From Zero-shot Learning to Conventional Supervised Classification: Unseen Visual Data Synthesis

Yang Long[1], Li Liu[2], Ling Shao[2], Fumin Shen[3], Guiguang Ding[4], and Jungong Han[5]

[1]Department of Electronic and Electrical Engineering, University of Sheffield, UK
[2]School of Computing Science, University of East Anglia, UK
[3]Center for Future Media, University of Electronic Science and Technology of China, China
[4]School of Software, Tsinghua University, China
[5]Department of Computer Science and Digital Technologies, Northumbria University, UK

[1]ylong2@sheffield.ac.uk, {li.liu, ling.shao}@uea.ac.uk, fumin.shen@gmail.com,
dinggg@tsinghua.edu.cn, jungong.han@northumbria.ac.uk

Abstract

Robust object recognition systems usually rely on powerful feature extraction mechanisms from a large number of real images. However, in many realistic applications, collecting sufficient images for ever-growing new classes is unattainable. In this paper, we propose a new Zero-shot learning (ZSL) framework that can synthesise visual features for unseen classes without acquiring real images. Using the proposed Unseen Visual Data Synthesis (UVDS) algorithm, semantic attributes are effectively utilised as an intermediate clue to synthesise unseen visual features at the training stage. Hereafter, ZSL recognition is converted into the conventional supervised problem, i.e. *the synthesised visual features can be straightforwardly fed to typical classifiers such as SVM. On four benchmark datasets, we demonstrate the benefit of using synthesised unseen data. Extensive experimental results suggest that our proposed approach significantly improve the state-of-the-art results.*

1. Introduction

Object Recognition is arguably one of the most fundamental tasks in computer vision field. Most of the conventional frameworks, *e.g.* Deep Neural Networks (DNN) [22], rely on a large number of training samples to build statistical models. However, such a premise is unattainable in many real-world situations. The main reasons can be summarised as follows: 1) Obtaining well-annotated training samples is expensive. Although abundant digital images and videos can be retrieved from the Internet, existing search engines crucially depend on user-defined keywords that are often vague and not suitable for learning tasks. 2) The number of newly defined classes is ever-growing. Meanwhile, fine-grained tasks make existing categories go deeper, *e.g.* to recognise a newly released handbag in a novel pattern. Training a particular model for each of them is infeasible. 3) Collecting instances for rare classes is difficult. For example, one might wish to detect an ancient or rare species automatically. It could be difficult to provide even a single example for them since available knowledge could be only textual descriptions or some distinctive attributes.

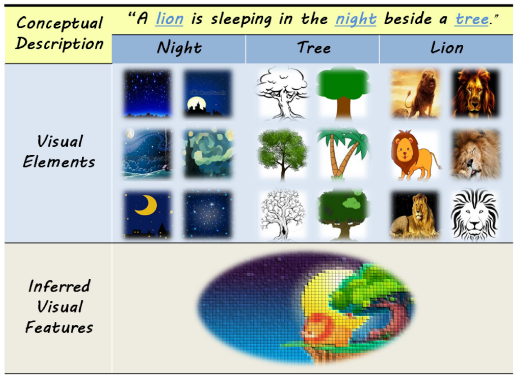

Figure 1. Given a conceptual description, human can imagine the outline of the scene by combining previous seen visual elements.

As a feasible solution, *Zero-shot Learning* (ZSL) aims to leverage a closed-set of semantic models that can generalise to unseen classes [25, 23]. The common paradigm of ZSL methods first train a prediction model that can map visual data to a semantic representation. Hereafter, new objects

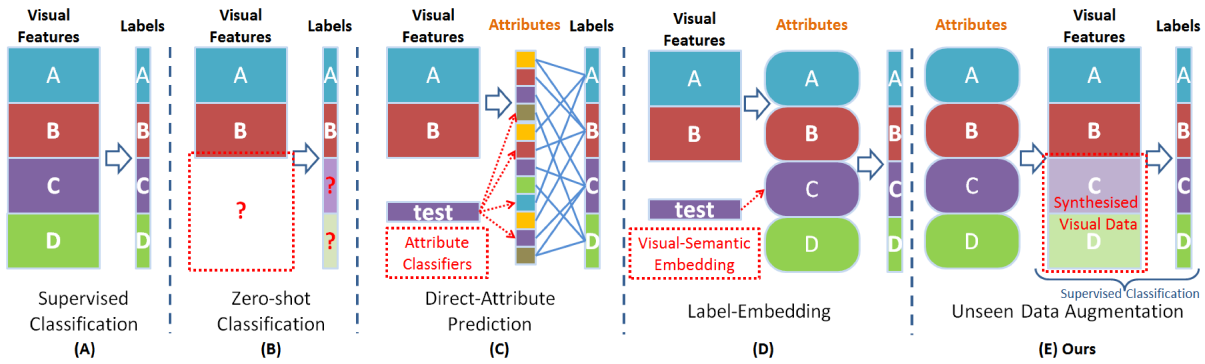

Figure 2. Comparison of supervised and zero-shot classifications and existing ZSL frameworks. (A) a typical supervised classification: the training samples and labels are in pairs; (B) a zero-shot learning problem: without training samples, the classes *C* and *D* cannot be predicted; (C) Direct-Attribute Prediction model uses attributes as intermediate clues to associate visual features to class labels; (D) label-embedding: the attributes are concatenated as a semantic embedding; (E) we inversely learn an embedding from the semantic space to visual space and convert the ZSL problem into conventional supervised classification.

can be recognised by only knowing their semantic descriptions. However, existing methods cannot expand the training data for new unseen classes. As illustrated in Fig. 2, such frameworks impede existing methods from scaling up since the fixed seen data is eventually limited to represent the ever-growing semantic concepts.

In this paper, we investigate to synthesise high-quality visual features from semantic attributes so that the ZSL problem can be converted into conventional supervised classification. Our idea is inspired by the ability of human imagination, as shown in Fig. 1. Given a semantic description, we human can associate familiar visual elements and then imagine an approximate scene. Accordingly, we synthesise discriminative low-level features from semantic attributes to substitute feature extraction from real images. Our contributions can be summarised as follows:

1) We provide a feasible framework to synthesise unseen visual features from given semantic attributes without acquiring real images. The synthesised data obtained at the training stage can be straightforwardly fed to conventional classifiers so that ZSL recognition is skilfully converted into the conventional supervised problem and leads to state-of-the-art recognition performance on four benchmark datasets.

