# Learning Class Prototypes via Structure Alignment for Zero-Shot Recognition

Huajie Jiang[1,2,3,4][0000−0002−1158−6321], Ruiping Wang[1,4][0000−0003−1830−2595], Shiguang Shan[1,4][0000−0002−8348−392X], and Xilin Chen[1,4][0000−0003−3024−4404]

[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2]Shanghai Institute of Microsystem and Information Technology, CAS, Shanghai, 200050, China
[3]ShanghaiTech University, Shanghai, 200031, China
[4]University of Chinese Academy of Sciences, Beijing, 100049, China
huajie.jiang@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

**Abstract.** Zero-shot learning (**ZSL**) aims to recognize objects of novel classes without any training samples of specific classes, which is achieved by exploiting the semantic information and auxiliary datasets. Recently most **ZSL** approaches focus on learning visual-semantic embeddings to transfer knowledge from the auxiliary datasets to the novel classes. However, few works study whether the semantic information is discriminative or not for the recognition task. To tackle such problem, we propose a coupled dictionary learning approach to align the visual-semantic structures using the class prototypes, where the discriminative information lying in the visual space is utilized to improve the less discriminative semantic space. Then, zero-shot recognition can be performed in different spaces by the simple nearest neighbor approach using the learned class prototypes. Extensive experiments on four benchmark datasets show the effectiveness of the proposed approach.

**Keywords:** Zero-Shot Learning, Visual-Semantic Structures, Coupled Dictionary Learning, Class Prototypes

## 1 Introduction

Object recognition has made tremendous progress in recent years. With the emergence of large-scale image database [28], deep learning approaches [17, 31, 29, 13] show their great power to recognize objects. However, such supervised learning approaches require large numbers of images to train robust recognition models and can only recognize a fixed number of categories, which limits their flexibility. It is well known that collecting large numbers of images is difficult. On one hand, the numbers of images often follow a long-tailed distribution [41] and it is hard to collect images for some rare categories. On the other hand, some fine-grained annotations require expert knowledge [33], which increases the difficulty of the annotation task. All these challenges motivate the rise of zero-shot learning, where no labeled examples are needed to recognize one category.

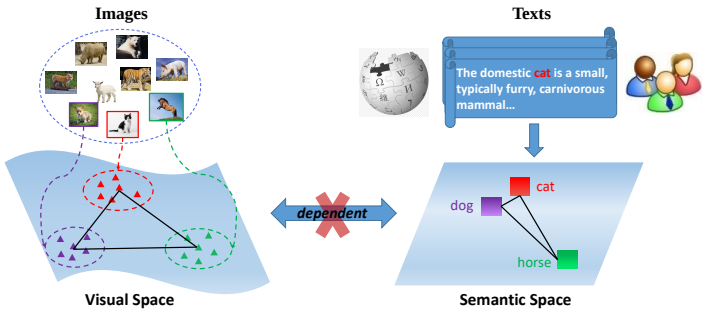


Fig. 1: Illustration diagram that shows the inconsistency of visual feature space and semantic space. The semantic information is manually defined or automatically extracted, which is independent of visual samples. The black lines in the two spaces show the similarities between different classes.

Zero-shot learning aims at recognizing objects that have not been seen in the training stage, where auxiliary datasets and semantic information are needed to perform such tasks. It is mainly inspired by the human's behavior to recognize new objects. For example, children have no problem recognizing *zebra* if they are told that *zebra* looks like a *horse* (auxiliary datasets) but has *stripes* (semantic information), even though they have never seen *zebra* before. Current **ZSL** approaches generally involve three steps. First, choose a semantic space to build up the relations between seen (auxiliary dataset) and unseen (test) classes. Recently the most popular semantic information includes attributes [19, 9] that are manually defined and wordvectors [10, 2] that are automatically extracted from the auxiliary text corpus. Second, learn general visual-semantic embeddings from the auxiliary dataset, where the images and class semantics could be projected into a common space [1, 5]. Third, perform the recognition task in the common space by different metric learning approaches.

Traditional **ZSL** approaches usually use fixed semantic information and pay much attention to learning more robust visual-semantic embeddings [19, 10, 1, 15, 24, 38]. However, most of these approaches ignore the fact that the semantic information, whether human-defined or automatically extracted, is incomplete and may be not discriminative enough to classify different classes because the descriptions about classes are limited. As is shown in Figure 1, some classes may locate quite close to each other in the semantic space due to the incomplete descriptions, *i.e. cat* and *dog*, thus it may be less effective to perform recognition task in this space. Since images are real reflections of different categories, they may contain more discriminative information that could not be described. Moreover, the semantic information is obtained independently from visual samples so the class structures between the visual space and semantic space are not consistent. In such cases, the visual-semantic embeddings would be too complicated to learn. Even if the embeddings are properly learned, they have large probabilities to overfit the seen classes and have less expansibility to the unseen classes.

In order to tackle such problems, we propose to learn the class prototypes by aligning the visual-semantic structures. The novelty of our framework lies in three aspects. First, different from traditional approaches which learn image embeddings, we perform the structure alignment on the class prototypes, which are automatically learned, to conduct the recognition task. Second, a coupled dictionary learning framework is proposed to align the class structures between visual space and semantic space, where the discriminative property lying in the visual space and the extensive property existing in the semantic space are merged in an aligned space. Third, semantic information of unseen classes is utilized for domain adaptation, which increases the expansibility of our model to the unseen classes. In order to demonstrate the effectiveness of the proposed approach, we perform experiments on four popular datasets for zero-shot recognition, where excellent results are achieved.

