This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

Zero-shot Learning Using Synthesised Unseen Visual Data with Diffusion Regularisation

Yang Long, Li Liu, Fumin Shen, Ling Shao, *Senior Member, IEEE,* and Xuelong Li, *Fellow, IEEE*

Abstract— Sufficient training examples are the fundamental requirement for most of the learning tasks. However, collecting well-labelled training examples is costly. Inspired by Zero-shot Learning (ZSL) that can make use of visual attributes or natural language semantics as an intermediate level clue to associate low-level features with high-level classes, in a novel extension of this idea, we aim to synthesise training data for novel classes using only semantic attributes. Despite the simplicity of this idea, there are several challenges. Firstly, how to prevent the synthesised data from over-fitting to training classes? Secondly, how to guarantee the synthesised data is discriminative for ZSL tasks? Thirdly, we observe that only a few dimensions of the learnt features gain high variances whereas most of the remaining dimensions are not informative. Thus, the question is how to make the concentrated information *diffuse* to most of the dimensions of synthesised data. To address the above issues, we propose a novel embedding algorithm named *Unseen Visual Data Synthesis* (UVDS) that projects semantic features to the high-dimensional visual feature space. Two main techniques are introduced in our proposed algorithm. (1) We introduce a latent embedding space which aims to reconcile the structural difference between the visual and semantic spaces, meanwhile preserve the local structure. (2) We propose a novel *Diffusion Regularisation* (DR) that explicitly forces the variances to diffuse over most dimensions of the synthesised data. By an orthogonal rotation (more precisely, an orthogonal transformation), DR can remove the redundant correlated attributes and further alleviate the over-fitting problem. On four benchmark datasets, we demonstrate the benefit of using synthesised unseen data for zero-shot learning. Extensive experimental results suggest that our proposed approach significantly outperforms the state-of-the-art methods.

Index Terms—Zero-shot learning, Data synthesis, Diffusion regularisation, Visual-semantic embedding, Object recognition.

1 Introduction

Classification is arguably one of the most fundamental tasks in the machine learning field. Most of the conventional classification frameworks rely on a sufficient number of training samples to build reliable models. However, such a condition is unattainable in many real world situations. First, obtaining annotations for 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. The second challenge is the explosive increase of concepts. 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. Another difficulty is collecting instances for rare classes. 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.

As a feasible solution, *Zero-shot Learning* (ZSL) aims to leverage a closed-set of semantic models that can generalise to an ever growing set of new classes [1], [2], [3], [4]. Since semantic information can be obtained through human knowledge, new classes can be dynamically created without collecting any new visual data. The common paradigm is inspired by that humans can identify new things by just knowing the conceptual descriptions since we could associate the concepts to our previous knowledge. Following such an idea, the first step of ZSL is to train a prediction model that can map visual data to a semantic representation. Hereafter, new categories can be recognised by only knowing their semantic descriptions. Existing ZSL studies fall into two main streams: prediction models and semantic representation designs. The former stream develops advanced models that aim to predict human knowledge accurately from visual data, *e.g.* the probabilistic model DAP and IAP [2], [5], [6]. More recent studies take advantages of an embedding approach as middle layers between low-level features and class labels [4], [7], [8], [9], [10], [11]. Besides, some novel works study how to directly construct classifiers for unseen classes [12], [13], [14]. The latter stream focuses on how to effectively represent human knowledge that can generalise to novel classes, such as human-nameable attributes [2], [15], [16], [17], [18], word vectors [3], [19], textual descriptions [20], and class similarities [21], [22].

The methods mentioned above share a common shortage that the training visual examples cannot be expanded while the semantic information is increasing and new classes are added. Since new concepts are ever growing, it is inevitable

- *Y. Long is with the OpenLab, School of Computing Science, The University of Newcastle, Newcastle upon Tune, NE4 5TG, UK, email: yang.long@ieee.org.*
- *L. Liu is with the Malong Technologies Co., Ltd., No. 610, Yinke Building, ZhongGuangCun, Haidian District, Beijing, 100080, China, email: li.liu@malongtech.cn.*
- *L. Shao is with the School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK, email: ling.shao@ieee.org.*
- *F. Shen is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, e-mail: fumin.shen@gmail.com.*
- *X. Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), the State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China.*

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

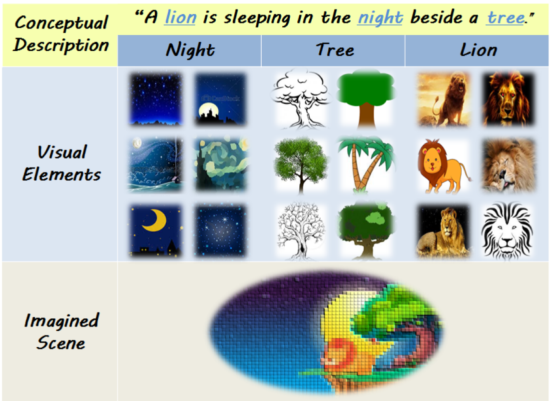


Fig. 1. Given a conceptual description, human can imagine the outline of the scene by combining previous seen visual elements.

to collect training data for new semantic models. In this paper, we propose to synthesise training data for unseen classes. Our idea is inspired by the ability of imagination of human beings. As illustrated in Fig. 1, given a semantic description, humans can associate familiar visual elements and then imagine an approximate scene. It is worth noting that our method differs from image synthesis in [1] since the synthesised images from semantics can hardly cover the large variation of visual appearances. Instead, we synthesise discriminative low-level features to train supervised classifiers for ZSL. Such an approach provides a direct interface between ZSL tasks and conventional supervised classifiers. Moreover, it enables the information mutually flow between high-level concepts and low-level visual features. In this way, the training set can be expanded to as large as the semantic representations.

Despite the simplicity of the idea, we confront two main technical issues. The first is the *visual-semantic discrepancy*. Since the visual and semantic features differ in the extracted sources and means, the data distributions of the two data spaces can be significantly discrepant. Two close points in one space can be far away in the other space. For example, as reported in [23], the same attribute 'HasTail' may have a great difference between the visual appearances of 'Zebra' and 'Pig'. However, rather than concerning the 'domain-shift problem' for the recognition task in [23], instead, we hope the model can effectively capture the semantic-visual correlation so that the synthesised visual data can preserve the intrinsic structure as close as the real data.

The second issue is the *variance decay*. Due to that the number of visual feature dimensions is usually much larger than that of semantic representations, the learnt projection is prone to be imbalanced, *i.e.* the variances of the projected dimensions vary severely [24]. As shown in Fig. 6, comparing to the real data, we observe that the synthesised data using linear projection suffers from remarkable variance decay. The variances of most of the projected dimensions are extremely low, which indicates they gain little information. Such projections can lead to degraded performance owing to the great number of redundant dimensions. Therefore, the challenge is how to make the information diffuse to most of the dimensions of the synthesised data with a balanced projection. To the best of our knowledge, this issue has not been identified in previous ZSL literature.

To address the above issues, we propose a novel embedding algorithm named *Unseen Visual Data Synthesis* (UVDS) that projects semantic features to the high-dimensional visual feature space. In particular, for the first issue, we introduce a latent embedding space to reconcile the structural difference between the visual and semantic spaces. We use a *dual-graph* (GR) to preserve the local structure of both visual and semantic spaces. For the second problem, we propose a novel *Diffusion Regularisation* (DR) that explicitly makes the information diffuse to all dimensions of the synthesised data. Specifically, we use the variances as the measurement to force information to diffuse over the dimensions of the synthesised data. We prove that such a scheme is equivalent to finding an orthogonal rotation transformation. Also, we discover an elegant form of such an orthogonal rotation using the $\ell_{2,1}$ norm regularisation with efficient solutions.

In addition to the above two problems, the synthesised data should also be discriminative for the ZSL task. A direct regression model tends to learn the principal components between the two spaces that lead to high bias towards the training classes. We view this as an *over-fitting problem*, *i.e.* the trained model can achieve high performance on the synthesised data of seen classes but will dramatically degrade on the synthesised unseen data. We empirically show that the above GR and DR can mitigate the over-fitting problem in a complementary manner: DR does not harm the local structure preservation but instead benefits the data synthesis by eliminating the redundant correlations in the semantic space through the orthogonal rotation. The main contributions of this paper are summarised below:

- An intuitive framework that enables us to synthesise unseen data from the given semantic attributes. The synthesised data can be straightforwardly fed to typical classifiers and lead to the state-of-the-art performance on four benchmark datasets.
- 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 data structural preservation and data reconstruction.