2) We introduce the *variance decay* problem during semantic-visual embedding and propose a novel *Diffusion Regularisation* that can explicitly make information diffuse to each dimension of the synthesised data. We achieve information diffusion by optimising an orthogonal rotation problem. We provide an efficient optimisation strategy to solve this problem together with the *structural difference* and *training bias* problem.

2. Related Work

Zero-shot Recognition Schemes: We summarise previous ZSL schemes in Fig. 2, in contrast to conventional supervised classification (Fig. 2(A)). Since collecting well-labelled visual data for novel classes is expensive, as shown in Fig. 2(B), zero-shot learning techniques [25, 23, 39, 35, 38, 32] are proposed to recognise novel classes without acquiring the visual data. Most of the early works are based on the Direct-Attribute Prediction (DAP) model [23]. Such a model utilises semantic attributes as intermediate clues. A test sample is classified by each attribute classifier alternately, and the class label is predicted by probabilistic estimation. Admitting the merit of DAP, there are some concerns about its deficiencies. [19] points out that the attributes may correlate to each other resulting in significant information redundancy and poor performance. The human labelling involved in attribute annotation may also be unreliable [18, 50].

To circumvent learning independent attributes, embedding-based ZSL frameworks (Fig.2(C)) are proposed to learn a projection that can map the visual features to all of the attributes at once. The class label is then inferred in the semantic space using various measurements [2, 34, 27, 4, 14, 45]. Since the attribute annotations are expansive to acquire, attributes are substituted by the visual similarity and data distribution information in transductive ZSL settings [40, 51, 13, 12, 28, 21, 54, 55, 56]. However, these methods involve the data of unseen classes to learn the model, which to some extent breaches the strict ZSL settings. Recent work [43, 49, 30] combines the embedding-inferring procedure into a unified framework and empirically demonstrates better performance. The closest related work is [7, 8, 31], which takes one-step further to synthesise classifiers or prototypes for unseen classes.

Our method takes the advantages of semantic embedding. However, the inference direction is different from existing work. Our method aims to inversely synthesise visual feature vectors to as many as the available semantic

instances rather than mapping visual data to the label space.

Semantic Side Information: ZSL tasks require to leverage side information as intermediate clues. Such frameworks not only broaden the classification settings but also enable various information to aid visual systems. Since textual sources are relatively easy to obtain from the Internet, [42, 33] propose to estimate the semantic relatedness of the novel classes from the text. [26, 10, 26] learn pseudo-concepts to associate novel classes using Wikipedia articles. Recently, lexical hierarchies in the ontology engineering are also exploited to find the relationships between classes [41, 5, 3].

Although various side information is studied, attribute-based ZSL methods still gain the most popularity. One reason is ZSL by learning attributes often gives prominent classification performance [53, 52, 17, 55, 54]. For another reason, attribute representation is a compact way that can further describe an image by concrete words that are human-understandable [11, 29, 15, 1]. Various types of attributes are proposed to enrich applicable tasks and improve the performance, such as relative attributes [36], class-similarity attributes [52], and augmented attributes [44]. Our main motivation of this paper not only aims to improve the ZSL performance, but also seeks for a reliable solution for synthesising high-quality visual features.

3. Approach

Preliminaries The training set contains *centralised* visual features, attributes, and seen class labels that are in 3-tuples: $(x_1, a_1, y_1), ..., (x_N, a_N, y_N) \subseteq \mathcal{X}_s \times \mathcal{A}_s \times \mathcal{Y}_s$, where N is the number of training samples; $\mathcal{X}_s = [x_{nd}] \in \mathbb{R}^{N \times D}$ is a D-dimensional feature space; $\mathcal{A}_s = [a_{nm}] \in \mathbb{R}^{N \times M}$ is an M-dimensional attribute space; and $y_n \in \{1, ..., C\}$ consists of C discrete class labels. Our framework can cope with either *class-level* or *image-level* attributes. For class-level, the instances in the same class share the attributes. Given $\hat{N}$ pairs of instances with semantic attributes from $\hat{C}$ unseen classes: $(\hat{a}_1, \hat{y}_1), ..., (\hat{a}_{\hat{N}}, \hat{y}_{\hat{N}}) \subseteq \mathcal{A}_u \times \mathcal{Y}_u$, where $\mathcal{Y}_u \cap \mathcal{Y}_s = \emptyset$, $\mathcal{A}_u = [a_{\hat{n}m}] \in \mathbb{R}^{\hat{N} \times M}$, the goal of zero-shot learning is to learn a classifier, $f : \mathcal{X}_u \rightarrow \mathcal{Y}_u$, where the samples in $\mathcal{X}_u$ are completely unavailable during training. We use *Calligraphic* typeface to indicate a space. Subscripts s and u refer to 'seen' and 'unseen'. *hat* denotes the variables that are related to 'unseen' samples.

Unseen Visual Data Synthesis: We aim to synthesise the visual features of unseen classes by the given semantic attributes. Specifically, we learn an embedding function on the training set $f' : \mathcal{A}_s \rightarrow \mathcal{X}_s$. After that, we are able to infer $\mathcal{X}_u$ through: $\mathcal{X}_u = f'(\mathcal{A}_u)$.

Zero-shot Recognition: Using the synthesised visual features, the ZSL recognition is converted to a typical classification problem. It is straightforward to employ conventional supervised classifiers, *e.g*. SVM, to predict the labels of unseen classes $f_{\text{SVM}} : \mathcal{X}_u \rightarrow \mathcal{Y}_u$.