## 2 Related Work

In this section, we review related works on zero-shot learning in three aspects, *i.e.* semantic information, visual-semantic embeddings, zero-shot recognition.

### 2.1 Semantic Information

Semantic information plays an important role in zero-shot learning. It builds up the relations between seen and unseen classes, thus making it possible for zero-shot recognition. Recently, the most popular semantic information includes attributes [19, 9, 1, 3, 14] and wordvectors [2, 7, 22]. Attributes are general descriptions of objects which can be shared among different classes. For example, *furry* can be shared among different animals. Thus it is possible to learn such attributes by some auxiliary classes and apply them to the novel classes for recognition. Wordvectors are automatically extracted from large numbers of text corpus, where the distances between different wordvectors show the relations between different classes, thus they are also capable of building up the relations between seen and unseen classes.

Since the knowledge that could be collected is limited, the semantic information obtained in general purpose is usually less discriminative to classify different classes in specific domains. To tackle such problem, we propose to utilize the discriminative information lying in the visual space to improve the semantic space.

### 2.2 Visual-Semantic Embeddings

Visual-semantic embedding is the key to zero-shot learning and most existing **ZSL** approaches focus on learning more robust visual-semantic embeddings. In the early stage, [19, 9] propose to use attribute classifiers to perform **ZSL** task. Such methods learn each attribute classifier independently, which is not applicable to large-scale datasets with lots of attributes. In order to tackle such problems, label embedding approaches emerge [1, 2], where all attributes are considered

as a whole for a class and label embedding functions are learned to maximize the compatibility of images with corresponding class semantics. To improve the performance of such embedding models, [35] proposes latent embedding models, where multiple linear embeddings are learned to approximate non-linear embeddings. Furthermore, [10, 30, 34, 26, 38, 22] exploit deep neural networks to learn more robust visual-semantic transformations.

Although some works pay attention to learning more complicated embedding functions, some other works deal with the visual-semantic transformation problem from different views. [23] forms the semantic information of unseen samples by a convex combination of seen-class semantics. [39, 40] utilize the class similarities and [14] proposes discriminative latent attributes to form more effective embedding space. [4] synthesizes the unseen-class classifiers by sharing the structures between the semantic space and the visual space. [5, 20] predicts the visual exemplars by learning embedding functions from the semantic space to the visual space. [3] exploits metric learning techniques, where relative distance is utilized, to improve the embedding models. [27] views the image classifier as a function of corresponding class semantic and uses additional regularizer to learn the embedding functions. [16] utilizes the auto-encoder framework to learn the visual-semantic embeddings. [8] uses low rank constraints to learn semantic dictionaries and [37] proposes a matrix tri-factorization approach with manifold regularizations. To tackle the embedding domain shift problem, [15, 11] use the transfer learning techniques to extend **ZSL** into transductive settings, where the unseen-class samples are also utilized in the training process.

Different from such existing approaches which learn image embeddings or synthesize image classifiers, we propose to learn the class prototypes by jointly aligning the class structures between the visual space and the semantic space.

### 2.3 Zero-Shot Recognition

The most widely used approaches for zero-shot recognition are probability models [19] and nearest neighbour classifiers [1, 39, 14]. To make use of the rich intrinsic structures on the semantic manifold, [12] proposes semantic manifold distance to recognize the unseen class samples and [4] directly synthesizes the image classifiers of unseen classes in the visual space by sharing the structures between the semantic space and the visual space. Considering more real conditions, [6] expands the traditional **ZSL** problem to the generalized **ZSL** problem, where the seen classes are also considered in the test procedure. Recently, [36] proposes more reasonable data splits for different datasets and evaluates the performance of different approaches under such experiment settings.

## 3 Approaches

The general idea of the proposed approach is to learn the class prototypes by sharing the structures between the visual space and the semantic space. However, the structures between these two spaces may be inconsistent, since the semantic

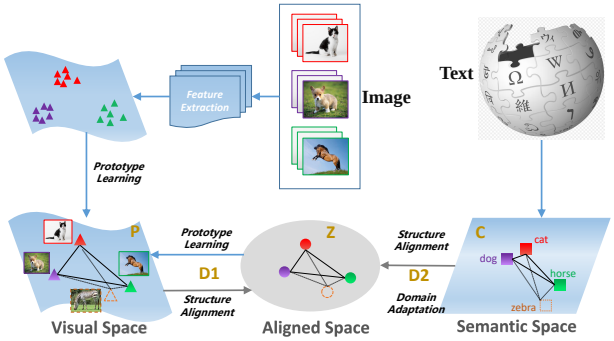


Fig. 2: Coupled dictionary learning framework to align the visual-semantic structure. The solid shapes represent the seen-class prototypes and the dotted shapes denote the prototypes of unseen classes. Black lines show the relationships between different classes. The brown characters are corresponding to the formulation of equations.

information is obtained independently of the visual examples. In order to tackle such problem, we propose a coupled dictionary learning (**CDL**) framework to simultaneously align the visual-semantic structures. Thus the discriminative information in the visual space and the relations in the semantic space can be shared to benefit each other. Figure 2 shows the framework of our approach. There are three key submodules of the proposed framework: prototype learning, structure alignment, and domain adaptation.

