss to exploit the unseen images for the purposes of training.

As illustrated in Fig. 5, during the training of the inductive ZSL, we sample positive and negative images both from the seen classes using the ground-truth labels. However, for the TZSL setting, unseen images have no label. To tackle this problem, for the training of TZSL, if the anchor image is from one of the seen classes, we sample its positive image from the same class and the negative image from both seen and unseen classes since the seen and unseen classes are disjoint. To cope with the condition that the anchor image is from one of the unseen classes, we utilize a random image augmentation

Figure 5: The triplet sampling strategy for the training phase under IZSL and TZSL settings. For the IZSL setting, we can sample triplets using the label of each seen image. For the TZSL setting, since unseen images have no label, we utilize the image augmentation technique on the anchor image of the unseen classes to obtain their positive samples and sample negative images from the disjoint seen classes.

technique (flip, crop, brightness) on the anchor image to obtain the positive image and then sample negative image from the disjoint seen classes. We then obtain triplets for all the training images from both the seen and unseen classes. The self-supervised triplet loss for the subspaces of unseen images is then defined as

$$\mathcal{L}_{SelfTri} = \sum_{i|y_i \in \mathcal{Y}^U} [D(\bar{\mathbf{S}}_i, \bar{\mathbf{S}}'_i) - D(\bar{\mathbf{S}}_i, \mathbf{S}_{neg}) + \alpha]_+ \tag{14}$$

where $\bar{\mathbf{S}}'_i$ is random augmentation of $\bar{\mathbf{S}}_i$.

For transductive ZSL, our model is trained with the overall loss using

$$\mathcal{L}_{TZSL} = \mathcal{L}_{CLS} + \beta_1 \mathcal{L}_{Tri} + \beta_2 \mathcal{L}_{MSE} + \mathcal{L}_{SelfTri} \tag{15}$$

when the input anchor image is from one of the unseen classes, since there is no ground-truth label, only $\mathcal{L}_{SelfTri}$ is used for training, and the remaining terms are set equal to 0.

After training the entire model, the inference phases of ZSL and GZSL are identical to that of inductive ZSL. With the unseen images exploited in training by our method, we can learn discriminative attribute subspace representations for unseen images, and this improves the performance of both ZSL and GZSL.

5. Experiments

For a fair comparison, our AS-ZSL was evaluated on three ZSL datasets, *i.e.*, CUB-200-2011 (CUB) [38], Animals with Attributes 2 (AwA2) [39], and SUN attribute (SUN) [40]. We followed the most widely used Proposed Split (PS) [39] in ZSL to divide the seen and unseen categories. Details of used datasets are listed in Table 1.

For the conventional ZSL setting, we adopt the average per-class Top-1 (**T1**) accuracy of unseen classes as evaluation metric. For GZSL setting, since the test set contains both seen and unseen categories, we calculate the Top-1 accuracy of seen classes $\mathbf{Acc_s}$ and unseen classes $\mathbf{Acc_u}$, respectively. Furthermore, the harmonic mean $\mathbf{H} = (2 \times \mathbf{Acc_s} \times \mathbf{Acc_u})/(\mathbf{Acc_s} + \mathbf{Acc_u})$ [39] is utilized to evaluate the comprehensive performance of GZSL.

5.1. Implementation Details

The Image Encoder of AS-ZSL is ResNet-101 which is pre-trained on ImageNet. We train the entire model in an end-to-end manner with the SGD optimizer. The momentum is set to 0.9, the weight decay is set to 10^{-5}, and the learning rate is set to 10^{-3}. The hyperparameters β_1 and β_2 are set to 0.5 and 1.0, respectively. The margin α is set to 0.2 for CUB and SUN datasets, and 0.3 for AwA2. The calibration factor γ is set to 0.9 for CUB and SUN, and 4.0 for AwA2. We use an episode-based training

Table 1: Attributes are the number of defined attributes in the dataset. $|\mathcal{Y}^S|$ and $|\mathcal{Y}^U|$ denote the category number of seen classes and unseen classes, respectively. TR, VAL, TE mean training set, validation set, and test set, respectively.

Datasets	Attributes	$\lvert\mathcal{Y}^S\rvert$	$\lvert\mathcal{Y}^U\rvert$	TR	VAL	TE
CUB	312	150	50	7057	1764	2967
AwA2	85	40	10	23527	5882	7913
SUN	102	645	72	10320	2580	1440

method [11] that samples M categories and N images for each category in a mini-batch. Each epoch contains 300 batches, and 20 epochs are trained. M and N are set to 16 and 2 consistently for all three datasets.