3.1. Unseen Visual Data Synthesis

To synthesise visual features, the most intuitive framework is to learn a mapping function from the semantic space to the visual feature space:

$$\min_P \mathcal{L}(\mathcal{A}_s P, \mathcal{X}_s) + \lambda \Omega(P), \tag{1}$$

where P is the projection matrix, $\mathcal{L}$ is a loss function, and Ω is a regularisation term with its hyper-parameter λ. It is common to choose $\Omega(P) = \|P\|_F^2$, where $\|.\|_F$ is the Frobenius norm of a matrix that estimates the Euclidean distance between two matrices. Before the test, we can synthesise unseen visual features from the attribute space by given attributes of the unseen instances:

$$\mathcal{X}_u = \mathcal{A}_u P. \tag{2}$$

Visual-Semantic Structure Preservation In spite of the simplicity of the above framework, we confront two main problems as follows. 1) *Structural difference*: in practice, there is often a huge gap between visual and semantic spaces. In pursuance of minimum reconstruction error, the model tends to learn principal components between the two spaces. Consequently, the synthesised data would be not discriminant enough for ZSL purposes. 2) *Training bias*: the synthesised unseen data can be biased towards the 'seen' data and gains a different data distribution to the real unseen data. This problem is due to the regression-based framework does not discover the intrinsic geometric structure of the semantic space and cannot capture the unseen-to-seen relationships. Thus, directly mapping from semantic to visual space can lead to inferior performance. We propose to introduce an auxiliary latent-embedding space $\mathcal{V}$ to reconcile the semantic space with the visual feature space, where $\mathcal{V} = [v_{nd}] \in \mathbb{R}^{N \times D}$. In this way, instead of $\Omega(P)$, we can let $\mathcal{V}$ preserve the intrinsic data structural information of both visual and semantic spaces:

$$J = \|\mathcal{X}_s - \mathcal{V}Q\|_F^2 + \|\mathcal{V} - \mathcal{A}_s P\|_F^2 + \lambda \Omega_1(\mathcal{V}), \tag{3}$$

where the latent-embedding space $\mathcal{V}$ is decomposed from $\mathcal{X}$ and $\mathcal{A}$ is then decomposed from $\mathcal{V}$. $Q = [q_{dd'}] \in \mathbb{R}^{D \times D}$ and $P = [p_{md}] \in \mathbb{R}^{M \times D}$ are two projection matrices. Ω_1 is a *dual-graph* that is introduced next.

We take the *Local Invariance* [6] assumption and solve the problem through a spectral *Dual-Graph* approach. This is a combination of two supervised graphs that aim to simultaneously estimate the data structures of both $\mathcal{X}$ and $\mathcal{A}$. The graph of visual space $W_{\mathcal{X}} \in \mathbb{R}^{N \times N}$ has N vertices $\{g_1, ..., g_N\}$ that correspond to N data points $\{x_1, ..., x_N\}$ in the training set. The semantic graph $W_{\mathcal{A}} \in \mathbb{R}^{N \times N}$ has

the same number of vertices as N instances of attributes $\{a_1, ..., a_N\}$. For *image-level attributes*, we construct k-nn graphs for both visual and semantic spaces, *i.e.* put an edge between each data point x_n (or a_n) and each of its k nearest neighbours. For each pair of the vertices g_i and g_j in the weight matrix (not differ in $W_{\mathcal{X}}$ and $W_{\mathcal{A}}$), the weight can be defined as

$$w_{ij} = \begin{cases} 1, & \text{if } g_i \text{ and } g_j \text{ are connected by an edge} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

As a result, we can separately compute the two weight matrices $W_{\mathcal{X}}$ and $W_{\mathcal{A}}$. It is noteworthy that, for *class-level attributes*, $W_{\mathcal{A}}$ is computed in a slightly different way. Every vertex in the same class is connected by a normalised edge, i.e. $w_{ij} = k/n_c$, if and only if a_i and a_j are from the same class c, where n_c is the size of class c.

In the embedding space $\mathcal{V}$, we expect that, if g_i and g_j in both graphs are connected, each pair of embedded points v_i and v_j are also close to each other. However, sometimes $W_{\mathcal{X}}$ and $W_{\mathcal{A}}$ are not always consistent due to the visual-semantic gap. To compromise such conflicts, we compute the mean of the visual and attribute graphs, i.e. $W = \frac{1}{2}(W_{\mathcal{X}} + W_{\mathcal{A}})$. The resulted regularisation is:

$$\begin{aligned} \Omega_1(\mathcal{V}) &= \frac{1}{2}\sum_{i,j=1}^{N} \|v_i - v_j\|^2 w_{ij} \\ &= Tr(\mathcal{V}^\top \mathbf{D}\mathcal{V}) - Tr(\mathcal{V}^\top W\mathcal{V}) = Tr(\mathcal{V}^\top L\mathcal{V}), \end{aligned} \tag{5}$$

where $\mathbf{D}$ is the degree matrix of W, $\mathbf{D}_{ii} = \sum_i w_{ij}$. L is known as graph Laplacian matrix $L = \mathbf{D} - W$ and $Tr(.)$ computes the trace of a matrix.

Diffusion Regularisation In this paper, we identify another fundamental problem: *variance decay*. When we learn visual features from the attributes, in particular when projecting $\mathcal{A}$ to $\mathcal{V}$ using $\mathcal{P}$, the dimension difference $D \gg M$ will lead the learning algorithm to pick the directions with low variances progressively. As shown in Fig. 3, most of the information (variance) is contained in a few projections. As a result, the remaining dimensions of the synthesised data suffers a dramatic variance decay, which indicates the learnt representation is severely redundant. To address the problem, we may expect the concentrated information can effectively diffuse to all of the learnt dimensions through an adjustment rotation [20]. Therefore, we modify the rotating matrix Q in Eq. (3). In this paper, we consider an orthogonal rotation, *i.e.* $QQ^\top = I$, since it is easy to show that $Tr(Q^\top P^\top \mathcal{A}^\top \mathcal{A}PQ) = Tr(P^\top \mathcal{A}^\top \mathcal{A}P)$ (I is an identity matrix). Such a property is reported in [16] that the orthogonal rotation can protect the properties captured in the semantic space. Next, we show how the rotation can control variance diffusion.