### 3.1 Problem Formulation

Assume a labeled training dataset contains $K$ seen classes with $n_s$ labeled samples $\mathcal{S} = \{(x_i, y_i)|x_i \in \mathcal{X}, y_i \in \mathcal{Y}^s\}_{i=1}^{n_s}$, where $x_i \in \mathbb{R}^d$ represents the image feature and $y_i$ denotes the class label in $\mathcal{Y}^s = \{s_1, ..., s_K\}$. In addition, a disjoint class label set $\mathcal{Y}^u = \{u_1, ..., u_L\}$, which consists $L$ unseen classes, is provided, *i.e.* $\mathcal{Y}^u \bigcap \mathcal{Y}^s = \emptyset$, but the corresponding images are missing. Given the class semantics $\mathcal{C} = \{\mathcal{C}^s \bigcup \mathcal{C}^u\}$, the goal of **ZSL** is to learn image classifiers $f_{zsl} : \mathcal{X} \rightarrow \mathcal{Y}^u$.

### 3.2 Framework

As is shown in Figure 2, our framework contains three submodules: prototype learning, structure alignment and domain adaptation.

**Prototype Learning** The structure alignment approach proposed by our framework is performed on the class prototypes. In order to align the class structures between the visual space and the semantic space, we must first obtain the class prototypes in both spaces. In the semantic space, we denote the class prototypes of seen/unseen classes as $C_s \in \mathbb{R}^{m \times K}/C_u \in \mathbb{R}^{m \times L}$, where $m$ is the dimension of the semantic space. Here, $C_s/C_u$ can be directly set as $\mathcal{C}^s/\mathcal{C}^u$.

However, in the visual space, only the seen-class samples $X_s \in \mathbb{R}^{d \times n_s}$ and their corresponding labels $Y_s$ are provided, so we should first learn the class prototypes $P_s \in \mathbb{R}^{d \times K}$ in the visual space, where $d$ is the dimension of the visual space. The basic idea for prototype learning is that samples should locate near their corresponding class prototypes in the visual space, so the loss function can be formulated as:

$$\mathcal{L}_p = \min_{P_s} \|X_s - P_s H\|_F^2 , \tag{1}$$

where each column in $H \in \mathbb{R}^{K \times n_s}$ is a one-hot vector indicating the class label of corresponding image.

**Structure Alignment** Due to the fact that the semantic information of classes is defined or extracted independently of the images, directly sharing the structures in the semantic space to form the prototypes of unseen classes in the visual space is not a good choice, where structure alignment should be performed first. Therefore, we propose a coupled dictionary learning framework to align the visual-semantic structures. The basic idea for our structure alignment approach is to find some bases in each space to represent each class and enforce the new representation to be the same in the two spaces, thus the structures can be aligned. The loss function is formulated as:

$$\begin{aligned} \mathcal{L}_s = & \min_{P_s, D_1, D_2, Z_s} \|P_s - D_1 Z_s\|_F^2 + \lambda \|C_s - D_2 Z_s\|_F^2 , \\ & s.t. \quad ||\boldsymbol{d}_1^i||_2^2 \leq 1, \quad ||\boldsymbol{d}_2^i||_2^2 \leq 1, \forall i. \end{aligned} \tag{2}$$

where $P_s$ and $C_s$ are the prototypes of seen classes in the visual and semantic space respectively. $D_1 \in \mathbb{R}^{d \times n_b}$ and $D_2 \in \mathbb{R}^{m \times n_b}$ are the bases in corresponding spaces, where $d, m$ are the dimensions of visual space and semantic space respectively and $n_b$ is the number of bases. $Z_s \in \mathbb{R}^{n_b \times K}$ is the common new representation of seen classes, and it just plays the key role to align the two spaces. $\lambda$ is a parameter controlling the relative importance of the visual space and semantic space. $\boldsymbol{d}_1^i$ denotes the $i$-th column of $D_1$ and $\boldsymbol{d}_2^i$ is the $i$-th column of $D_2$. By exploring new representation bases in each space to reformulate each class, we obtain the same class representations for the visual and semantic spaces, thus the class structures in the two spaces will be consistent.

**Domain Adaptation** In the structure alignment process, only seen-class prototypes are utilized and this may cause the domain shift problem [11]. In other words, a general structure alignment approach learned on seen classes may not be appropriate for the unseen classes, since there are some differences between seen and unseen classes. To tackle such problem, we further propose a domain adaptation term, which automatically learns the unseen-class prototypes in the visual space and uses the unseen prototypes to assist the structure learning process. The loss function can be formulated as:

$$\begin{aligned} \mathcal{L}_u = & \min_{P_u, D_1, D_2, Z_u} \|P_u - D_1 Z_u\|_F^2 + \lambda \|C_u - D_2 Z_u\|_F^2 , \\ & s.t. \quad ||\boldsymbol{d}_1^i||_2^2 \leq 1, \quad ||\boldsymbol{d}_2^i||_2^2 \leq 1, \forall i. \end{aligned} \tag{3}$$

where $P_u \in \mathbb{R}^{d\times L}$ and $C_u \in \mathbb{R}^{m\times L}$ are the prototypes of unseen classes in the visual and semantic space respectively, and $Z_u \in \mathbb{R}^{n_b\times L}$ is the common new representation of unseen classes.

In a whole, our full objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_s + \alpha\mathcal{L}_u + \beta\mathcal{L}_p, \tag{4}$$

where $\alpha$ and $\beta$ are the parameters controlling the relative importance.

### 3.3 Optimization

The final loss function of the proposed framework can be formulated as:

$$\begin{aligned}\mathcal{L} = \min_{P_s,P_u,D_1,D_2,Z_s,Z_u} &(\|P_s - D_1Z_s\|_F^2 + \lambda\|C_s - D_2Z_s\|_F^2)+\\ &\alpha(\|P_u - D_1Z_u\|_F^2 + \lambda\|C_u - D_2Z_u\|_F^2) + \beta(\|X_s - P_sH\|_F^2),\\ &s.t. \quad ||\boldsymbol{d}_1^i||_2^2 \le 1, \quad ||\boldsymbol{d}_2^i||_2^2 \le 1, \forall i.\end{aligned} \tag{5}$$

It is obvious that Eq.5 is not convex for $P_s, P_u, D_1, D_2, Z_s$ and $Z_u$ simultaneously, but it is convex for each of them separately. We thus employ an alternating optimization method to solve the problem.