The rest of the paper is organised as follows. We review existing ZSL methods and related work in Section 2. The proposed algorithm is described in detail in Section 3. The experimental results are demonstrated in Section 4. Finally, we make a conclusion and discuss possible future works in Section 5.

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 [1], [2], [3] 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 [2]. Such a model utilises semantic attributes as intermediate clues. A test sample is classified

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

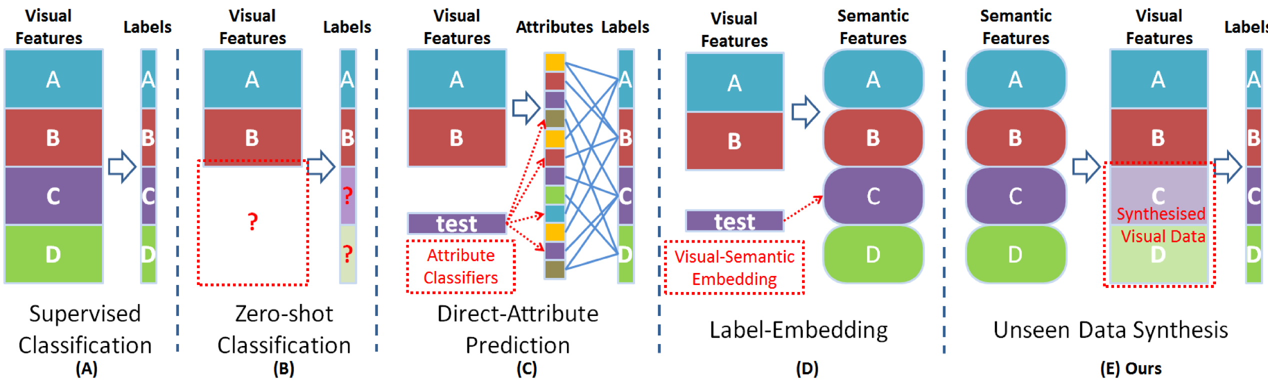


Fig. 2. Comparison of supervised and zero-shot classifications and existing ZSL frameworks. (A) a typical supervised classification: the training samples and labels are in pairs; (B) a zero-shot learning problem: without training samples, the classes *C* and *D* cannot be predicted; (C) Direct-Attribute 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 use semantic embedding to synthesise unseen visual data.

by each attribute classifier in turn, and the class label is predicted by probabilistic estimation. Admitting the merit of DAP, there are some concerns about its deficiencies. [18] 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 [25].

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 [7], [10], [19], [26], [27], [28]. Since the attribute vectors are regarded as whole semantic representations, attributes are used for transductive ZSL settings [11], [23], [29], [30], [31], [32]. However, these methods involve the data of unseen classes to learn the model, which to some extent breaches the strict ZSL settings. Recent work [4], [33] combines the embedding-inferring procedure into a unified framework and empirically demonstrates better performance. The closest related work is [34], which takes one-step further to synthesise classifiers for unseen classes. Our method is also different from DS-SJE [35], in terms of learning objective, regularisation, and the potential applications. DS-SJE seeks to learn a compatibility function for both images and texts, whereas our objective function aims to reconstruct the visual features from semantic attributes. Also, our method learns with GR and DR that are not considered in DS-SJE. The inferred visual features can be applied to conventional supervised classifiers, which differs our method from other previous work.

Our method takes the advantages of semantic embedding. However, our purpose is entirely different from existing work. As discussed earlier, owing to the fact that the semantic information is ever growing, it is inevitable to collect visual training data for newly added concepts. Since it is easier to obtain semantic information from the Internet, our method can expand the number of visual feature vectors to as many as the semantic instances.

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, [14], [20] propose to estimate the semantic relatedness of the novel classes from the text. [13], [36], [36] 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 [37], [38], [39].

Although various side information is studied, attribute-based ZSL methods still gain the most popularity. One reason is that attributes often give prominent classification performance [21], [22], [40], [41], [42]. For another reason, attribute representation is a compact way that can further describe an image by concrete words that are human-understandable [16], [43], [44], [45]. Various types of attributes are proposed to enrich applicable tasks and improve the performance, such as relative attributes [15], class-similarity attributes [21], and augmented attributes [17].

In this paper, we evaluate our method using attributes and Word2vectors. Since our proposed framework is embedding-based, it can easily exploit most of existing semantic side information.

Structure-Preserving Projection: Structure-preserving projection is well-studied in unsupervised learning [46]. A spectral graph is constructed to preserve the original data structure. [47] extends such an idea to multi-view classification to preserve the intrinsic data structures of multiple modalities. The most common approach is to use local neighbourhood graphs for each view independently [31]. [48] generalises a single graph to a multi-graph with random walks between the connections. The graph-based approach is adopted in [23] for transductive ZSL. They estimate the pairwise similarity between training data and unlabelled unseen data using heterogeneous hyper-graphs.

In contrast to these methods, we strictly follow the ZSL setting that excludes data of unseen classes from the training set. Such a setting increases the difficulty since the visual structure of unseen classes can be distinctive from the given semantic data structure. As a solution, we propose to insert a latent embedding space to reconcile the data structure discrepancy between the visual and semantic spaces. A dual-graph is then constructed to find a balanced structure between the two spaces.

Data Rotation for Information Diffusion: Because information diffusion has not aroused attentions in the ZSL field, we discuss related work in a broader context. Data Rotation aims to find a balanced projection that makes the informa-

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

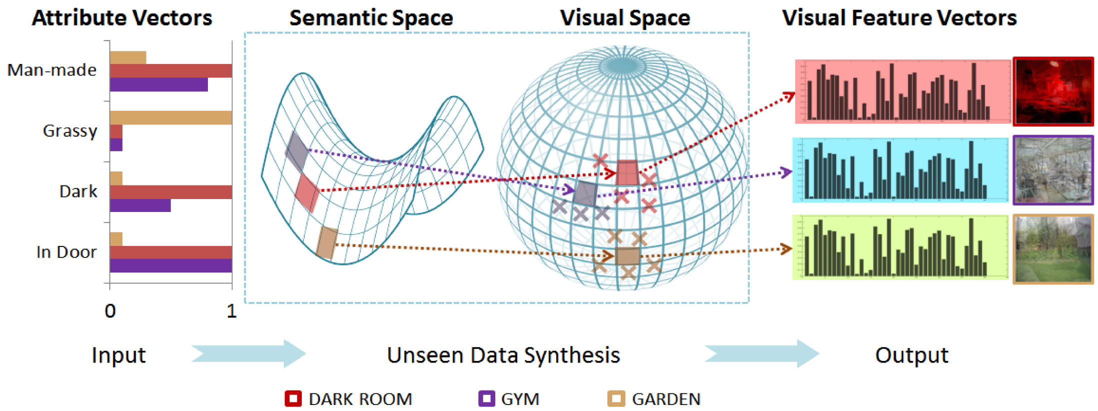


Fig. 3. An illustration of our framework of unseen data synthesis. Unseen classes are represented by semantic attributes as inputs. We train a model that maps the semantic space to the visual data space to synthesise training data for these unseen classes. The crosses in the visual spaces denote test feature points.

tion diffuse to all dimensions of the synthesised data. Such an issue is initially concerned with unsupervised learning methods [49], [50], [51] since imbalanced projection can lead to inferior retrieval performance. In [51], data rotation is adopted to minimise the quantisation error. [49] achieves information diffusion by minimising the reconstruction error of the covariance matrix. [52] uses perfectly diffused data as referencing source to find the rotation so that the projected data can also be well diffused.

We share the consideration of these previous works, yet our proposed method is entirely different from them. Firstly, our ZSL task is fully supervised. We aim to synthesise visual features rather than finding an optimal subspace of original features. Secondly, none of the previous works utilise variance as measurement and explicitly control the information diffusion. In our experiments, we demonstrate that the synthesised data can achieve more balanced dimensions even comparing to the real data. The improved performance can also prove the effectiveness of our method.

3 Approach

ZSL tasks generally involve three steps: training, inference, and test. Some of previous methods may combine inference with training or test. In our framework, the training only requires data of seen classes. The attributes of unseen classes are required at the inference stage to synthesis visual features. Finally, we use the synthesised features for ZSL classification.