From Eq. (3), the optimal synthesised data is $\mathcal{X} = \mathcal{V}Q$, where $\mathcal{V} = \mathcal{A}P$. We first prove that the overall variance does not change after rotation. Before rotation, $\mathcal{V}$ is centralised, i.e. $\sum_{n=1}^{N} v_n = \mathbf{0}$. The original overall variance Γ of $\mathcal{V}$ is $\Gamma = N\sum_{d=1}^{D}\sigma_d$, where $\sigma_d = (\sum_{n=1}^{N} v_{nd}^2)/N$ denotes the variance of the d-th dimension. After rotation Q, we have the new variance of each dimension σ'_d and the sum of variance of each dimension is Γ'. We show $\Gamma = \Gamma'$ in the following:

$$\begin{aligned} \Gamma &= \sum_{d=1}^{D}\sum_{n=1}^{N} v_{nd}^2 = \|\mathcal{V}\|_F^2 = Tr(\mathcal{V}\mathcal{V}^\top) \\ &= Tr(\mathcal{V}QQ^\top\mathcal{V}^\top) = \|\mathcal{V}Q\|_F^2 \\ &= \sum_{d=1}^{D}\sum_{n=1}^{N} x_{nd}^2 = N\sum_{d=1}^{D}\sigma'_d = \Gamma'. \end{aligned} \tag{6}$$

We hope the overall variance Γ tends to equally diffuse to all of the learnt dimensions in order to recover the real data distribution of $\mathcal{X}$. In other words, the variance of diffused standard deviations Π in the synthesised data should be small ($\Pi = \frac{1}{D}\sum_{d=1}^{D}(\pi_d - \bar{\pi})^2$, where $\pi_d = \sqrt{\sigma'_d}$ and $\bar{\pi}$ is the mean of all standard deviations). According to the above Eq. (6), we have $\sum_{d=1}^{D}\pi_d^2 = \sum_{d=1}^{D}\sigma'_d = \sum_{d=1}^{D}\sigma_d = \epsilon$. Next, we show how to minimise Π in our learning framework to find the orthogonal rotation:

$$\begin{aligned} \Pi &= \frac{1}{D}\sum_{d=1}^{D}(\pi_d - \bar{\pi})^2 \\ &= \frac{1}{D}\sum_{d=1}^{D}\pi_d^2 + \bar{\pi}^2 - \frac{2}{D}\sum_{d=1}^{D}\pi_d\bar{\pi} \\ &= \frac{\epsilon}{D} - \frac{1}{D^2}(\sum_{d=1}^{D}\pi_d)^2. \end{aligned} \tag{7}$$

The above equation shows that to minimise Π is equivalent to maximise the sum of diffused standard deviations. Such a deduction is intuitive because our goal is a higher overall sum of standard deviation so that the synthesised data can gain more information. Moreover, we discover a novel relationship between the sum of diffused standard deviations and the orthogonal rotation:

$$\begin{aligned} \sum_{d=1}^{D}\pi_d &= \sum_{d=1}^{D}\sqrt{\sigma'_d} = \sum_{d=1}^{D}\sqrt{\sum_{n=1}^{N} x_{nd}^2/N} \\ &= \frac{1}{\sqrt{N}}\|\mathcal{X}^\top\|_{2,1} = \frac{1}{\sqrt{N}}\|Q^\top\mathcal{V}^\top\|_{2,1}, \end{aligned} \tag{8}$$

where $\|.\|_{2,1}$ is the $\ell_{2,1}$ norm of a matrix. According to Eq. (7) and Eq. (8), we can simply maximise $\|Q^\top\mathcal{V}^\top\|_{2,1}$ to maximise Π for the purpose of information diffusion. Finally, we can combine the diffusion regularisation with Eq.

(3) and Eq. (5) to form the overall loss function. Such a function aims to minimise the reconstruction error from attributes to visual features, meanwhile preserve the data structure and enable the information to diffuse to all dimensions:

$$\min_{P,Q,\mathcal{V}} \quad J = \|\mathcal{X}_s - \mathcal{V}Q\|_F^2 + \|\mathcal{V} - \mathcal{A}_s P\|_F^2 + \lambda Tr(\mathcal{V}^\top L\mathcal{V}) \\ -\beta\|Q^\top \mathcal{V}^\top\|_{2,1}, \ \ s.t.\ QQ^\top = I. \tag{9}$$

3.2. Optimisation Strategy

The problem raised in Eq. (9) is a non-convex optimisation problem. To the best of our knowledge, there is no direct way to find its optimal solution. Similar to [**?**], in this paper, we propose an iterative scheme by using the alternating optimisation to obtain the local optimal solution. Specifically, we initialise $Q = I$ and $\mathcal{V} = \mathcal{X}_s$.The initialisation of P can be obtained via $P = (\mathcal{A}_s^\top \mathcal{A}_s)^{-1}\mathcal{A}_s^\top \mathcal{V}$. The whole alternate procedure of the proposed UVDS is listed as follows.

1. $\mathcal{V}$-step: By fixing P and Q, we can reduce Eq. (9) to the following sub-problem:

$$\min_{\mathcal{V}} \|\mathcal{X}_s - \mathcal{V}Q\|_F^2 + \|\mathcal{V} - \mathcal{A}_s P\|_F^2 + \lambda Tr(\mathcal{V}^\top L\mathcal{V}) \\ - \beta\|Q^\top \mathcal{V}^\top\|_{2,1} + \gamma\|\mathbf{1}\mathcal{V}\|_2^2, \tag{10}$$

where the extra term $\gamma\|\mathbf{1}\mathcal{V}\|_2^2$ constrains the learnt $\mathcal{V}$ to be centralised according to Eq. 6. The minimal $\mathcal{V}$ can be obtained by setting the partial derivative of Eq. (10) to zero and we have

$$\begin{aligned} \frac{\partial J}{\partial \mathcal{V}} &= 2(\mathcal{V}Q - \mathcal{X})Q^\top + 2(\mathcal{V} - \mathcal{A}P) \\ &+ 2\lambda L\mathcal{V} - \beta\mathcal{V}QEQ^\top + \gamma\mathbf{1}^\top\mathbf{1}\mathcal{V} = 0, \end{aligned} \tag{11}$$

where $E = diag(e_1, \ldots, e_d, \ldots, e_D) \in \mathbb{R}^{D\times D}$ and the d-th element of E is $e_d = 1/(\sqrt{N}\pi_d)$. By merging the like terms, Eq. (11) can be rewritten as

$$\mathcal{V}(2QQ^\top + 2\alpha I + \beta QEQ^\top) + (2\lambda L + \gamma\mathbf{1}^\top\mathbf{1})\mathcal{V} \\ -(XQ^\top + 2\mathcal{A}P) = 0, \tag{12}$$

which is a typical Sylvester equation so that $\mathcal{V}$ can be efficiently solved by the *lyap()* function in the MATLAB. Afterwards, the leant $\mathcal{V}$ needs to be further centralised: $v_n \leftarrow v_n - (\sum_{n=1}^N v_n)/N$ to satisfy Eq. 6.