**Initialization** In our framework, we set the number of dictionary bases $n_b$ as the number of seen classes $K$ and enforces each column of $Z$ to be the similarities to all seen classes. First, we initialize $Z_u \in \mathbb{R}^{K\times L}$ as the similarities of unseen classes to the seen classes, *i.e.* cosine distances between unseen and seen class prototypes in the semantic space. Second, we get $D_2$ by the second term of Eq.3, which has closed-form solution. Third, we get $Z_s$ by the second term of Eq.2. Next, we initialize $P_s$ as the mean of samples in each class. Then, we get $D_1$ by the first term of Eq.2. In the end, we get $P_u$ by the first term in Eq.3. In this way, all the variables in our framework are initialized.

**Joint Optimization** After all variables in our framework are initialized separately, we jointly optimize them as follows:

(1) Fix $D_1, Z_s$ and update $P_s$. The subproblem can be formulated as:

$$\arg\min_{P_s} \|P_s - D_1Z_s\|_F^2 + \beta\|X_s - P_sH\|_F^2 \tag{6}$$

(2) Fix $P_s, D_1, D_2$ and update $Z_s$ by Eq.2.

(3) Fix $P_s, P_u, Z_s, Z_u$ and update $D_1$. The subproblem can be formulated as:

$$\arg\min_{D_1} \|P_s - D_1Z_s\|_F^2 + \alpha\|P_u - D_1Z_u\|_F^2 \quad s.t. \quad ||\boldsymbol{d}_1^i||_2^2 \le 1, \forall i. \tag{7}$$

(4) Fix $Z_s, Z_u$ and update $D_2$. The subproblem can be formulated as:

$$\arg\min_{D_1} \|C_s - D_2Z_s\|_F^2 + \alpha\|C_u - D_2Z_u\|_F^2 \quad s.t. \quad ||\boldsymbol{d}_2^i||_2^2 \le 1, \forall i. \tag{8}$$

(5) Fix $P_u, D_1, D_2$ and update $Z_u$ by Eq.3.

(6) Fix $D_1, Z_u$ and update $P_u$ by the first term of Eq.3.

In our experiments, we set the maximum iterations as 100 and the optimization always converges after tens of iterations, usually less than 50. [1]

### 3.4 Zero-Shot Recognition

In the proposed framework, we can obtain the prototypes of unseen classes in different spaces (*i.e.* visual space $P_u$, aligned space $Z_u$, semantic space $C_u$), where we can perform zero-shot recognition task using nearest neighbour approach.

**Recognition in the Visual Space**. In the test process, we can directly compute the similarities $Sim_v$ of test samples ($X_i$) to the unseen class prototypes ($P_u$), *i.e.* cosine distance, and classify the images to the classes corresponding to their most similar prototypes.

**Recognition in the Aligned Space**. To perform recognition task in this space, we must first obtain the representations of images in this space by

$$\arg\min_{Z_i} \|X_i - D_1 Z_i\|_F^2 + \gamma \|Z_i\|_F^2 \tag{9}$$

where $X_i$ represents the test images and $Z_i$ is the corresponding representation in the aligned space. Then we can obtain the similarities $Sim_a$ of test samples ($Z_i$) to the unseen-class prototypes ($Z_u$) and use the same recognition approach as that in the visual space.

**Recognition in the Semantic Space**. First, we should get the semantic representations of images by $C_i = D_2 Z_i$. Then the similarities $Sim_s$ can be obtained by computing the distances between the test samples ($C_i$) and the unseen-class prototypes ($C_u$). The recognition task can be performed the same way as that in the visual space.

**Combining Multiple Spaces**. Due to the fact that the visual space is discriminative, the semantic space is more generative, and the aligned space is a compromise, combining multiple spaces would improve the performance. In our framework, we simply combine the similarities obtained in each space, *i.e.* combining the visual space and aligned space by $Sim_{va} = Sim_v + Sim_a$, and use the same nearest neighbour approach to perform recognition task.

### 3.5 Difference from Relevant Works

Among prior works, the most relevant one to ours is [4], where the structures in the semantic space and visual space are also utilized. However, the key ideas of the two works are quite different. [4] uses fixed semantic information and directly shares its structure to the visual space to form unseen classifiers. It doesn't consider whether the two spaces are consistent or not since the semantic information is obtained independently of the visual exemplars. While our approach focuses on aligning the visual-semantic structure and then shares the aligned structures to form unseen-class prototypes in different spaces. Moreover,

[1] Source code of CDL is available at *http://vipl.ict.ac.cn/resources/codes*.