5.2. Comparison with the State-of-the-Art

We compared our AS-ZSL with abundant recent state-of-the-art ZSL methods. These include non end-to-end methods such as SP-AEN [41], PSR [42], TCN [43], IIR [44], DAZLE [45], E-PGN [46], and generative model based methods cycle-CLSWGAN [20], f-CLSWGAN [4], CADA-VAE [21], IZF [22], IB-ZSL [5], SRSA [3], DAGAN [18], and end-to-end methods LFGAA [47], AREN [23], APN [10], GEM-ZSL [11], MSDN [8], and HRT [9].

Both the conventional ZSL and GZSL results are reported in Table 2. The methods in the top block of the table are non end-to-end and the middle block shows generative model based methods. The bottom block gives end-to-end methods. From the table, our AS-ZSL can outperform all the compared methods on the CUB and AwA2 datasets for the harmonic mean accuracy. Additionally, the conventional ZSL result of AS-ZSL on CUB is also the best. This is because CUB is a challenging fine-grained bird image dataset which has stronger requirement on local discriminative attributes learning. These results validate that the proposed attribute composition learning in AS-ZSL has a more substantial ability in essential local feature learning. On the SUN dataset, generative model based methods achieve significant advantage. The performance of our method is not so satisfactory. Since there are more than 700 categories in SUN dataset,

Table 2: ZSL and GZSL results (%) of compared methods on different datasets. The top 6 methods are non end-to-end, the middle 7 are generative model-based methods, and the bottom 7 methods are end-to-end.

Methods	CUB				AwA2				SUN			
	ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
	T1	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**	**T1**	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**	**T1**	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**
SP-AEN(CVPR'18) [41]	55.4	34.7	70.6	46.6	-	-	-	-	59.2	24.9	38.6	30.3
PSR(CVPR'18) [42]	56.0	24.6	54.3	33.9	63.8	20.7	73.8	32.3	61.4	20.8	37.2	26.7
TCN(ICCV'19) [43]	59.5	52.6	52.0	52.3	71.2	61.2	65.8	63.4	61.5	31.2	37.3	34.0
IIR(ICCV'19) [44]	63.8	55.8	52.3	53.0	67.9	48.5	83.2	61.3	63.5	47.9	30.4	36.8
DAZLE(CVPR'20) [45]	65.9	56.7	59.6	58.1	-	60.3	75.7	67.1	-	52.3	24.3	33.2
E-PGN(CVPR'20) [46]	72.4	52.0	61.1	56.2	73.4	52.6	83.5	64.6	-	-	-	-
cycle-CLSWGAN(ECCV'18) [20]	58.4	45.7	61.0	52.3	-	-	-	-	60.0	49.4	33.6	40.0
f-CLSWGAN(CVPR'18) [4]	57.3	43.7	57.7	49.7	-	-	-	-	60.8	42.6	36.6	39.4
CADA-VAE(CVPR'19) [21]	-	51.6	53.5	52.4	-	55.8	75.0	63.9	-	47.2	35.7	40.6
IZF(ECCV'20) [22]	67.1	52.7	68.0	59.4	**74.5**	60.6	77.5	68.0	**68.4**	52.7	57.0	**54.8**
IB-ZSL(ML'22) [5]	62.2	52.2	56.2	54.1	70.1	56.0	80.0	65.9	64.2	43.8	37.8	40.6
SRSA(PR'22) [3]	59.9	27.5	55.6	36.8	68.3	38.1	59.6	46.5	64.3	25.3	37.9	30.3
DAGAN(PR'22) [18]	62.4	49.0	59.5	53.7	-	-	-	-	66.5	49.9	38.8	43.3
LFGAA(ICCV'19) [47]	67.6	36.2	80.9	50.0	68.1	27.0	93.4	41.9	61.5	18.5	40.0	25.3
AREN(CVPR'19) [23]	71.8	63.2	69.0	66.0	67.9	54.7	79.1	64.7	60.6	40.3	32.3	35.9
APN(NeurIPS'20) [10]	72.0	65.3	69.3	67.2	68.4	56.5	78.0	65.5	61.6	41.9	34.0	37.6
GEM-ZSL(CVPR'21) [11]	77.8	64.8	77.1	70.4	67.3	64.8	77.5	70.6	62.8	38.1	35.7	36.9
MSDN(CVPR'22) [8]	76.1	68.7	67.5	68.1	70.1	62.0	74.5	67.7	65.8	52.2	34.2	41.3
HRT(PR'23) [9]	71.7	63.5	62.1	62.8	67.3	78.7	58.9	67.4	63.9	26.9	53.2	35.7
AS-ZSL(Ours)	**78.5**	65.8	78.2	**71.5**	68.9	66.5	78.3	**71.9**	62.2	39.5	37.2	38.3