3.1 Preliminaries

The training set contains 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. During the test, the given attributes can be either *category-level* or *instance-level*. In our framework, we aim to cope with both of the scenarios using a unified paradigm. Given $\hat{N}$ pairs of unseen instances with semantic attributes from $\hat{C}$ discrete categories: $(\hat{a}_1, \hat{y}_1), ..., (\hat{a}_{\hat{N}}, \hat{y}_{\hat{N}}) \subseteq \mathcal{A}_u \times \mathcal{Y}_u$, where $\mathcal{Y}_u \cap \mathcal{Y}_s = \varnothing$, $\mathcal{A}_u = [a_{\hat{n}m}] \in \mathbb{R}^{\hat{N} \times M}$, the goal of zero-shot learning is to learn a classifier, $f : \mathcal{X}_u \to \mathcal{Y}_u$, 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 categories by the given semantic attributes. Specifically, we learn an embedding function on the training set $f' : \mathcal{A}_s \to \mathcal{X}_s$. 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, it can directly estimate the probability distribution of the unseen categories. It is straightforward to employ conventional supervised classifiers, *e.g.* SVM, to predict the labels of unseen classes $f_{\text{SVM}} : \mathcal{X}_u \to \mathcal{Y}_u$.

3.2 Unseen Visual Data Synthesis

Traditional ZSL methods minimise the single classification error of each attribute. Due to that, the attributes are separately learnt, as aforementioned, such a framework highly depends on the quality of the designed attributes. Recently, there is a new scheme that addresses ZSL by an embedding approach [7]. In particular, an objective function is learnt to minimise the multi-class error simultaneously and consider the relationship between different attributes. A typical multi-attributes classifier can be learnt by the following problem:

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

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 $\Omega(P) = \|P\|_F^2$. During the test, an unseen instance can be directly mapped to the attribute space by $\hat{a} = \hat{x}P$.

However, due to the fact that P is learnt using only the training data, the inferred attributes $\hat{a}$ are prone to be biased towards the 'seen' attributes $\mathcal{A}_s$. Inspired by the idea that a human can imagine the visual appearance of an unseen object through given semantic descriptions, we propose to synthesise visual features by reversely learning a mapping function from the semantic space to the visual feature space:

$$\min_P \mathcal{L}(\mathcal{A}_s P, \mathcal{X}_s) + \lambda \Omega(P). \tag{2}$$

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

The loss term accounts for the reconstruction error between the semantic input and the visual output; whereas the regularisation ensures the discrimination to unseen classes. Such a framework provides a direct mapping to the visual space without computing a pseudo-inverse matrix and therefore avoids information loss. Before the test, it is straightforward to infer the visual features of unseen classes using their class attributes:

$$\mathcal{X}_u = \mathcal{A}_u P. \tag{3}$$

Visual-Semantic Structure Preservation In spite of the simplicity of the above framework, several problems are worth noting. Firstly, 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. Secondly, such a regression-based framework does not discover the intrinsic topological structure. As a result, the synthesised data may gain an entirely different feature distribution to the original visual features. 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:

$$J = \|\mathcal{X}_s - \mathcal{V}Q\|_F^2 + \|\mathcal{V} - \mathcal{A}_s P\|_F^2 + \lambda \Omega_1(\mathcal{V}), \tag{4}$$

where $\|.\|_F$ is the Frobenius norm of a matrix, which estimates the Euclidean distance between two matrices. The latent-embedding space $\mathcal{V}$ is decomposed from $\mathcal{X}$ and $\mathcal{A}$ is then decomposed from $\mathcal{V}$, where $Q = [q_{d'd}] \in \mathbb{R}^{D \times D}$ and $P = [p_{md}] \in \mathbb{R}^{M \times D}$ are two projection matrices. Ω_1 is a *dual-graph* that is introduced next.

In detail, we take the *Local Invariance* [46] 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 mentioned earlier, the attributes for ZSL tasks can be instance-level or category-level. In particular, for *instance-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 $W_{\mathcal{A}}$), $w_{ij} = 1$ if and only if g_i and g_j are connected by an edge, otherwise, $w_{ij} = 0$. As a result, we can separately compute the two weight matrices $W_{\mathcal{X}}$ and $W_{\mathcal{A}}$. It is noteworthy that, for *category-level attributes*, $W_{\mathcal{A}}$ is computed in a slightly different way. Every vertex in the same category 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 category c, where n_c is the size of category 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. $W_{ij} = \frac{1}{2}(W_{\mathcal{X}_{ij}} + W_{\mathcal{A}_{ij}})$. The resulted regularisation is:

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

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 Another fundamental problem is *Redundant projections*. Compared to the compact attributes, the variance of visual data is usually larger and more informative. However, 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. 6, most of the information (variance) is contained in a few projections. As a result, the remaining dimensions of the synthesised data experience 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 [53]. Therefore, we modify the rotating matrix Q in Eq. (4). In this paper, we consider an orthogonal rotation, *i.e.* $QQ^T = I$, since it is easy to show that $Tr(Q^T P^T \mathcal{A}^T \mathcal{A} P Q) = Tr(P^T \mathcal{A}^T \mathcal{A} P)$. This is an intuitive idea that we can rotate the whole feature space by changing the coordinates through the orthogonal transformation. In this way, the high-variance can diffuse to lower-variance dimensions without changing the overall variance. Such a property is reported in [54] that is known as ITQ, which aims to learn similarity-preserving binary codes. By solving an orthogonal Procrustes problem, the whole feature space is rotated according to the coordinates without changing the structure. Although the values of each dimension are changed, the overall data structure in the semantic $\mathcal{A}$ is completely preserved. Next, we show how the rotation can control variance diffusion.

From Eq. (4), 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. The attribute data $\mathcal{A}_s$ is centralised, i.e. $\sum_{n=1}^{N} a_n = \mathbf{0}$. The original variance Γ of $\mathcal{V}$ is $\Gamma = N\sigma_d$, 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:

$$\begin{aligned} \Gamma &= N \sum_{d=1}^{D} \sigma_d = \sum_{d=1}^{D} \sum_{n=1}^{N} v_{nd}^2 = \|\mathcal{V}\|_F^2 = Tr(\mathcal{V}\mathcal{V}^T) \\ &= Tr(\mathcal{V}QQ^T\mathcal{V}^T) = \|\mathcal{V}Q\|_F^2 \\ &= \sum_{d=1}^{D} \sum_{n=1}^{N} x_{nd}^2 = N \sum_{d=1}^{D} \sigma'_d = \Gamma'. \end{aligned} \tag{6}$$

We hope the overall variance Γ tends to equally diffuse to all of the learnt dimensions in order to recover the real

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

data distribution of $\mathcal{X}$. In other words, the variance of diffused standard deviations Π in the synthesised data should be small, *i.e.* $\Pi = 1/D\sum_{d=1}^{D}(\pi_d - \bar{\pi})^2$, where $\pi_d = \sqrt{\sigma'_d}$ and $\bar{\pi}$ is the mean of all standard deviations. According to the above Eq. (6), we have ϵ, *i.e.* $\sum_{d=1}^{D}\pi_d^2 = \sum_{d=1}^{D}\sigma'_d = \sum_{d=1}^{D}\sigma_d = \epsilon$. Since the sum of standard deviations does not change after rotation Eq. (6), minimising the variance of diffused standard deviations can make high variances diffuse to dimensions with low variances. Next, we show how to minimise Π in our learning framework to find the orthogonal rotation. We first rewrite Π:

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

The first term $\frac{\epsilon}{D}$ of the above equation is a constant. Thus, the problem of minimising Π is equivalent to maximise the sum of diffused standard deviations in the bracket of the second term of Eq. (7). Furthermore, such a maximisation can be converted into the problem of optimising the orthogonal rotation:

$$\begin{aligned}\sum_{d=1}^{D}\pi_d &= \sum_{d=1}^{D}\sqrt{\sigma'_d} = \sum_{d=1}^{D}\sqrt{\sum_{n=1}^{N}x_{nd}^2/N} \\ &= \frac{1}{\sqrt{N}}\|\mathcal{X}^T\|_{2,1} = \frac{1}{\sqrt{N}}\|Q^T\mathcal{V}^T\|_{2,1}, \end{aligned} \quad (8)$$

where $\|.\|_{2,1}$ is the $\ell_{2,1}$ norm of a matrix. According to Eq. (7) and Eq. (8), we can simply maximise $\|Q^T\mathcal{V}^T\|_{2,1}$ to maximise Π with the optimal Q for the purpose of information diffusion. Finally, we can combine the diffusion regularisation with Eq. (4) 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:

$$\begin{aligned}\min_{P,Q,\mathcal{V}} \quad & J = \|\mathcal{X}_s - \mathcal{V}Q\|_F^2 + \|\mathcal{V} - \mathcal{A}_sP\|_F^2 + \lambda Tr(\mathcal{V}^TL\mathcal{V}) \\ & -\beta\|Q^T\mathcal{V}^T\|_{2,1}, \quad s.t.\ QQ^T = I. \end{aligned} \quad (9)$$