Table 1: Statistics for attribute datasets: aPY , AwA , CUB and SUNA in terms of image numbers (*Img*), attribute numbers (*Attr*), training + validation seen class numbers (*Seen*) and unseen class numbers (*Unseen*)

| **Dataset** | *Img* | *Attr* | *Seen* | *Unseen* |
|---|---|---|---|---|
| **aPY** [9] | 15,339 | 64 | 15 + 5 | 12 |
| **AwA** [19] | 30,475 | 85 | 27 + 13 | 10 |
| **CUB** [32] | 11,788 | 312 | 100 + 50 | 50 |
| **SUNA** [25] | 14,340 | 102 | 580 + 65 | 72 |

[4] learns visual classifiers independently of the semantic information while our approach automatically learns the class prototypes in the visual space by jointly leveraging the semantic information. Furthermore, to make the model more suitable to the unseen classes to tackle the challenging domain shift problem, which is not addressed in [4], we propose to utilize the unseen-class semantics to make domain adaptation. Another work [34] also uses structure constraints to learn visual-semantic embeddings. However, it deals with the sample structure, where the distances among samples are preserved. While our approach aligns the class structures, which aims to learn more robust class prototypes.

## 4 Experiments

### 4.1 Datasets and Settings

**Datasets**. Following the new data splits proposed by [36], we perform experiments on four bench-mark **ZSL** datasets, *i.e.* aPascal & aYahoo (aPY) [9], Animals with Attributes (AwA) [19], Caltech-UCSD Birds-200-2011 (CUB) [32], SUN Attribute (SUNA) [25], to verify the effectiveness of the proposed framework. The statistics of all datasets are shown in Table 1.

**Settings**. To make fair comparisons, we use the class semantics and image features provided by [36]. Specifically, the attribute vectors are utilized as the class semantics and the image features are extracted by the 101-layered ResNet [13]. Parameters $(\lambda, \alpha, \beta, \gamma)$ in the proposed framework are fine-tuned in the range $[0.001, 0.01, 0.1, 1, 10]$ using the train and validation splits provided by [36]. More details about the parameters can be seen in the supplementary material. We use the average per-class top-1 accuracy to measure the performance of our models.

### 4.2 Evaluations of Different Spaces

The proposed framework involves three spaces, *i.e.* visual space (v), aligned space (a) and semantic space (s). As is described above, zero-shot recognition can be performed in each space independently or in the combined space, and the

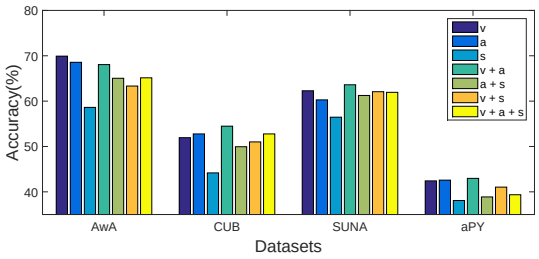


Fig. 3: Zero-shot recognition results via different evaluation spaces, *i.e.* visual space (v), aligned space (a), semantic space (s), combination of visual space and aligned space (v + a) and other combinations, as is described in Section 3.4.

recognition results are shown in Figure 3. It can be seen that the performance in the visual space is higher than that in the semantic space, which indicates that the incomplete semantic information is usually less discriminative. By aligning the visual-semantic structures, the discriminative property of the semantic space improves a lot, which can be inferred from the comparisons between the aligned space and the semantic space. Moreover, the recognition performance will be further improved by combining the visual space and the aligned space, since the visual space is more discriminative and the aligned space is more extensive. For AwA, the best performance is obtained in the visual space. Perhaps the visual space is discriminative enough and it is not complementary with other spaces, so combining it with others will pull down its performance.

### 4.3 Comparison with State-of-the-Art

To demonstrate the effectiveness of the proposed framework, we compare our method with several popular approaches and the recognition results on the four datasets are shown in Table 2. We report our results in the best space for each dataset, as is analyzed in Section 4.2. It can be seen that our framework achieves the best performance on three datasets and is comparable to the best approach on CUB, which indicates the effectiveness of our framework. SAE [16] gets poor performance on aPY probably due to that it is not robust to the weak relations between seen and unseen classes. We owe the success of **CDL** to the structure alignment procedure. Different from other approaches, where fixed semantic information is utilized to perform the recognition task, we automatically adjust the semantic space by aligning the visual-semantic structures. Since the visual space is more discriminative and the semantic space is more extensive, it will benefit each other by aligning the structures for the two spaces. Compared with [4], we get slightly lower result on CUB and this may be caused by the less discriminative class structures. CUB is a fine-grained dataset, where most classes are very similar, so less discriminative class relations could be obtained in the visual space. While [4] learns more complicated image classifiers to enhance the discriminative property in the visual space.

Table 2: Zero-shot recognition results on aPY, AwA, CUB and SUNA (%)

| **Method** | **aPY** | **AwA** | **CUB** | **SUNA** |
|---|---|---|---|---|
| DAP [19] | 33.8 | 44.1 | 40.0 | 39.9 |
| IAP [19] | 36.6 | 35.9 | 24.0 | 19.4 |
| CONSE [23] | 26.9 | 45.6 | 34.3 | 38.8 |
| CMT [30] | 28.0 | 39.5 | 34.6 | 39.9 |
| SSE [39] | 34.0 | 60.1 | 43.9 | 51.5 |
| LATEM [35] | 35.2 | 55.1 | 49.3 | 55.3 |
| ALE [1] | 39.7 | 59.9 | 54.9 | 58.1 |
| DEVISE [10] | 39.8 | 54.2 | 52.0 | 56.5 |
| SJE [2] | 32.9 | 65.6 | 53.9 | 53.7 |
| EZSL [24] | 38.3 | 58.2 | 53.9 | 54.5 |
| SYNC [4] | 23.9 | 54.0 | **55.6** | 56.3 |
| SAE [16] | 8.3 | 53.0 | 33.3 | 40.3 |
| **CDL**(Ours) | **43.0** | **69.9** | 54.5 | **63.6** |