2. Q-step: By fixing P and V, we can reduce Eq. (9) to the following sub-problem:

$$\min_{\mathcal{Q}} \|\mathcal{X}_s - \mathcal{V}Q\|_F^2 - \beta\|Q^\top \mathcal{V}^\top\|_{2,1}, \ \ s.t.\ QQ^\top = I \tag{13}$$

Since we need to solve Q with the orthogonality constraint in Eq. (13), in this paper, we adopt the gradient flow in [47] which is an iterative scheme for optimising generic orthogonal problems with a feasible solution. Specifically, given the orthogonal rotation Q_t during the t-th iterative optimisation, a better solution of Q_{t+1} is updated via *Cayley transformation*:

$$Q_{t+1} = H_t Q_t, \tag{14}$$

where H_t is the *Cayley transformation* matrix and defined as

$$H_t = (\mathrm{I} + \frac{\tau}{2}\Phi_t)^{-1}(\mathrm{I} - \frac{\tau}{2}\Phi_t), \tag{15}$$

where I is the identity matrix, $\Phi_t = \Delta_t Q_t^\top - Q_t\Delta_t^\top$ is the skew-symmetric matrix, τ is an approximate minimiser satisfying Armijo-Wolfe conditions [48] and Δ is the partial derivative of Eq. (13) with respect to Q as

$$\Delta_t = \mathcal{V}^\top(\mathcal{V}Q_t - \mathcal{X}_s) - \beta\mathcal{V}^\top\mathcal{V}Q_t E., \tag{16}$$

where the diagonal matrix E is defined the same as that in Eq. (11). In this way, for the Q-step, we repeat the above formulation to update Q until achieving convergence.

3. P-step: By fixing Q and V, we can reduce Eq. (9) to the following sub-problem:

$$\min_{P} \alpha\|\mathcal{V} - \mathcal{A}_s P\|_F^2. \tag{17}$$

The resulted equation is derived by a standard least squares problem with the following analytical solution:

$$P = (\mathcal{A}_s^\top \mathcal{A}_s)^{-1}\mathcal{A}_s^\top \mathcal{V}. \tag{18}$$

In this way, we sequentially update $\mathcal{V}$, $\mathbf{Q}$ and $\mathbf{P}$ to optimise UVDS with T times based on coordinate descent. For each variable, either global or local optimum is achieved and thus the overall objective is lower bounded, which guarantees the convergence of our method. In practice, UVDS can well converge with $T = 5 \sim 10$.

3.3. Zero-shot Recognition

Once we obtain the embedding matrices P and Q, the visual features of unseen classes can be easily synthesised from their semantic attributes:

$$\mathcal{X}_u = \mathcal{A}_u PQ. \tag{19}$$

It is noticeable that for image-level attributes, $\mathcal{X}_u$ contains as many instances as the test set. The zero-shot recognition task now becomes a typical classification problem. Thus, any existing supervised classifier, *e.g.* SVM, can be applied. For class-level, only a prototype feature of each class is synthesised. Either few-shot learning techniques or the simplest Nearest Neighbour (NN) algorithm can be adopted. Since we focus on the quality of the synthesised features, we simply use NN and SVM for image-level tasks and NN for class-level tasks.

Table 1. Comparison with State-of-the-art methods.

Methods	Feature	Animals with Attributes	Caltech-UCSD Birds	aPascal&aYahoo	SUN Attribute
DAP [24]	$\mathcal{L}$	40.50	-	18.12	52.50
ALE [2]	$\mathcal{L}$	43.50	18.00	-	-
Jayaraman and Grauman [18]	$\mathcal{L}$	43.01± 0.07	-	26.02± 0.05	56.18± 0.27
Romera-Paredes and Torr [43]	$\mathcal{L}$	49.30± 0.21	-	27.27± 1.62	-
Ours+CA	$\mathcal{L}$	**53.45± 0.30**	**43.52± 0.69**	36.98± 0.62	53.46± 1.32
Ours+SVM	$\mathcal{L}$	-	40.88± 1.34	**44.21± 0.28**	**66.03± 0.74**
DAP [24]	$\mathcal{V}$	57.23	-	38.16	72.00
Akata [3]	$\mathcal{A}$	61.9	40.3	-	-
Romera-Paredes and Torr [43]	$\mathcal{V}$	75.32± 2.28	-	24.22± 2.89	82.10± 0.32
Zhang and Saligrama [54]	$\mathcal{V}+T$	76.33± 0.83	30.41± 0.20	46.23± 0.53	82.50± 1.32
Zhang and Saligrama [55]	$\mathcal{V}+T$	80.46± 0.53	42.11± 0.55	50.35± 2.97	83.83± 0.29
Zhang and Saligrama [56]	$\mathcal{V}+T$	**90.25 ± 8.08**	**53.30± 33.39**	**65.36± 37.29**	86.00± 14.97
Ours+CA	$\mathcal{V}$	82.12± 0.12	44.90± 0.88	42.25± 0.54	80.50± 0.75
Ours+SVM	$\mathcal{V}$	-	45.72± 1.23	53.21± 0.62	**86.50± 1.75**

$\mathcal{L}$: Low-level feature, $\mathcal{A}$: Deep feature using AlexNet, and $\mathcal{V}$: VGG-19, CA: class-level attributes. T: transductive.