generative model-based methods can accommodate more features for generalization to the unseen classes. Even though, our AS-ZSL can still outperform the other part or locality-based methods such as AREN, APN, GEM-ZSL, and HRT.

5.3. Further Analysis

Ablation study. To further verify the effectiveness of the proposed attribute subspace learning module. We conducted ablation experiments on all the three datasets. The results are shown in Table 3. The first row of the result is baseline model that contains the image encoder with a cross-entropy loss. The second and third rows are the performance of AS-ZSL when adding the mean square error loss $\mathcal{L}_{MSE}$ and triplet loss $\mathcal{L}_{Tri}$, respectively. The last row is results of the full model with all losses. The ablation experimental results verify that the proposed AS-ZSL significantly promotes the ZSL

Table 3: Ablation study of the proposed AS-ZSL on different datasets.

Methods	CUB				AwA2				SUN			
	ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
	T1	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**	**T1**	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**	**T1**	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**
Baseline	70.4	62.4	72.4	67.0	64.2	58.0	**84.0**	68.7	58.1	35.3	33.6	34.4
$+\mathcal{L}_{MSE}$	76.8	65.6	75.7	70.3	68.1	65.8	74.4	69.8	61.6	38.9	32.9	35.6
$+\mathcal{L}_{Tri}$	74.8	62.3	78.0	69.3	66.7	60.8	83.3	70.3	61.2	36.8	35.1	35.9
Full Model	**78.5**	**65.8**	**78.2**	**71.5**	**68.9**	**66.5**	78.3	**71.9**	**62.2**	**39.5**	**37.2**	**38.3**

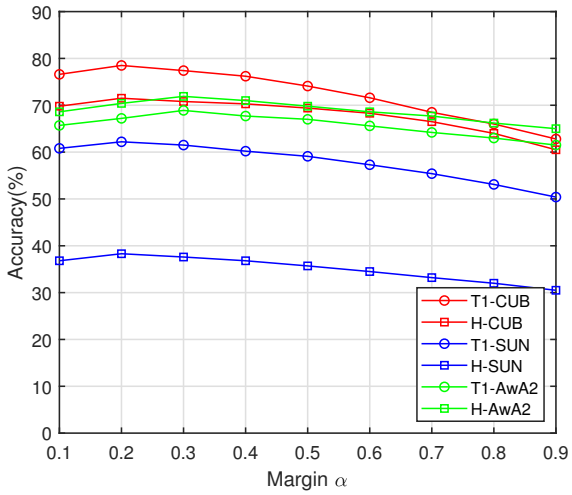


Figure 6: The effect of margin α in the triplet loss.

and GZSL performance. Specifically, AS-ZSL gains an improvement for ZSL by 8.1% (CUB), 4.7% (AwA2), 4.1% (SUN), and for the GZSL by 4.5% (CUB), 3.2% (AwA2), 3.9% (SUN). Thus, the proposed attribute subspace learning module can significantly promote the local discriminative feature learning which is beneficial to ZSL.

Effect of margin α in $\mathcal{L}_{Tri}$. Fig. 6 shows the results of **T1** and **H** when varying α from 0.1 to 0.9 under ZSL/GZSL settings for our method. It shows that when α is 0.2, AS-ZSL obtains best results on CUB and SUN, and when α is 0.3, AS-ZSL performs best on AwA2.

Training method analysis. In the training phase of AS-ZSL, to improve the gener-

Table 4: Results of different mini-batch sampling methods on GZSL (**H**). $\mathcal{R}$ denotes random sampling, $\mathcal{E}$ denotes different pairs of M and N.