### 4.4 Effectiveness of the Proposed Framework

In order to demonstrate the effectiveness of each component proposed in our framework, we compare our approach with different submodels. The recognition task is performed in the best space according to the datasets. Specifically, for CUB, SUNA, aPY, we evaluate the performance by combining the visual space and the aligned space; for AwA, we evaluate the performance in the visual space. Figure 4 shows the zero-shot recognition results of different submodels. By comparing the performance of "NA" and "CDL", we can figure out that the models will improve a lot by aligning the visual-semantic structures and the less discriminative semantic space will be improved with the help of discriminative visual space. However, if the seen-class prototypes are fixed, it becomes difficult to align the structures between the two spaces and the models degrade seriously, which can be seen through the comparisons of "CDL" and "CDL-Pr". Moreover, the models will be more suitable to the unseen classes by utilizing the unseen-class semantic information to adapt the learning procedure, which is indicated by the comparisons of "CDL" and "CDL-Ad".

### 4.5 Visualization of the Class Structures

In order to have an intuitive understanding of structure alignment, we visualize the class prototypes in the visual space and semantic space on aPY, since the classes in aPY are more easy to understand. In the visual space, we obtain the class prototypes by the mean feature vector of all samples belonging to each class. In the semantic space, we get the class prototypes directly from the semantic representations. Then we use multidimensional scaling (MDS) approach [18] to visualize the class prototypes, where the relations of all classes are preserved. The original class structures in the semantic space and the visual space are shown

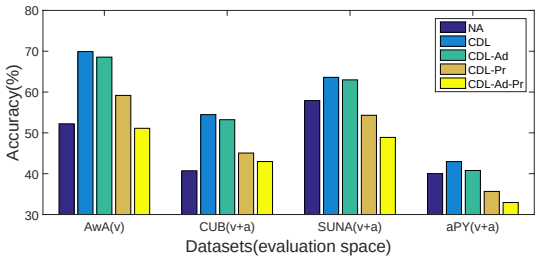


Fig. 4: Comparisons of different baseline methods. NA: not aligning the visual-semantic structure, as is done in the initialization period. CDL: The proposed framework. CDL-Ad: CDL without the adaptation term (second term). CDL-Pr: CDL without the prototype learning term (third term), where $P_s$ is fixed as the means of visual samples in each class. CDL-Ad-Pr: CDL without the adaptation term and the prototype learning term.

in the first row of Figure 5. To make the figure more intuitive, we manually gathered the classes into three groups, *i.e.* Vehicle, Animal and House. We can figure out that the class structures in the semantic space are not discriminative enough, as can be seen by the tight structures among animals, while those in the visual space are more discriminative. Moreover, the structures between these two space are seriously inconsistent, so directly sharing the structures from the semantic space to the visual space to synthesize the unseen-class prototypes will degrade the model. Therefore, we propose to learn the representation bases in each space to reformulate the class prototypes and align the class structures in a common space. It can be seen that the semantic structures become more discriminative after structure alignment. For example, in the original semantic space, *dog* and *cat* are mostly overlapped and they are separated after structure alignment with the help of their relations in the visual space. Thus the aligned semantic space becomes more discriminative to different classes. Moreover, the aligned structures in the two spaces become more consistent than those in the original spaces.

### 4.6 Visualization of Class Prototypes

The prototype of one class should locate near the samples belonging to the corresponding class. In order to check whether the prototypes are properly learned, we visualize the prototypes and corresponding samples in the visual space. To have more intuitive understanding, we choose 10 seen classes and 5 unseen classes from AwA. Then we use t-SNE [21] to project the visual samples and class prototypes to a 2-D plane. The visualization results are shown in Figure 6. It can be seen that most prototypes locate near the samples belonging to the same classes. Although the unseen prototypes deviate from the centers of corresponding samples due to the fact that no corresponding images are provided for training, they are still discriminative enough to classify different classes, which shows the

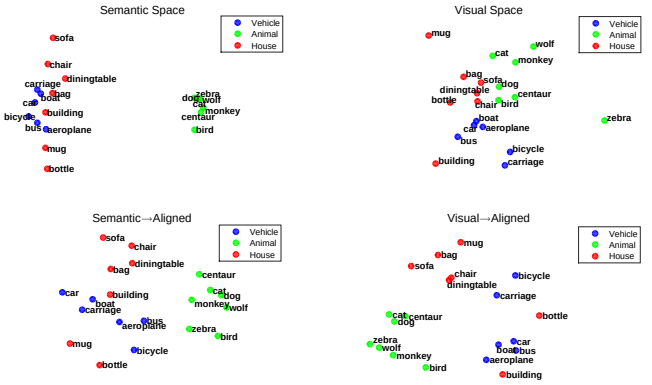


Fig. 5: Visualization of the seen-class prototypes in the semantic space and visual space before and after structure alignment on aPY. To make it intuitive, the classes are manually clustered into three groups, *i.e.*Vehicle, Animal and House.

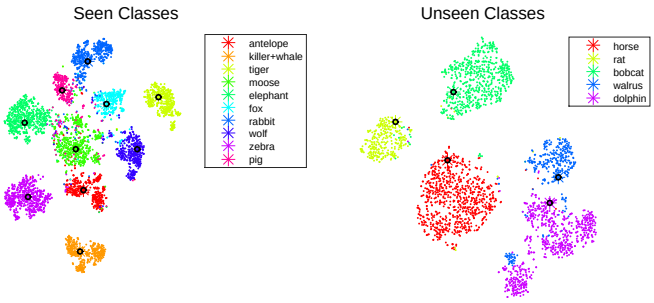


Fig. 6: Visualization of class prototypes on AwA in the feature space by t-SNE. The prototypes are represented by "*" with colors corresponding to the classes. To make them visible, we use black circles to mark them.

expansibility of our structure alignment approach for prototype learning. More visualization results can be seen in the supplementary material.