Algorithm 1: Unseen Visual Data Synthesis (UVDS)

Input: Training set $\{\mathcal{X}_s,\ \mathcal{A}_s,\ \mathcal{Y}_s\}$, k for k-nn graph
Output: $P,\ Q$ and $\mathcal{V}$

Initialise $Q = \mathrm{I}$, $\mathcal{V} = \mathcal{X}_s$ and $P = (\mathcal{A}_s^\top \mathcal{A}_s)^{-1}\mathcal{A}_s^\top \mathcal{V}$, where $\mathrm{I} \in \mathbb{R}^{D\times D}$ is the identity matrix.
Repeat
  $\mathcal{V}$**-Step:** Fix P, Q and update $\mathcal{V}$ using Eq. (12).
  Q**-Step:** Fix P, $\mathcal{V}$ and update $\mathcal{Q}$ by following steps:
  for $t = 1 : \max$ iterations **do**
    Compute the gradient Δ_t using Eq. (16);
    Compute the the skew-symmetric matrix Φ_t;
    Compute the Cayley matrix H_t using Eq. (15);
    Compute the Q_{t+1} using Eq. (14);
    if convergence, **break**;
  end
  P**-Step:** Fix $\mathcal{V}$, Q and update $\mathcal{P}$ using Eq. (18).
Until convergence
Return $f_{UVDS}(x) = xPQ$

4. Experiments

Settings We evaluate our method on four benchmark datasets and strictly follow the published seen/unseen splits. For AwA [23] and aPY [11], we follow the standard 40/10 and 20/12 splits like most of existing methods. For CUB, we follow [2] to use the 150/50 setting. For SUN, we use the simple 707/10 setting as reported in [18, 43, 54]. Methods under different settings [40, 13, 7, 9], or using other various semantic information [36, 52, 1, 3] are not compared with.

Semantic Attributes Existing attributes are divided into image-level and class-level. On CUB, aPY, and SUN datasets, image-level attributes are provided. Our approach can synthesise the visual features for all unseen instances. We compute class-level attributes by averaging the image-level attributes for each class. For the AwA dataset, only class-level attributes are provided.

Visual Features For low-level visual features, we use those provided by the four datasets [23, 11, 37, 46]. For deep learning features, we adopt CNN features released by[54] for the four datasets using the VGG-19 model.

Implementation Parameters Half of the data in each class in the training sets are used as the validation set. We use 10-fold cross-validation to obtain the optimal hyper-parameters λ and β. k is fixed to 10 for the k-nn graph.

4.1. Comparison with the State-of-the-art methods

Table 1 summarises our comparison to the published results of state-of-the-art methods. The hyphens indicate that the compared methods were not tested on the corresponding datasets in the original papers. In the first section, all of the compared methods were tested using conventional low-level features. In the second section, deep learning features are employed. For all of the four datasets, we first evaluate our method using class-level attributes (CA). In this scenario, each unseen class gains a synthesised visual feature prototype from the class attribute signature. The unseen test images are predicted by the NN classification using these prototypes. When image-level attributes are available in CUB, aPY, and SUN, we further conduct experiments using SVM classifiers. The visual feature vector of each unseen image is synthesised by the proposed UVDS and then fed to train SVM models. During the test, visual features that are extracted from the unseen images are fed to the trained SVM to get the prediction. Our method can steadily outperform the state-of-the-art methods on conventional ZSL scenarios. Our results also exceed two of the results base on transductive settings [56, 54], which sufficiently support our synthesised visual features are highly discriminative. While deep learning features can boost the

Table 2. Comparison with baseline methods.

Scenario	Dataset	CUB				SUN				aPY			
	Test Domain	Seen		Unseen		Seen		Unseen		Seen		Unseen	
Prototype-based	**Baseline**	CA	MF	CA	MF	CA	MF	CA	MF	CA	MF	CA	MF
	Linear Regression	66.82	64.34	27.28	30.31	88.85	89.12	63.00	64.50	52.42	55.35	17.96	21.63
	GR-only ($\beta = 0$)	65.79	65.53	38.82	40.42	89.67	88.41	75.50	76.00	59.38	57.75	25.75	28.86
	DR-only ($\lambda = 0$)	66.32	67.98	37.75	40.64	90.31	89.85	74.00	77.50	57.96	58.32	30.28	32.46
	Ours	**67.47**	**68.43**	**44.90**	**44.90**	**92.32**	**89.88**	**80.50**	**78.50**	**62.75**	**64.88**	**42.25**	**41.97**
Sample-based	**Baseline**	NN	SVM	NN	SVM	NN	SVM	NN	SVM	NN	SVM	NN	SVM
	Linear Regression	**64.57**	67.44	22.36	26.57	**90.79**	92.27	72.50	77.00	43.75	44.42	13.48	15.96
	GR-only ($\beta = 0$)	61.38	66.88	32.65	38.58	88.42	91.91	74.50	80.00	53.34	57.08	22.74	25.59
	DR-only ($\lambda = 0$)	62.44	68.94	36.93	42.24	88.34	90.47	78.00	84.00	**55.05**	53.41	23.68	24.22
	Ours	63.78	**70.32**	**39.82**	**45.72**	89.85	**93.23**	**78.50**	**86.50**	54.35	**69.75**	**38.49**	**53.21**

CA: Class-level attributes, MF: Mean of synthesised features, GR: Graph regularisation, and DR: Diffusion regularisation. Best results are in bold.