3.3 Optimisation Strategy

The key of our optimisation is to find a proper solution for the latent-embedding space $\mathcal{V}$. From the above Eq. (9), it can be seen that $\mathcal{V}$ simultaneously accounts for the reconstruction error, structure preservation, and diffusion regularisation. However, 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. In this paper, we propose an iterative scheme by using the alternating optimisation to obtain the local optimal solution. Specifically, we iteratively update $\mathcal{V}$, Q, and P in an alternate manner. In this way, the optimisation becomes analytic and tractable for each variable with the associated sub-problem. It is noted that some variables are first heuristically initialised before our proposed optimisation. Specifically, we initialise $Q = I$ and $\mathcal{V} = \mathcal{X}_s$. Such an initialisation equals to start from the simple problem in Eq. (2). The initialisation of P can be obtained via $P = (\mathcal{A}_s^T\mathcal{A}_s)^{-1}\mathcal{A}_s^T\mathcal{V}$. The whole alternate procedure of the proposed UVDS is listed as follows.

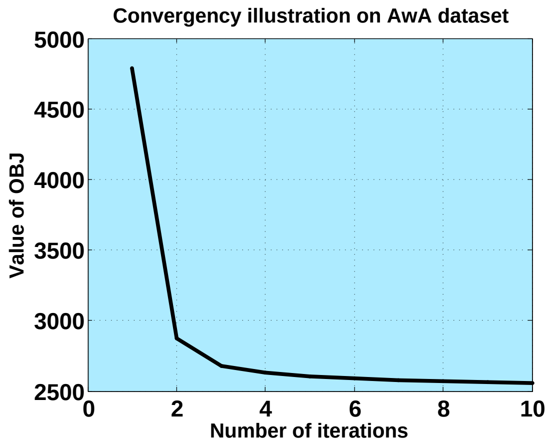


Fig. 4. Objective function convergence on the AwA dataset.

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

$$\begin{aligned}\min_{\mathcal{V}}\|\mathcal{X}_s - \mathcal{V}Q\|_F^2 + \|\mathcal{V} - \mathcal{A}_sP\|_F^2 + \lambda Tr(\mathcal{V}^TL\mathcal{V}) \\ - \beta\|Q^T\mathcal{V}^T\|_{2,1} \end{aligned} \quad (10)$$

The minimal $\mathcal{V}$ can be obtained by setting the partial derivative of Eq. (10) to zero and we have

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

where $E = diag(e_1, \dots, e_d, \dots, e_D) \in \mathbb{R}^{D\times D}$ 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

$$\mathcal{V}(2QQ^T + 2\alpha I + \beta QEQ^T) + (2\lambda L)\mathcal{V} - (XQ^T + 2\mathcal{A}P) = 0, \quad (12)$$

which is a typical Sylvester equation so that $\mathcal{V}$ can be efficiently solved by the *lyap()* function in the MATLAB toolbox.

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

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

Since we need to solve Q with the orthogonality constraint in Eq. (13), we adopt the gradient flow in [55] which is an iterative scheme that can optimise orthogonal problems with a feasible solution. Such an iterative scheme can minimise Eq. (13) until it arrives at a stationary 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*:

$$Q_{t+1} = H_tQ_t, \quad (14)$$

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

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

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1
Key statistics of the four datasets.

Dataset	# of attributes	Attribute Type	Annotation Level	# of Seen Classes	# of unseen classes	# of total images
AwA	85	Both	per class	40	10	30475
CUB	312	Binary	Both	150	50	11788
aPY	64	Binary	per image	20	12	15339
SUN	102	Continues	per image	707	10	14340

Fig. 5. Some random image and attribute examples of the 4 datasets.

where I is the identity matrix, $\Phi_t = \Delta Q_t^T - Q_t \Delta^T$ is the skew-symmetric matrix, τ is an approximate minimiser satisfying Armijo-Wolfe conditions [56] and Δ is the partial derivative of Eq. (13) with respect to Q as

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

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. Generally, we set $t = 30$ for Q updating in the Q-step. A similar proof of the updating procedure with the orthogonality constraint can be observed in [55].

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:

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

Note that $(\mathcal{A}_s^T \mathcal{A}_s)^{-1}$ is not always full rank, especially in which all of the instances share the class-level attributes. Therefore, we use Moore-Penrose pseudo inverse of matrix instead. We have so far described our optimisation of each step for Eq. (9) in detail. As mentioned above, to obtain a local optimal solution, we adopt an alternate optimisation scheme, in which we repeat t times to solve $\mathcal{V}$ sub-problem, Q sub-problem and P sub-problem in sequence. In our experiments, ten iterations in overall alternate optimisation are proved to be enough for convergence as shown in Fig. 4. The proposed UVDS approach is depicted in Algorithm. 1.

3.4 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:

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

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 $P = (\mathcal{A}_s^T \mathcal{A}_s)^{-1} \mathcal{A}_s^T \mathcal{V}$, 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 $f_{UVDS}(x) = xPQ$

It is noticeable that for instance-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 to learn a mapping function: $\mathcal{Y}_u = f_{svm}(\mathcal{X}_u)$.

For category-level, only a prototype feature of each category is synthesised. Either few-shot learning techniques or the simplest Nearest Neighbour (NN) algorithm can be adopted: $\hat{y} = \arg\min_i \|\hat{x} - \hat{a}_i PQ\|_2^2$, where $\hat{x}$ is a test image, $\hat{a}_i$ is the class-level attribute vector of the i-th unseen class, and $\hat{y}$ is the final prediction. Since we focus on the quality of the synthesised features, we simply use NN and SVM for instance-level tasks and NN for category-level tasks.

4 Experiments

We provide a comprehensive comparison with both classic and recent state-of-the-art methods on four benchmark

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 2
Comparison with state-of-the-art methods.

Method	Feature	AI	EP	Animals with Attributes	Caltech-UCSD Birds	aPascal&aYahoo	SUN Attribtue
Lampert et al. [2]	V	CA	PC	57.23	-	38.16	72.00
Romera-Paredes and Torr [4]	V	CA	PC	75.32±2.28	-	24.22±2.89	**82.10±0.32**
GAN	V	CA	PC	62.40±0.85	40.52±0.95	24.28±0.44	68.85±0.72
Ours	V	CA	PC	**82.12±0.12**	**44.90±0.88**	**42.25±0.54**	80.50±0.75
Lampert *et al.* [2]	V	W2V	PC	42.82±0.81	24.52±0.68	24.52±0.28	65.28±0.57
Akata *et al.* [39]	V	W2V	PC	56.25±0.74	30.28±0.56	29.28±0.86	70.70±0.65
Romera-Paredes and Torr [4]	V	W2V	PC	58.29±0.58	28.47±0.76	32.67±0.58	72.65±0.78
Zhang and Saligrama [22]	V	W2V	PC	57.49±1.82	29.68±0.84	34.95±1.47	74.19±0.83
GAN	V	W2V	PC	48.34±0.69	25.33±0.82	27.48±0.74	68.58±0.89
Ours	V	W2V	PC	**62.88±0.76**	**32.14±0.47**	**35.82±0.45**	**76.98±0.46**
Lampert et al. [2]	G	CA	PC	57.19±0.62	37.41±0.63	35.26±0.95	38.94±0.90
Romera-Paredes and Torr [4]	G	CA	PC	74.72±0.81	55.10±0.48	34.48±0.96	57.36±0.29
Akata *et al.* [39]	G	CA	PC	76.72±0.12	55.34±0.27	32.02±0.15	57.18±0.80
Ours	G	CA	PC	**80.28±0.14**	**57.52±0.54**	**38.65±0.28**	**60.82±0.91**
Zhang and Saligrama [22]	V	CA	PI	76.33±0.83	30.41±0.20	46.23±0.53	82.50±1.32
Zhang and Saligrama [42]	V	CA	PI	80.46±0.53	42.11±0.55	50.35±2.97	83.83±0.29
Ours	V	CA	PI	**85.28±0.49**	**46.48±0.82**	**52.48±0.79**	**87.50±0.75**
Romera-Paredes and Torr [4]	V	IA	PC	-	42.82±0.73	39.69±0.45	79.85±1.02
Ours+SVM	V	IA	PC	-	**45.72±1.23**	**53.21±0.62**	**86.50±1.75**

Feature: VGG-19 (V) and GoogLeNet-1K (G); Auxiliary Information (AI): Class-level Attributes (CA), Instance-level Attributes (IA), and Word2Vec (W2V); Evaluation Protocol (EP): Per-class accuracy (PC) and Per-image accuracy (PI).

datasets: Animals with Attributes (AwA) [2], aPascal & aYahoo (aPY) [16], Caltech-UCSD Birds-200-2011 (CUB) [57], and SUN Attribute (SUN) [58]. Key characteristics of these datasets are summarised in Table 1. Furthermore, we verify the statements we made in this paper by comparing to a variety of baselines.