### 4.7 Generalized Zero-Shot Learning

To demonstrate the effectiveness of the proposed framework, we also apply our method to the generalized zero-shot learning (**GZSL**) task, where the seen class are also considered in the test procedure. The task for **GZSL** is to learn images classifiers $f_{gzsl} : \mathcal{X} \rightarrow \mathcal{Y}^s \bigcup \mathcal{Y}^u$. We adopt the data splits provided by [36] and compare our method with several popular approaches. Table 3 shows the generalized zero-shot recognition results on the four datasets. It can be seen that most approaches get low accuracy on the unseen-class samples because of overfitting the seen classes, while our framework gets better results on the unseen

Table 3: Generalized zero-shot learning results on aPY, AwA, CUB and SUNA. ts = Top-1 accuracy of the test unseen-class samples, tr = Top-1 accuracy of the test seen-class samples, H = harmonic mean (CMT*: CMT with novelty detection). We measure top-1 accuracy in %.

| Method | aPY | | | AwA | | | CUB | | | SUNA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ts | tr | H | ts | tr | H | ts | tr | H | ts | tr | H |
| DAP [19] | 4.8 | 78.3 | 9.0 | 0.0 | **88.7** | 0.0 | 1.7 | 67.9 | 3.3 | 4.2 | 25.1 | 7.2 |
| IAP [19] | 5.7 | 65.6 | 10.4 | 2.1 | 78.2 | 4.1 | 0.2 | 72.8 | 0.4 | 1.0 | 37.8 | 1.8 |
| CONSE [23] | 0.0 | **91.2** | 0.0 | 0.4 | 88.6 | 0.8 | 1.6 | 72.2 | 3.1 | 6.8 | 39.9 | 11.6 |
| CMT [30] | 1.4 | 85.2 | 2.8 | 0.9 | 87.6 | 1.8 | 7.2 | 49.8 | 12.6 | 8.1 | 21.8 | 11.8 |
| CMT* [30] | 10.9 | 74.2 | 19.0 | 8.4 | 86.9 | 15.3 | 4.7 | 60.1 | 8.7 | 8.7 | 28.0 | 13.3 |
| SSE [39] | 0.2 | 78.9 | 0.4 | 7.0 | 80.5 | 12.9 | 8.5 | 46.9 | 14.4 | 2.1 | 36.4 | 4.0 |
| LATEM [35] | 0.1 | 73.0 | 0.2 | 7.3 | 71.7 | 13.3 | 15.2 | 57.3 | 24.0 | 14.7 | 28.8 | 19.5 |
| ALE [1] | 4.6 | 73.7 | 8.7 | 16.8 | 76.1 | 27.5 | 23.7 | 62.8 | **34.4** | **21.8** | 33.1 | 26.3 |
| DEVISE [10] | 4.9 | 76.9 | 9.2 | 13.4 | 68.7 | 22.4 | **23.8** | 53.0 | 32.8 | 16.9 | 27.4 | 20.9 |
| SJE [2] | 3.7 | 55.7 | 6.9 | 11.3 | 74.6 | 19.6 | 23.5 | 59.2 | 33.6 | 14.1 | 30.5 | 19.8 |
| EZSL [24] | 2.4 | 70.1 | 4.6 | 6.6 | 75.6 | 12.1 | 12.6 | 63.8 | 21.0 | 11.0 | 27.9 | 15.8 |
| SYNC [4] | 7.4 | 66.3 | 13.3 | 8.9 | 87.3 | 16.2 | 11.5 | **70.9** | 19.8 | 7.9 | **43.3** | 13.4 |
| SAE [16] | 0.4 | 80.9 | 0.9 | 1.8 | 77.1 | 3.5 | 7.8 | 54.0 | 13.6 | 8.8 | 18.0 | 11.8 |
| **CDL**(Ours) | **19.8** | 48.6 | **28.1** | **28.1** | 73.5 | **40.6** | 23.5 | 55.2 | 32.9 | 21.5 | 34.7 | **26.5** |

classes and achieves more balanced results between the seen and unseen classes. By jointly aligning the visual-semantic structures and utilizing the semantic information of unseen classes to make an adaption, our model has less tendency to overfit the seen classes.

## 5 Conclusions

In this paper, we propose a coupled dictionary learning framework to align the visual-semantic structures for zero-shot learning, where unseen-class prototypes are learned by sharing the aligned structures. Extensive experiments on four bench-mark datasets show the effectiveness of the proposed approach. The success of **CDL** should be owing to three characters. First, instead of using the fixed semantic information to perform recognition task, our structure alignment approach shares the discriminative property lying in the visual space and the extensive property lying in the semantic space, which benefits each other and improves the incomplete semantic space. Second, by utilizing the unseen-class semantics to adapt the learning procedure, our model is more suitable for the unseen classes. Third, the class prototypes are automatically learned by sharing the aligned structures, which makes it possible to directly perform recognition task using simple nearest neighbour approach. Moreover, we combine the information of multiple spaces to improve the recognition performance.

**Acknowledgements.** This work is partially supported by Natural Science Foundation of China under contracts Nos. 61390511, 61772500, 973 Program under contract No. 2015CB351802, Frontier Science Key Research Project CAS No. QYZDJ-SSW-JSC009, and Youth Innovation Promotion Association CAS No. 2015085.