Training Method	M-way	N-shot	CUB	AwA2	SUN
$\mathcal{R}$	mini-batch	random 64	62.8	67.2	34.7
$\mathcal{E}$	8	2	60.8	47.1	27.5
	8	3	64.8	62.5	28.6
	8	4	60.2	64.3	29.5
	12	2	62.9	68.8	33.4
	12	3	68.2	64.3	34.9
	12	4	67.6	65.5	34.7
	16	2	**71.5**	**71.9**	**38.3**
	16	3	70.1	70.2	37.5
	16	4	69.3	69.6	36.7

alization ability, we followed [11] to exploit an episode-based training strategy. Specifically, we sampled M categories and N images for each category in a batch. To analysis the influence of the different sampling strategies, we conducted experiments with different pairs of M and N from the ranges of $\{8, 12, 16\}$ and $\{2, 3, 4\}$, respectively. The comparison is shown in Table 4. It is clear that our AS-ZSL obtains the best performance when $M = 16$ and $N = 2$.

Visualization results. To further demonstrate the advantage of our AS-ZSL method, we used the t-SNE [48] to visualize the attribute response value $\hat{a}(\mathbf{x})$ and the global feature $h(\mathbf{x})$ used for the final classification. Fig. 7(a) and (b) show the distributions of the attribute response features and the global features of unseen classes on the three datasets. We can see that the feature distribution is very consistent for each class of CUB and AwA2. This verifies that the implicit attribute representation for each class is well learned by our subspace learning method. For the SUN dataset, the visualization results are not so satisfactory since there are 72 categories for the unseen images. The result also reflects the poorer performance of AS-ZSL on SUN.

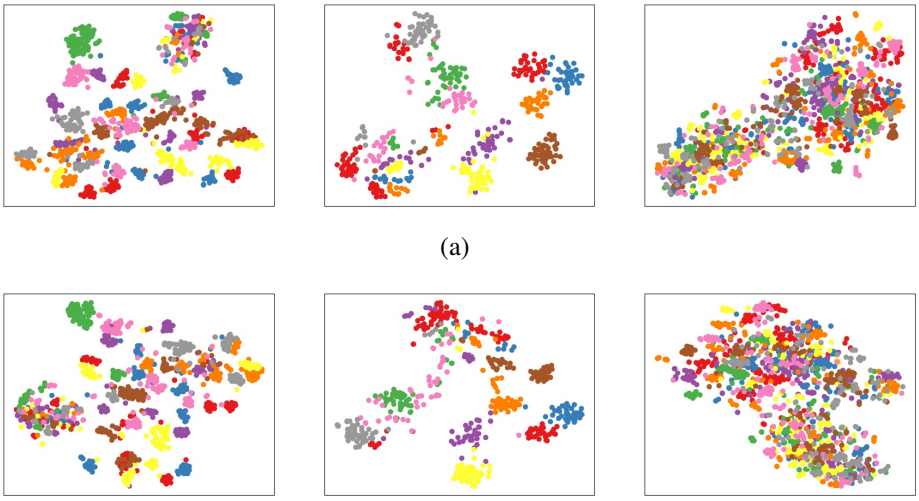


Figure 7: The feature distribution visualization results on the unseen classes of CUB, AwA2 and SUN (from left to right). (a) The attribute response value $\hat{a}(\mathbf{x})$. (b) The global feature $h(\mathbf{x})$.

5.4. Transductive ZSL

For the experiments of TZSL, the images of unseen classes are available for training. Thus, our AS-TZSL can achieve better performance compared with IZSL by alleviating the data imbalance problem between seen and unseen classes. To verify the effectiveness of AS-TZSL, we selected state-of-the-art TZSL methods for comparison, including ALE-trans [49], GFZSL [50], QFSL [51], GXE [27], GMN [52], f-VAEGAN [53], WDVSc [54], Zero-VAE-GAN [29], DeGAN [55], VMAN [30], IB-TZSL [5].

Table 5 shows the performance of GZSL on all the three datasets. AS-TZSL achieves the best results on both the CUB and AwA2 datasets. Especially on CUB, AS-TZSL outperform the second best method with a large margin of 11.4%. For the SUN dataset, AS-TZSL can also achieve competitive performance. The conventional ZSL results are reported in Table 6. Our AS-TZSL also obtains superior classification accuracy on the CUB and AwA2 datasets. These results validate the superiority of our proposed self-supervised training strategy.