4.1 Setup

Settings ZSL is a complicated system that involves multiple key steps. Existing methods differ in the experimental setup, in terms of visual feature extraction, semantic auxiliary information, modelling, and evaluation protocols. We strictly follow published seen/unseen splits. For AwA [2] and aPY [16], we follow the standard 40/10 and 20/12 splits like most of existing methods. For CUB, we follow [7] to use the 150/50 setting. For SUN, we use the simple 707/10 setting as reported in [4], [22], [25]. Methods under different settings [23], [29], [34], or using other various semantic information [15], [21], [39], [45] are not compared with. For fair and comprehensive comparison with existing state-of-the-art methods, we divided our main comparisons into five groups. The details are introduced as follows.

Visual Features The adopted visual features of existing methods mainly differ in deep models. Since most of previous methods are based on the 4096-dimensional CNN features extracted by [22] for the four datasets using the "Image-net-vgg-verydeep-19" model [59], most of our evaluations are based on the same model. In order to see the effect of different visual features, we also conduct experiments using the GoogLeNet-1K feature and compared to the results evaluated by [60].

Auxiliary Information The attribute annotation levels of the four datasets are different. In CUB, aPY, and SUN, each image is annotated by a unique attribute signature. In AwA, all of the images within one class share the same attribute signature. We compute such class-level attributes (CA) for aPY and SUN by averaging the image-level attributes for each class. Yet, it is impossible to get the image-level attribute descriptions for AwA. The resultant class-level attributes for the four datasets are in real numbers, whereas the image-level attribute (IA) signatures of CUB, aPY, and SUN are binary. We also implement evaluations using Word2Vec features [61] as the auxiliary information. Each class name is encoded into a vector as the class-level semantic representation.

Evaluation Protocols The first comprehensive ZSL comparison [22] adopts the Per-image accuracy (PI) as the evaluation criteria. Namely, they measure whether the Top-1 prediction is the correct class label for each image. However, it is recently argued that such a criterion may encourage biased prediction on densely populated classes [60]. Therefore, most of our evaluation is based on the Per-class accuracy (PC) which is the mean value of all of the test classes. Without loss of generality, we also calculate the corresponding PI for comparison.

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.

4.2 Comparison with the State-of-the-art methods

Table 2 summarises our comparison to the published results of state-of-the-art methods on the benchmark datasets. The hyphens indicate that the compared methods were not tested on the corresponding datasets in the original papers. The comparisons are mainly divided into five sections. In the first section, all of the compared methods were tested using human-annotated attributes. In the second section, W2V class-label embeddings [61] are employed as the class-level semantic features. We implement the state-of-the-art methods using their published codes. Section three demonstrates the effect of input visual features. We alter the VGG-19 model by GoogLeNet-1K and keep the rest the same as that in section one. We also calculate the Per-image accuracy of our method in section four for comparison. For all of the above four sections, we evaluate our method using class-level attributes. In this scenario, each unseen class gains a synthesised visual feature prototype from the class attribute signature. The test unseen images are predicted by the NN classification using these prototypes.

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3
Detailed analysis of key aspects of the proposed method.

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.

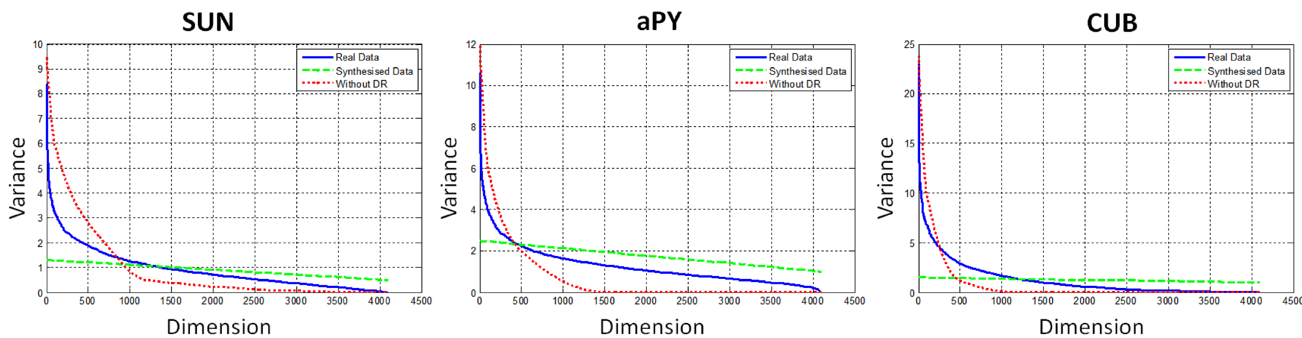


Fig. 6. Normalised variances of the synthesised data *w.r.t.* dimensions. Variance of each dimension is sorted in descending order. We make a comparison between the synthesised data variances 'with' (green) and 'without' (red) diffusion regularisation. The variances of real data (blue) are computed from real unseen data as references.

In section five, when image-level attributes (IA) 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 real unseen image are fed to the trained SVM to get the prediction. A similar setup can be found in [4], which assumes each training attribute signature as a class in its own right. The IA-based ZSL has its unique applications. For instance, the unseen class may be unknown for humans and we do not have a CA signature for it. Alternatively, we may know the class. But it could be difficult to summarise a CA signature, *e.g.* restaurants may have distinctive styles and attributes. In both of the cases, we could describe those 'hard' unseen classes by exemplars with sparse attributes. This does not violates the spirit of ZSL since the test images are still unseen for the machine.

Our method outperforms most of the published results on the four datasets. Note that on aPY, using synthesised instance-level features with SVM provides a significant performance boost. The evidence can also be found on SUN. This is because that on aPY and SUN, the class-level attributes may not well conclude the features of all of the instances in each class, *e.g.* different style of room. Thus, the individualised synthesised visual features with the SVM classifier can make significant improvement. However, using finer-defined attributes, such as on AwA and CUB, CA can also results in similar performance to that of using instance-level features with SVM. In the second section, the performance based on W2V degraded severely due to the coarse description of the class labels. Our method achieves the best results on all of the datasets. The success can be considered from two aspects. Firstly, although the W2V feature space is heterogeneous to the visual space, our GR can adjust the synthesised features to mitigate such a difference. Secondly, from the Fig. (8) can be seen, the synthesised features are more discriminative than the real visual features, which can withstand some performance degradation. In section three, different results can be seen using GoogLeNet-1K features. It can be observed that CUB gains the most significant benefit. [2], [4] also achieve improvements on aPY. But in general, we can conclude that the VGG-19 feature can better fit most of the approaches although GoogLeNet-1K has its own advantage on the CUB dataset.

In addition, we also consider the recent Generative Adversarial Network (GAN) [62] as a comparable baseline, which can also synthesise unseen visual features using attributes. In order to preserve the discrimination for different classes, we adopt a similar framework as that in the InfoGan [63]. As shown in Fig. 7, the difference to InfoGan is that the generative net (GN) is conditioned on the class attributes (or W2V) instead of the one-hot class vector so as to achieve the inter-class transfer for ZSL. The implementation details can be found in [63]. During the inference, we input the class-level semantic representations (attributes or W2V) of unseen classes to generate unseen visual features. We then use the generated samples to train SVM to classify unseen instances at the test phase. As shown in Table 2, due to the sizes of the datasets are not very large, the results of GAN are inferior to conventional methods. Therefore, how to apply GAN on ZSL task requires further investigation.

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

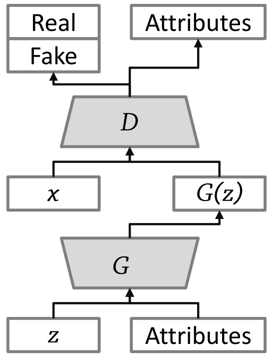


Fig. 7. Framework of the compared GAN model.