## References


1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for attribute-based classification. In: Proc. of Computer Vision and Pattern Recognition. pp. 819–826 (2013)
2. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: Proc. of Computer Vision and Pattern Recognition. pp. 2927–2936 (2015)
3. Bucher, M., Herbin, S., Jurie, F.: Improving semantic embedding consistency by metric learning for zero-shot classification. In: Proc. of European Conference on Computer Vision (2016)
4. Changpinyo, S., Chao, W.L., Gong, B., Sha, F.: Synthesized classifiers for zero-shot learning. Proc. of Computer Vision and Pattern Recognition pp. 5327–5336 (2016)
5. Changpinyo, S., Chao, W.L., Sha, F.: Predicting visual exemplars of unseen classes for zero-shot learning. Proc. of International Conference on Computer Vision pp. 3496–3505 (2017)
6. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: Proc. of European Conference on Computer Vision (2016)
7. Demirel, B., Cinbis, R.G., Ikizler-Cinbis, N.: Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. Proc. of International Conference on Computer Vision pp. 1241–1250 (2017)
8. Ding, Z., Shao, M., Fu, Y.: Low-rank embedded ensemble semantic dictionary for zero-shot learning. Proc. of Computer Vision and Pattern Recognition pp. 6005–6013 (2017)
9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Proc. of Computer Vision and Pattern Recognition. pp. 1778–1785 (2009)
10. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Proc. of Advances in Neural Information Processing Systems. pp. 2121–2129 (2013)
11. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Transductive multi-view zero-shot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**, 2332–2345 (2015)
12. Fu, Z.Y., Xiang, T.A., Kodirov, E., Gong, S.: Zero-shot object recognition by semantic manifold distance. Proc. of Computer Vision and Pattern Recognition pp. 2635–2644 (2015)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of Computer Vision and Pattern Recognition. pp. 770–778 (2016)
14. Jiang, H., Wang, R., Shan, S., Yang, Y., Chen, X.: Learning discriminative latent attributes for zero-shot classification. Proc. of International Conference on Computer Vision pp. 4233–4242 (2017)
15. Kodirov, E., Xiang, T., Fu, Z.Y., Gong, S.: Unsupervised domain adaptation for zero-shot learning. Proc. of International Conference on Computer Vision pp. 2452–2460 (2015)
16. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. Proc. of Computer Vision and Pattern Recognition pp. 4447–4456 (2017)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proc. of Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)

18. Kruskal, J.B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**(1), 1–27 (1964)
19. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: Proc. of Computer Vision and Pattern Recognition. pp. 951–958 (2009)
20. Long, Y., Liu, L., Shen, F., Shao, L., Li, X.: Zero-shot learning using synthesised unseen visual data with diffusion regularisation. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2018)
21. van der Maaten, L., Hinton, G.E.: Visualizing data using t-sne. In: Journal of Machine Learning Research. vol. 9, pp. 2579–2605 (2008)
22. Morgado, P., Vasconcelos, N.: Semantically consistent regularization for zero-shot recognition. Proc. of Computer Vision and Pattern Recognition pp. 2037–2046 (2017)
23. Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J.: Zero-shot learning by convex combination of semantic embeddings. Proc. of International Conference on Learning Representations (2014)
24. Paredes, B.R., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: Proc. of International Conference on Machine Learning. pp. 2152–2161 (2015)
25. Patterson, G., Xu, C., Su, H., Hays, J.: The SUN attribute database: Beyond categories for deeper scene understanding. International Journal of Computer Vision **108**(1-2), 59–81 (2014)
26. Reed, S.E., Akata, Z., Schiele, B., Lee, H.: Learning deep representations of fine-grained visual descriptions. Proc. of Computer Vision and Pattern Recognition pp. 49–58 (2016)
27. Romera-Paredes, B., Torr, P.H.S.: An embarrassingly simple approach to zero-shot learning. In: Proc. of International Conference on Machine Learning (2015)
28. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
30. Socher, R., Ganjoo, M., Sridhar, H., Bastani, O., Manning, C.D., Ng, A.Y.: Zero-shot learning through cross-modal transfer. In: Proc. of Advances in Neural Information Processing Systems. pp. 935–943 (2013)
31. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proc. of Computer Vision and Pattern Recognition. pp. 1–9 (2015)
32. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. rep. (2011)
33. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.J.: The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology (2011)
34. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. Proc. of Computer Vision and Pattern Recognition pp. 5005–5013 (2016)
35. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q.N., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. Proc. of Computer Vision and Pattern Recognition pp. 69–77 (2016)
36. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning - the good, the bad and the ugly. In: Proc. of Computer Vision and Pattern Recognition (2017)

37. Xu, X., Shen, F., Yang, Y., Zhang, D., Shen, H.T., Song, J.: Matrix tri-factorization with manifold regularizations for zero-shot learning. Proc. of Computer Vision and Pattern Recognition pp. 2007–2016 (2017)
38. Zhang, L., Xiang, T., Gong, S.: Learning a deep embedding model for zero-shot learning. Proc. of Computer Vision and Pattern Recognition pp. 3010–3019 (2017)
39. Zhang, Z., Saligrama, V.: Zero-shot learning via semantic similarity embedding. Proc. of International Conference on Computer Vision pp. 4166–4174 (2015)
40. Zhang, Z., Saligrama, V.: Zero-shot learning via joint latent similarity embedding. Proc. of Computer Vision and Pattern Recognition pp. 6034–6042 (2016)
41. Zhu, X., Anguelov, D., Ramanan, D.: Capturing long-tail distributions of object subcategories. In: Proc. of Computer Vision and Pattern Recognition. pp. 915–922 (2014)