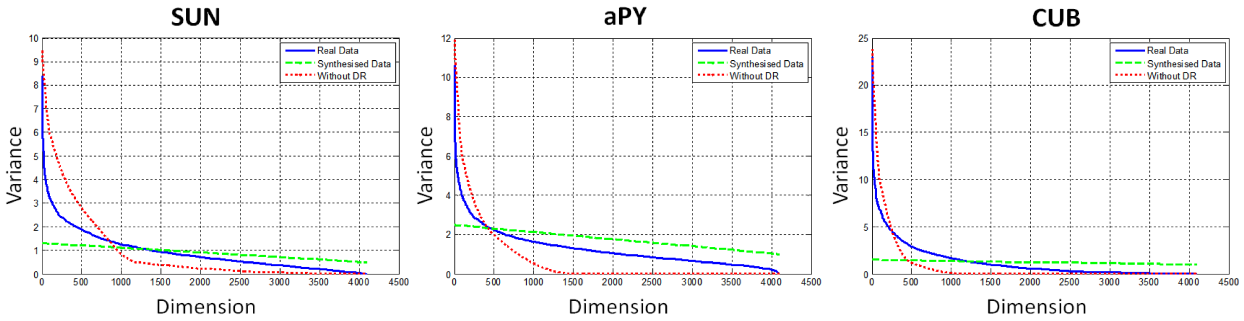

Figure 3. Normalised variances of the synthesised data w.r.t. dimensions. Variance of each dimension is sorted in descending order. We make a comparison between the synthesised data variances 'with' (green) and 'without' (red) diffusion regularisation. The variances of real data (blue) are computed from real unseen data as references.

performance, our method can also achieve acceptable results with low-level features. In most cases, using SVM can further improve the recognition rates, especially when the class-level attributes are noisy, *e.g*. on aPY and SUN. However, if the class-level attributes are more precise, *e.g*. CUB, the class-level NN classifier can be better than SVM.

4.2. Detailed Evaluations

Baseline methods To understand the effect of each term in Eq. (9), we compare our method to several baseline methods in Table 2. Since AwA only provides class-level attributes, the following experiments are conducted on CUB, SUN, and aPY only. The first method is simply *Linear Regression* that we solve Eq. (1) and synthesise prototypes of unseen classes using Eq. (2). The second and third methods are denoted as *Graph-Regularisation (GR) only* ($\beta = 0$) and *Diffusion-Regularisation (DR) only* ($\lambda = 0$). For the training bias problem, we use the validation set to test the methods on seen classes. We also investigate ZSL under both class-level and image-level attributes scenarios. The first scenario is *prototype-based*, *i.e*. each unseen class gains only one visual prototype. We compare two possible ways to obtain the class-level visual prototype: 1) we compute the mean of image-level attributes in each class and use the averaged class-level attributes (CA) to synthesise one visual prototype for each class; 2) we first synthesise the visual features from the image-level attributes and use the mean of the features (MF) as the class prototype. During the test, we use NN classification to predict the label for the test image. The second scenario is *sample-based*, *i.e*. each unseen image has one unique attribute description. In this scenario, we fully synthesise all of the visual features of unseen classes and use them as training examples. We show how an advanced classifier, *e.g*. SVM, can further boost the performance.

In summary, our method can effectively prevent the training bias whereas the linear regression without regularisation suffers from 30% performance degradation in average from seen to unseen. DR is complementary to GR and can further boost the performance. There is no significant difference between the CA and MF scenarios. Therefore, our proposed method can be reliably applied to both image-level and class-level attributes. Another advantage is that the synthesised visual data can be fed to typical supervised classifiers to achieve better performance, which can be supported by the results using SVM.

Further Discussion There are two more questions: (1) what are the outcomes of the diffusion regularisation? (2) What kind of visual features are synthesised? In Fig. 3, we show the variance of each dimension of the synthesised data. The variances are sorted in descending order. We compare with the real unseen data and the synthesised data without diffusion regularisation ($\beta = 0$). Note that, in the synthesised data without DR (red), most variances are con-

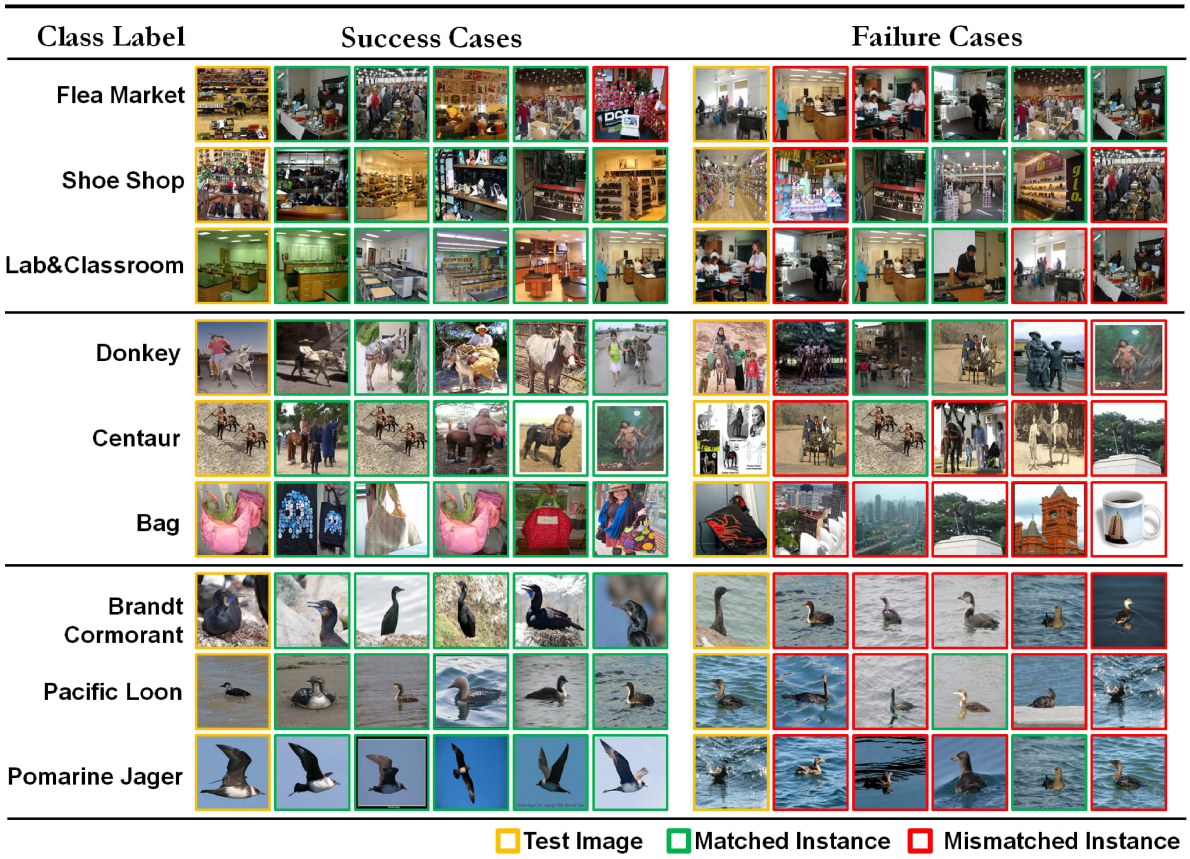

Figure 4. Success and Failure cases of nearest neighbour matching. The query visual feature is synthesised from its attribute description. We find top-5 nearest neighbours of the query feature from the real instances. It is a match if the nearest instance and the test image have the same label.