4.3 Detailed Evaluations

To further understand the success of our UVDS algorithm and verify our statements that are made in this paper, we compare to variations of our methods as baselines under different scenarios. Since AwA only provides class-level attributes, we conduct the remaining experiments on CUB, SUN, and aPY.

Baseline methods The primary purpose of our comparison is to understand the effect of each term in Eq. (9). The first baseline method is simply *Linear Regression* ($\beta = 0, \lambda = 0$) that we solve Eq. (2) and synthesise prototypes of unseen classes using Eq. (3). The second and third methods are denoted as *Graph-Regularisation (GR) only* ($\beta = 0$) and *Diffusion-Regularisation (DR) only* ($\lambda = 0$). In this pair of comparison, we aim to study the characters of each term and how they contribute to the overall performance. For both of the methods, we use the same cross-validation as our proposed method to tune λ and β. In order to discuss the over-fitting problem, we also use the validation set as test for seen classes.

Since existing zero-shot learning methods differ in the annotation level of the semantic attributes, we also investigate how such scenarios can affect the performance. The first scenario is *prototype-based*, *i.e.* each unseen class gains only one visual prototype. There are two possible ways to obtain the class-level prototype: (1) we can 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 can 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 can 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. We summarise the results of our self-comparison in Table 3. Based on the outcomes, we can verify the following statements that are made in this paper.

Generalisation to Unseen Data From Table 3, we can see that linear regression can achieve acceptable performance when tested on seen classes. On two datasets, CUB and SUN, the synthesised visual features by the linear regression method are even better than the comparative methods using simple NN classifiers. However, a remarkable drop of recognition rates (32.21% on CUB and 18.29% on SUN) can be found when tested on unseen classes. In average, the performance degradation of unseen class recognition using the linear regression method is about 20%. This is a typical over-fitting problem since we tune the best parameters on the seen set but the trained model cannot well generalise to unseen classes. In comparison, the proposed method can achieve the best performance in most of the situations. Meanwhile, the proposed method can also smoothly generalise to unseen classes. In the case of the SUN dataset, the recognition rate of unseen classes using the SVM classifier is only 3.38% lower than the MF scenario on seen classes (89.88%). The other two baseline methods GR-only and DR-only achieve similar performances on the seen classes and once is higher than the proposed method (55.05% of DR-only on aPY using NN classifier). On unseen classes, the two baseline methods are all better than linear regression without regularisation but lower than the proposed method using both regularisations. Such results suggest that the proposed method can significantly eliminate the bias to the seen training data.

Effect of Regularisations In Table 3, we can see both of the regularisations can significantly boost the performance comparing to the linear regression method. In most cases, the DR-only method is slightly better than the GR-only method. This suggests the importance of the balanced features. Also, we observe the performance of using both of the regularisations is always better than using one of them on the unseen set. To further understand the relationships between GR and DR, in Fig. 9, we fix $\lambda = 0.001$ and show the performance varies with β. In turn, we fix $\beta = 0.1$ to see the trend of performance with respect to λ. It can be seen that in most cases, adding the other regularisation can benefit the performance (compared to the case of $\beta = 0$ or $\lambda = 0$ at the beginning of the curve). The exception is only when the other regularisation is over-weighted, *e.g.* $\lambda = 10$. Such a result indicates the two regularisations are not redundant but well complementary to each other.

Class-level attributes or Mean of Features In the case that only class-level attributes are provided, there is no other optional scenario. However, if the provided attributes are image-level, we could use the mean of the attributes for each class to compute prototypes (CA). Alternatively, we could synthesise visual feature for each image first and then compute the mean of the features for each class (MF). When comparing these two scenarios in Table 3, interestingly, the performance difference between the two methods is insignificant. The results of MF on the aPY dataset tend to be better than those of CA, whereas, on the SUN dataset, the results of CA are slightly higher than those of MF. We assume the potential reason is due to the quality of the attribute annotations since the attributes in aPY are reported not very reliable [25]. Such results also show the positive side of our method that we could confidently use the class-level attributes even though there are no image-level attributes available, *e.g.* the AwA dataset.

Advance of using SVM One encouraging reason for synthesising unseen data is to be used for training supervised classifiers. In Table 3, the performance of using NN classification under the sample-based scenario is somewhat worse than that under the prototype-based scenarios (CA and MF). After using SVM classifiers, the performance is remark-

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

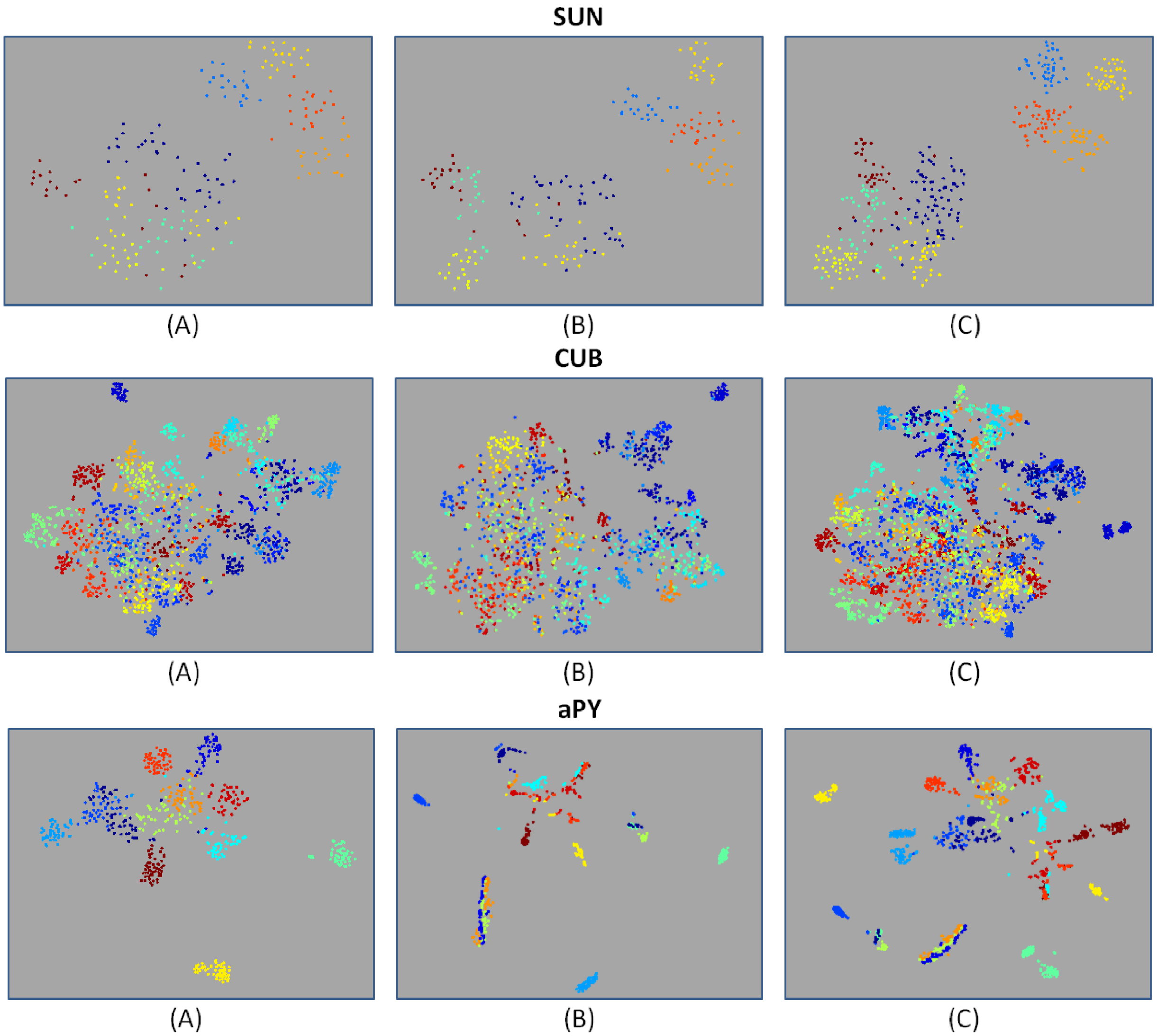


Fig. 8. T-SNE of the real and synthesised visual features of unseen classes: (A) real visual features; (B) synthesised visual features; (C) Since t-SNE of different data is not aligned, we also show the distribution of mixed real and synthesised visual features.