Table 5: Results of the GZSL for TZSL on CUB, AwA2, and SUN.

Methods	CUB			AwA2			SUN		
	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**
ALE-trans(CVPR'15) [49]	23.5	45.1	30.9	12.6	73.0	21.5	19.9	22.6	21.2
GFZSL(ECML-PKDD'17) [50]	24.9	45.8	32.2	31.7	67.2	43.1	-	-	-
QFSL(CVPR'18) [51]	17.3	39.0	24.0	20.8	74.7	32.6	17.7	25.0	20.7
GXE(ICCV'19) [27]	57.0	68.7	62.3	80.2	90.0	84.8	45.4	**58.1**	**51.0**
GMN(CVPR'19) [52]	60.2	70.6	65.0	-	-	-	57.1	40.7	47.5
f-VAEGAN(CVPR'19) [53]	61.4	65.1	63.2	84.8	88.6	86.7	**60.6**	41.9	49.6
WDVSc(NeurIPS'19) [54]	43.3	**85.4**	57.5	76.4	88.1	81.8	-	-	-
Zero-VAE-GAN(TIP'20) [29]	64.1	57.9	60.8	70.2	87.0	77.6	53.1	35.8	42.8
DeGAN(WACV'21) [55]	59.1	68.4	63.4	-	-	-	57.2	44.3	49.9
VMAN(TIP'21) [30]	65.6	54.9	59.8	72.9	84.9	78.4	59.3	32.0	41.6
IB-TZSL(ML'22) [5]	63.5	66.5	65.9	82.7	89.2	85.8	57.5	44.6	50.2
AS-TZSL	**73.2**	76.5	**74.8**	**85.2**	**90.2**	**87.6**	55.6	42.5	48.2

5.5. Further Analyses for Transductive Setting

Table 6: Results of conventional ZSL for transductive setting on CUB, AwA2, and SUN.

Methods	CUB	AwA2	SUN
ALE-trans(CVPR'15) [49]	54.5	70.7	55.7
GFZSL(ECML-PKDD'17) [50]	49.3	78.6	64.0
QFSL(CVPR'18) [51]	72.1	79.7	58.3
GXE(ICCV'19) [27]	61.3	83.2	63.5
GMN(CVPR'19) [52]	64.6	-	64.3
f-VAEGAN(CVPR'19) [53]	71.7	89.8	**70.1**
WDVSc(NeurIPS'19) [54]	73.4	87.3	63.4
Zero-VAE-GAN(TIP'20) [29]	68.9	85.4	66.8
VMAN(TIP'21) [30]	72.9	89.3	69.3
IB-TZSL(ML'22) [5]	73.5	88.1	67.6
AS-TZSL	**79.3**	**90.1**	64.5

Different data augmentation methods. Under the transductive ZSL setting, we generated triplets for anchor images from unseen classes by sampling negative from seen classes and utilizing the data augmentation method on itself to obtain positive.

Table 7: Results (%) of ZSL and GZSL under transductive ZSL setting with different data augmentation methods in the triplets generation.

Methods	CUB				AwA2				SUN			
	ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
	T1	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**	**T1**	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**	**T1**	$\mathbf{Acc_u}$	$\mathbf{Acc_s}$	**H**
Flip	77.3	71.9	75.8	73.8	88.6	83.3	89.2	86.1	63.0	53.8	41.2	46.7
Crop	78.1	72.5	76.2	74.3	89.2	84.1	89.4	86.7	63.3	54.3	41.7	47.2
Brightness	77.5	72.2	76.1	74.1	88.9	83.7	89.3	86.4	63.2	54.1	41.7	47.1
Random augmentation	**79.3**	**73.2**	**76.5**	**74.8**	**90.1**	**85.2**	**90.2**	**87.6**	**64.5**	**55.6**	**42.5**	**48.2**

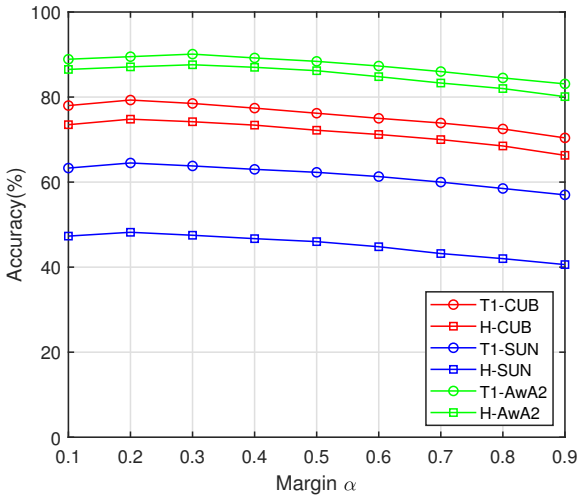


Figure 8: The effect of margin α in the self-supervised triplet loss for transductive ZSL.