centrated in a few dimensions (roughly 1000, 1500, and 500 on SUN, aPY, and CUB) while most of the remaining dimensions gain very low variances. In comparison, the variances of our proposed synthesised data (green) and real data are more informative. Furthermore, thanks to the DR, the variances in our proposed data are more balanced than real data, *i.e*. each of the dimension gains the equal amount of information. Such quantitative evidence explains the success of our proposed method in ZSL recognition.

Finally, we provide some qualitative results of our method. We use the synthesised features as queries and retrieve real images from the unseen datasets. In Fig. 4, we show some success cases that most of the top-5 results are with the same class labels. Particularly, the third result of *Bag* is the same paired image of the attributes that are used to synthesise the data. Such results demonstrate that the synthesised data is close to the samples from the same class in the feature space. On the contrary, we also provide some failure cases that the top-1 retrieval result is not with the same class label. Some of them are due to the ambiguity of the semantic meaning, *e.g*. the *flea market* has many similar attributes to the *shoe shop*. Some other cases, *e.g*. the CUB dataset, the real data of the birds are not distinctive to the other classes. Therefore, the NN-based retrieval gives a mixture of true-positives and false-positives. Such failures due to the ambiguity of the visual feature are not common cases. We can still achieve 45.72% overall recognition rate on the CUB dataset.

5. Conclusion

In this paper, we proposed a novel algorithm that synthesises visual data for unseen classes using semantic attributes. From the experiments, we can see that directly embedding using regression-based models can lead to low recognition rates owing to three main problems, in terms of structural difference, training bias, and variance decay. In correspondence, we introduced a latent structure-preserving space with the diffusion regularisation. Our approach outperformed the state-of-the-art methods on all of the four benchmark datasets. For future work, a worthy attempt is to substitute the semantic attributes by automatic word vectors that are driven from the text. In this way, the cost of synthesising data can be further reduced.

References

[1] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele. Multi-cue zero-shot learning with strong supervision. In *CVPR*, 2016.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.

[3] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

[4] Z. Al-Halah, T. Gehrig, and R. Stiefelhagen. Learning semantic attributes via a common latent space. In *VISAPP*, 2014.

[5] Z. Al-Halah and R. Stiefelhagen. How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes. In *WACV*, 2015.

[6] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1548–1560, 2011.

[7] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016.

[8] S. Changpinyo, W.-L. Chao, and F. Sha. Predicting visual exemplars of unseen classes for zero-shot learning. *arXiv preprint arXiv:1605.08151*, 2016.

[9] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. *arXiv preprint arXiv:1605.04253*, 2016.

[10] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *CVPR*, 2013.

[11] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[12] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*. 2014.

[13] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(11):2332–2345, 2015.

[14] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, 2015.

[15] C. Gan, M. Lin, Y. Yang, G. de Melo, and A. G. Hauptmann. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In *AAAI*, 2016.

[16] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*, 2011.

[17] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang. Learning hypergraph-regularized attribute predictors. In *CVPR*, 2015.

[18] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *NIPS*, 2014.

[19] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, 2014.

[20] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010.

[21] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[23] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[24] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3):453–465, 2014.

[25] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI*, 2008.

[26] J. Lei Ba, K. Swersky, S. Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, 2015.

[27] X. Li and Y. Guo. Max-margin zero-shot learning for multi-class classification. In *AISTATS*, 2015.

[28] X. Li, Y. Guo, and D. Schuurmans. Semi-supervised zero-shot classification with label representation learning. In *ICCV*, 2015.

[29] K. Liang, H. Chang, S. Shan, and X. Chen. A unified multiplicative framework for attribute learning. In *ICCV*, 2015.

[30] Y. Long, L. Liu, and L. Shao. Attribute embedding with visual-semantic ambiguity removal for zero-shot learning. In *BMVC*, 2016.

[31] Y. Long, L. Liu, and L. Shao. Towards fine-grained open zero-shot learning: Inferring unseen visual features from attributes. In *WACV*, 2017.

[32] Y. Long and L. Shao. Describing unseen classes by exemplars: Zero-shot learning using grouped simile ensemble. In *WACV*, 2017.

[33] T. Mensink, E. Gavves, and C. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014.

[34] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *ECCV*. 2012.

[35] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[36] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[37] G. Patterson, C. Xu, H. Su, and J. Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.

[38] J. Qin, L. Liu, L. Shao, F. Shen, B. Ni, J. Chen, and Y. Wang. Zero-shot action recognition with error-correcting output codes. In *CVPR*, 2017.

[39] J. Qin, Y. Wang, L. Liu, J. Chen, and L. Shao. Beyond semantic attributes: Discrete latent attributes learning for zero-shot recognition. *IEEE Signal Processing Letters (SPL)*, 2016.

[40] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *NIPS*, 2013.

[41] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011.

[42] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010.

[43] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015.

[44] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Augmented attribute representations. In *ECCV*. 2012.

[45] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[46] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[47] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.

[48] S. Wright and J. Nocedal. Numerical optimization. *Springer Science*, 35:67–68, 1999.

[49] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016.

[50] X. Xu, F. Shen, Y. Yang, D. Zhang, T. Shen, and J. Song. Matrix tri-factorization with manifold regularizations for zero-shot learning. In *CVPR*, 2017.

[51] Y. Yang and T. M. Hospedales. A unified perspective on multi-domain and multi-task learning. *ICLR*, 2015.

[52] F. Yu, L. Cao, R. Feris, J. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013.

[53] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *ECCV*. 2010.

[54] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015.

[55] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016.

[56] Z. Zhang and V. Saligrama. Zero-shot recognition via structured prediction. In *ECCV*, 2016.