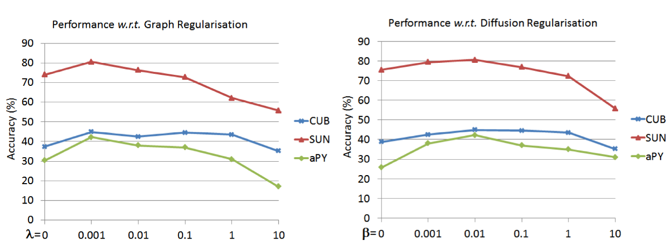


Fig. 9. The performance with respect to the Graph regularisation and Diffusion regularisation. The results are under the scenario of CA and using NN classifier.

ably boosted and achieves the highest ZSL recognition rate among all of the scenarios. This is a promising result that substantially demonstrates the advantage of using synthesised training data for advanced classifiers.

Efficiency Our method is very efficient at the test phase since it only requires to search among several unseen prototypes or to make prediction using SVM. The experiment is conducted in Matlab 2016b environment with Core i7-6820 Processors. As shown in Table 4, the averaged computation time for training is also practical for both conventional and large-scale datasets.

TABLE 4
Computation Time on Each Dataset.

AwA	CUB	aPY	SUN	ImageNet
1.56×10^3	1.47×10^3	0.84×10^3	1.03×10^3	1.44×10^4

4.4 Further Discussions

This section mainly investigates three key aspects of the proposed method: (1) what are the outcomes of the diffusion regularisation? (2) what kind of visual features are synthesised? and (3) how is the performance on other ZSL scenarios, *e.g.* Generalised and large-scale ZSL? We answer these questions based on the following experimental analysis.

In Fig. 6, 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 ($\beta = 0$). It is noticeable that, in the synthesised data without DR, most variances are concentrated in a few dimensions (roughly

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

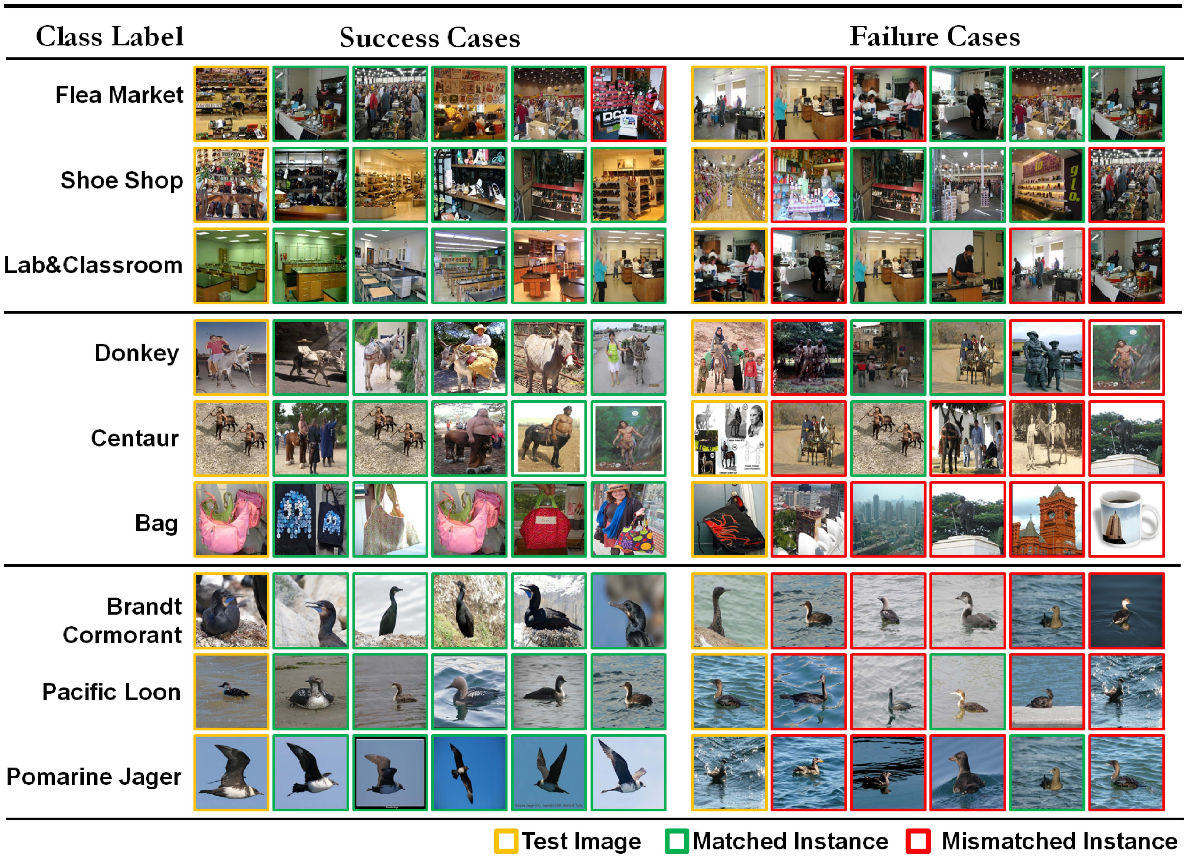


Fig. 10. Success and Failure cases of nearest neighbour matching. The query visual feature is synthesised from its attribute description. We find top-5 nearest neighbours of the query feature from the real instances. It is a match if the nearest instance and the test image have the same label.

TABLE 5
Comparison with published results on the ImageNet Dataset.

	Hierarchy		Most Populated			Least Populated			AH
Method	2H	3H	500	1K	5K	500	1K	5k	20K
ConSE [9]	7.63	2.18	12.33	8.31	3.22	3.53.	2.69	1.05	0.95
DEVISE [8]	5.25	1.29	10.36	6.68	1.94	4.23	2.86	0.78	0.49
SJE [39]	5.31	1.33	9.88	6.53	1.99	4.93	2.93	0.78	0.52
ESZSL [4]	6.35	1.51	11.91	7.69	2.34	4.50	3.23	0.94	0.62
SYNC [34]	9.26	2.29	15.83	10.75	3.42	5.83	3.52	1.26	0.96
Ours	**10.15**	**2.47**	**15.96**	**11.28**	**4.12**	**6.06**	**3.74**	**1.49**	**1.02**

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 and real data are more informative. Furthermore, thanks to the DR, the variances in our proposed data are even more balanced than real data. In other words, each of the dimension gains the equal amount of information. Such quantitative evidence explains the success of our proposed method in the ZSL recognition task.

In Fig.(10), 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. 10, 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.

Fig.(8) shows the distribution of the synthesised (B) and real features (A) of the unseen classes using t-SNE. On SUN and CUB, after mixing both of the features together (C), most classes are discriminative, which means the synthesised features capture the same distribution of the real unseen classes. On aPY, however, the synthesised features look more discriminative than the real features. This can be ascribed to the orthogonal constraint that makes the structure-preserving of the graph constraint sacrifice for the performance. After mixing the real and synthesised features together, intraclass points can be easily discriminated, which supports the effectiveness of the synthesised features.

Finally, we evaluate our method under Generalise ZSL (GZSL) scenarios (Table 6) and on large scale datasets (Table 5) using the class-level attributes (CA). For the former one, we strictly follow the four protocols proposed in [64]. For consistency, we use the GoogLeNet-1K feature as [39], [60],

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 6
Comparison with published results on GZSL.

	AwA				CUB			
Method	U-U	S-S	U-T	S-T	U-U	S-S	U-T	S-T
DAP [2]	51.1	78.5	2.4	77.9	38.8	56.0	4.0	55.1
IAP [2]	56.3	77.3	1.7	76.8	36.5	69.6	1.0	69.4
ConSE [9]	63.7	76.9	9.5	75.9	35.8	70.5	1.8	69.9
SynC [64]	73.4	81.0	0.4	**81.0**	54.4	73.0	13.2	72.0
Ours	**80.3**	**86.7**	**15.3**	79.5	**57.52**	**75.4**	**23.8**	**76.5**

U: Unseen classes; S: Seen classes; T=S+U.

[64]. The attributes are the same as conventional ZSL. *U-U* is the conventional unseen-to-unseen ZSL; *S-S* is the traditional supervised classification; *U-T* and *S-T* are two types of GZSL that evaluate whether learnt unseen/seen models are confused to each other. On AwA, our method outperforms the state-of-the-art methods on three of the four scenarios. Only on *S-T* our result is slightly lower than that of [64]. The seen/unseen balance can be viewed as an over-fitting problem: while we sacrifice the performance on seen classes (*S-T*), the performance on GZSL on unseen classes *U-T* is significantly boosted. The evidence can also be found on CUB dataset. Although our model performs slightly worse on the seen classes, a better trade-off is achieved, which results in 6.2% performance gain on the *U-T* scenario on CUB.