Here, we conducted an experiment to demonstrate the effect of different data augmentation methods. Table 7 shows the results of three widely used data augmentation methods, *i.e.*, flip, crop, and brightness. The random augmentation randomly adopts one of the three methods for each unseen image. It s clear that AS-TZSL achieves the best performance with random augmentation. The reason for this may be that random augmentation can bring better diversity for images from different classes.

Effect of margin α in $\mathcal{L}_{SelfTri}$. Fig. 8 shows the results of **T1** and **H** when varying

α from 0.1 to 0.9 under ZSL/GZSL for transductive ZSL. From the figure, we reach the same conclusion as with the inductive setting, *i.e.*. that AS-TZSL obtains the best performance on CUB and SUN when α is 0.2, and 0.3 for AwA2.

6. Conclusion

In this paper, we have proposed a novel attribute subspace learning method for the zero-shot recognition task. It benefits from the property of subspace learning that can capture the underlying structure of diversified samples from similar classes. Our AS-ZSL method can learn the attribute composition with only class-level supervision. Compared with locality-based ZSL methods, learning the attribute composition provides both discriminative and robust features, capturing the diversity of spatial locality and the visual appearance of attributes in different images. To our knowledge, AS-ZSL is the first method to investigate the subspace representation for attribute composition learning. Furthermore, we naturally extend our method to cope with the transductive ZSL problem using a self-supervised triplet loss. The designed triplet sampling strategy can also be used for alternative transductive learning methods to solve the problem of unseen classes in training. Extensive experiments under different ZSL settings validate the effectiveness of our method, especially compared with the state-of-the-art locality-based methods APN and GEM-ZSL. Our future work will explore the potential for further investigation of attribute composition learning using alternative manifold learning methods. In addition, the attribute subspace learning method could also be applied to few-shot learning and fine-grained recognition tasks to facilitate discriminative feature learning.

Acknowledgement

This work was supported by the National Natural Science Foundation of China project no. 62276016, and supported by "the Fundamental Research Funds for the Central Universities (WUT:233110004)".

References

[1] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: Computer Vision and Pattern Recognition, 2009, pp. 951–958.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for image classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (7) (2015) 1425–1438.

[3] Y. Liu, X. Gao, J. Han, L. Liu, L. Shao, Zero-shot learning via a specific rank-controlled semantic autoencoder, Pattern Recognition 122 (2022) 108237.

[4] Y. Xian, T. Lorenz, B. Schiele, Z. Akata, Feature generating networks for zero-shot learning, in: Computer Vision and Pattern Recognition, 2018, pp. 5542–5551.

[5] L. Zhou, Y. Liu, P. Zhang, X. Bai, L. Gu, J. Zhou, Y. Yao, T. Harada, J. Zheng, E. Hancock, Information bottleneck and selective noise supervision for zero-shot learning, Machine Learning (2022) 1–23.

[6] Y. Liu, L. Zhou, P. Zhang, X. Bai, L. Gu, X. Yu, J. Zhou, E. R. Hancock, Where to focus: Investigating hierarchical attention relationship for fine-grained visual classification, in: European Conference on Computer Vision, Springer, 2022, pp. 57–73.

[7] G.-S. Xie, L. Liu, F. Zhu, F. Zhao, Z. Zhang, Y. Yao, J. Qin, L. Shao, Region graph embedding network for zero-shot learning, in: European Conference on Computer Vision, 2020, pp. 562–580.

[8] S. Chen, Z. Hong, G.-S. Xie, W. Yang, Q. Peng, K. Wang, J. Zhao, X. You, Msdn: Mutually semantic distillation network for zero-shot learning, in: Computer Vision and Pattern Recognition, 2022, pp. 7612–7621.

[9] D. Cheng, G. Wang, B. Wang, Q. Zhang, J. Han, D. Zhang, Hybrid routing transformer for zero-shot learning, Pattern Recognition 137 (2023) 109270.

[10] W. Xu, Y. Xian, J. Wang, B. Schiele, Z. Akata, Attribute prototype network for zero-shot learning, in: Neural Information Processing Systems, 2020, pp. 21969–21980.

[11] Y. Liu, L. Zhou, X. Bai, Y. Huang, L. Gu, J. Zhou, T. Harada, Goal-oriented gaze estimation for zero-shot learning, in: Computer Vision and Pattern Recognition, 2021, pp. 3794–3803.

[12] L. Zhou, X. Bai, X. Liu, J. Zhou, Binary coding by matrix classifier for efficient subspace retrieval, in: International Conference on Multimedia Retrieval, 2018, pp. 82–90.

[13] L. Zhou, X. Zhang, J. Wang, X. Bai, L. Tong, L. Zhang, J. Zhou, E. Hancock, Subspace structure regularized nonnegative matrix factorization for hyperspectral unmixing, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13 (2020) 4257–4270.

[14] M. Jiang, S. Zhou, C. Li, Y. Lei, Jsl3d: Joint subspace learning with implicit structure supervision for 3d pose estimation, Pattern Recognition 132 (2022) 108965.

[15] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, Devise: A deep visual-semantic embedding model, in: Neural Information Processing Systems, 2013, pp. 2121–2129.

[16] M. Radovanovic, A. Nanopoulos, M. Ivanovic, Hubs in space: Popular nearest neighbors in high-dimensional data, Journal of Machine Learning Research 11 (sept) (2010) 2487–2531.

[17] Q. Wang, K. Chen, Zero-shot visual recognition via bidirectional latent embedding, International Journal of Computer Vision 124 (3) (2017) 356–383.

[18] H. Kim, J. Lee, H. Byun, Discriminative deep attributes for generalized zero-shot learning, Pattern Recognition 124 (2022) 108435.

[19] X. Kong, Z. Gao, X. Li, M. Hong, J. Liu, C. Wang, Y. Xie, Y. Qu, Encompactness: Self-distillation embedding & contrastive generation for general-

ized zero-shot learning, in: Computer Vision and Pattern Recognition, 2022, pp. 9306–9315.

[20] R. Felix, V. B. Kumar, I. Reid, G. Carneiro, Multi-modal cycle-consistent generalized zero-shot learning, in: European Conference on Computer Vision, 2018, pp. 21–37.

[21] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, Z. Akata, Generalized zero-and few-shot learning via aligned variational autoencoders, in: Computer Vision and Pattern Recognition, 2019, pp. 8247–8255.

[22] Y. Shen, J. Qin, L. Huang, L. Liu, F. Zhu, L. Shao, Invertible zero-shot recognition flows, in: European Conference on Computer Vision, 2020, pp. 614–631.

[23] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, L. Shao, Attentive region embedding network for zero-shot learning, in: Computer Vision and Pattern Recognition, 2019, pp. 9384–9393.

[24] H. Zhang, H. Bai, Y. Long, L. Liu, L. Shao, A plug-in attribute correction module for generalized zero-shot learning, Pattern Recognition 112 (2021) 107767.

[25] T. Sylvain, L. Petrini, D. Hjelm, Locality and compositionality in zero-shot learning, in: International Conference on Learning Representations, 2020.

[26] A. Paul, N. C. Krishnan, P. Munjal, Semantically aligned bias reducing zero shot learning, in: Computer Vision and Pattern Recognition, 2019, pp. 7056–7065.

[27] K. Li, M. R. Min, Y. Fu, Rethinking zero-shot learning: A conditional visual classification perspective, in: International Conference on Computer Vision, 2019, pp. 3583–3592.

[28] J. Wu, T. Zhang, Z.-J. Zha, J. Luo, Y. Zhang, F. Wu, Self-supervised domain-aware generative network for generalized zero-shot learning, in: Computer Vision and Pattern Recognition, 2020, pp. 12767–12776.

[29] R. Gao, X. Hou, J. Qin, J. Chen, L. Liu, F. Zhu, Z. Zhang, L. Shao, Zero-vae-gan: Generating unseen features for generalized and transductive zero-shot learning, IEEE Transactions on Image Processing 29 (2020) 3665–3680.

[30] G.-S. Xie, X.-Y. Zhang, Y. Yao, Z. Zhang, F. Zhao, L. Shao, Vman: A virtual mainstay alignment network for transductive zero-shot learning, IEEE Transactions on Image Processing 30 (2021) 4316–4329.

[31] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2008) 210–227.

[32] L. Zhou, X. Bai, X. Liu, J. Zhou, E. R. Hancock, Learning binary code for fast nearest subspace search, Pattern Recognition 98 (2020) 107040.

[33] L. Zhou, B. Xiao, X. Liu, J. Zhou, E. R. Hancock, et al., Latent distribution preserving deep subspace clustering, in: International Joint Conference on Artificial Intelligence, 2019, pp. 4440–4446.

[34] A. Cherian, S. Sra, S. Gould, R. Hartley, Non-linear temporal subspace representations for activity recognition, in: Computer Vision and Pattern Recognition, 2018, pp. 2197–2206.

[35] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (10) (2015) 2010–2023.

[36] C. Simon, P. Koniusz, R. Nock, M. Harandi, Adaptive subspaces for few-shot learning, in: Computer Vision and Pattern Recognition, 2020, pp. 4136–4145.

[37] C. F. Van Loan, G. Golub, Matrix computations (johns hopkins studies in mathematical sciences).

[38] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona, Caltech-ucsd birds 200.

[39] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata, Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly, IEEE Transactions on Pattern Analysis and Machine Intelligence (2018) 2251–2265.

[40] G. Patterson, J. Hays, Sun attribute database: Discovering, annotating, and recognizing scene attributes, in: Computer Vision and Pattern Recognition, 2012, pp. 2751–2758.

[41] L. Chen, H. Zhang, J. Xiao, W. Liu, S.-F. Chang, Zero-shot visual recognition using semantics-preserving adversarial embedding networks, in: Computer Vision and Pattern Recognition, 2018, pp. 1043–1052.

[42] Y. Annadani, S. Biswas, Preserving semantic relations for zero-shot learning, in: Computer Vision and Pattern Recognition, 2018, pp. 7603–7612.

[43] H. Jiang, R. Wang, S. Shan, X. Chen, Transferable contrastive network for generalized zero-shot learning, in: International Conference on Computer Vision, 2019, pp. 9765–9774.

[44] Y. L. Cacheux, H. L. Borgne, M. Crucianu, Modeling inter and intra-class relations in the triplet loss for zero-shot learning, in: International Conference on Computer Vision, 2019, pp. 10333–10342.

[45] D. Huynh, E. Elhamifar, Fine-grained generalized zero-shot learning via dense attribute-based attention, in: Computer Vision and Pattern Recognition, 2020, pp. 4483–4493.

[46] Y. Yu, Z. Ji, J. Han, Z. Zhang, Episode-based prototype generating network for zero-shot learning, in: Computer Vision and Pattern Recognition, 2020, pp. 14035–14044.

[47] Y. Liu, J. Guo, D. Cai, X. He, Attribute attention for semantic disambiguation in zero-shot learning, in: International Conference on Computer Vision, 2019, pp. 6698–6707.

[48] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (Nov) (2008) 2579–2605.

[49] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: Computer Vision and Pattern Recognition, 2015, pp. 2927–2936.

[50] V. K. Verma, P. Rai, A simple exponential family framework for zero-shot learning, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2017, pp. 792–808.

[51] J. Song, C. Shen, Y. Yang, Y. Liu, M. Song, Transductive unbiased embedding for zero-shot learning, in: Computer Vision and Pattern Recognition, 2018, pp. 1024–1033.

[52] M. B. Sariyildiz, R. G. Cinbis, Gradient matching generative networks for zero-shot learning, in: Computer Vision and Pattern Recognition, 2019, pp. 2168–2178.

[53] Y. Xian, S. Sharma, B. Schiele, Z. Akata, f-vaegan-d2: A feature generating framework for any-shot learning, in: Computer Vision and Pattern Recognition, 2019, pp. 10275–10284.

[54] Z. Wan, D. Chen, Y. Li, X. Yan, J. Zhang, Y. Yu, J. Liao, Transductive zero-shot learning with visual structure constraint, in: Neural Information Processing Systems, 2019, pp. 9972–9982.

[55] F. Marmoreo, J. Cavazza, V. Murino, Transductive zero-shot learning by decoupled feature generation, in: Winter Conference on Applications of Computer Vision, 2021, pp. 3109–3118.