For the large scale ZSL, we follow the settings of [60] on the ImageNet dataset. We extracted the same VGG-19 features as that for the four ZSL benchmarks. For class-level attributes, we use the W2V features provided by [34]. Our method consistently outperforms the published results, from which we can see the prominence synthesised features. However, there is still a large room for improvements. We argue that, for most of ZSL scenarios, the number of unseen classes should be at least smaller than that of training classes. Such inverted ZSL with significantly larger number of test classes requires reconsideration of the framework. One possible way is to incrementally synthesise unseen visual features and then fine-tune the model using both real and synthesised features like a semi-supervised learning framework.

5 Conclusion

In this paper, we proposed a novel algorithm that synthesises visual data for unseen classes using semantic attributes. The attributes are regarded as a full representation and embedded into the visual feature space. From the experiments, we can see that directly embedding using regression-based models can lead to low zero-shot recognition rates. We ascribed such direct synthesised data to three problems, in terms of imbalanced variances, over-fitting, and indiscrimination. In correspondence, we introduced a latent structure-preserving space with the diffusion regularisation as the objective function. As a result, we observed that the proposed algorithm could significantly benefit the performance on unseen class recognition. Our approach outperformed the state-of-the-art methods on all of the four benchmark datasets.

For future work, a worthy attempt is to synthesise instance-level features so that the SVM-based framework can be widely applied. For another, our qualitative experiments give positive results since we have shown the synthesised features are close to the real features in the same class. In the future, the synthesised data can be leveraged for more applications such as image retrieval or unseen image reconstruction. Also, how to address the inverted ZSL with larger number of test classes requires further investigation.

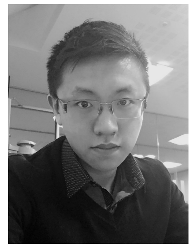
References

[1] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks." in *AAAI*, 2008.
[2] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*, 2009.
[3] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *NIPS*, 2009.
[4] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *ICML*, 2015.
[5] 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*, vol. 36, no. 3, pp. 453–465, 2014.
[6] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa, "Online incremental attribute-based zero-shot learning," in *CVPR*, 2012.
[7] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *CVPR*, 2013.
[8] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *NIPS*, 2013.
[9] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," in *ICLR*, 2014.
[10] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, "Zero-shot object recognition by semantic manifold distance," in *CVPR*, 2015.
[11] X. Li, Y. Guo, and D. Schuurmans, "Semi-supervised zero-shot classification with label representation learning," in *ICCV*, 2015.
[12] X. Wang and Q. Ji, "A unified probabilistic approach modeling relationships between attributes and objects," in *ICCV*, 2013.
[13] M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a classifier: Zero-shot learning using purely textual descriptions," in *CVPR*, 2013.
[14] T. Mensink, E. Gavves, and C. Snoek, "Costa: Co-occurrence statistics for zero-shot classification," in *CVPR*, 2014.
[15] D. Parikh and K. Grauman, "Relative attributes," in *ICCV*, 2011.
[16] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.
[17] V. Sharmanska, N. Quadrianto, and C. H. Lampert, "Augmented attribute representations," in *ECCV*, 2012.
[18] D. Jayaraman, F. Sha, and K. Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," in *CVPR*, 2014.
[19] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *NIPS*, 2013.
[20] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, "What helps where–and why? semantic relatedness for knowledge transfer," in *CVPR*, 2010.
[21] F. Yu, L. Cao, R. Feris, J. Smith, and S.-F. Chang, "Designing category-level attributes for discriminative visual recognition," in *CVPR*, 2013.
[22] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *ICCV*, 2015.
[23] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 11, pp. 2332–2345, 2015.
[24] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for scalable image retrieval," in *CVPR*, 2010.
[25] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," in *NIPS*, 2014.
[26] 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.
[27] X. Li and Y. Guo, "Max-margin zero-shot learning for multi-class classification." in *AISTATS*, 2015.
[28] Z. Al-Halah, T. Gehrig, and R. Stiefelhagen, "Learning semantic attributes via a common latent space," in *VISAPP*, 2014.
[29] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer learning in a transductive setting," in *NIPS*, 2013.

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

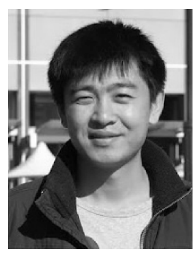
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2762295, IEEE Transactions on Pattern Analysis and Machine Intelligence

[30] Y. Yang and T. M. Hospedales, "A unified perspective on multi-domain and multi-task learning," *ICLR*, 2015.
[31] 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.
[32] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Unsupervised domain adaptation for zero-shot learning," in *ICCV*, 2015.
[33] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *CVPR*, 2016.
[34] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *CVPR*, 2016.
[35] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *CVPR*, 2016.
[36] J. Lei Ba, K. Swersky, S. Fidler *et al.*, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *ICCV*, 2015.
[37] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *CVPR*, 2011.
[38] Z. Al-Halah and R. Stiefelhagen, "How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes," in *WACV*, 2015.
[39] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *CVPR*, 2015.
[40] X. Yu and Y. Aloimonos, "Attribute-based transfer learning for object categorization with zero/one training example," in *ECCV*, 2010.
[41] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang, "Learning hypergraph-regularized attribute predictors," in *CVPR*, 2015.
[42] Z. Zhang and V. Saligrame, "Zero-shot learning via joint latent similarity embedding," in *CVPR*, 2016.
[43] K. Liang, H. Chang, S. Shan, and X. Chen, "A unified multiplicative framework for attribute learning," in *ICCV*, 2015.
[44] 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.
[45] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele, "Multi-cue zero-shot learning with strong supervision," in *CVPR*, 2016.
[46] 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*, vol. 33, no. 8, pp. 1548–1560, 2011.
[47] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *NIPS*, 2004.
[48] D. Zhou and C. J. Burges, "Spectral clustering and transductive learning with multiple views," in *ICML*, 2007.
[49] W. Kong and W.-J. Li, "Isotropic hashing," in *NIPS*, 2012.
[50] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *ICML*, 2011.
[51] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *CVPR*, 2011.
[52] B. Xu, J. Bu, Y. Lin, C. Chen, X. He, and D. Cai, "Harmonious hashing." in *IJCAI*, 2013.
[53] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010.
[54] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *CVPR*, 2011.
[55] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Mathematical Programming*, vol. 142, no. 1-2, pp. 397–434, 2013.
[56] S. Wright and J. Nocedal, "Numerical optimization," *Springer Science*, vol. 35, pp. 67–68, 1999.
[57] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.
[58] G. Patterson, C. Xu, H. Su, and J. Hays, "The sun attribute database: Beyond categories for deeper scene understanding," *International Journal of Computer Vision*, vol. 108, no. 1-2, pp. 59–81, 2014.
[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint:1409.1556*, 2014.
[61] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.
[60] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning-the good, the bad and the ugly," *CVPR*, 2017.
[62] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.
[63] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NIPS*, 2016.
[64] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *ECCV*, 2016.

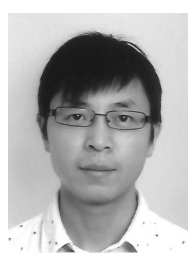


Yang Long is an research assistant with the OpenLab, the School of Computing Science, The University of Newcastle. He received the M.Sc. degree from the Department of Electronic and Electrical Engineering, The University of Sheffield, UK, in 2014. He is currently a Ph.D. student in the same department. His research interests include image classification, deep feature learning, probabilistic modelling for zero-shot learning, and large-scale ontology design.

Li Liu is with the Malong Technologies Co., Ltd. He received the B.Eng. degree in electronic information engineering from Xian Jiaotong University, Xian, China, in 2011, and the Ph.D. degree from the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, U.K., in 2014. His current research interests include computer vision, machine learning, and data mining.



Fumin Shen received his Bachelors degree in 2007 and the Ph.D. degree in 2014 from Shandong University and Nanjing University of Science and Technology, China, respectively. Now, he is an Associate Professor with University of Electronic Science and Technology of China. His major research interests include computer vision and machine learning, including face recognition, image analysis, and hashing methods.



Ling Shao (M'09-SM'10) is a professor with the School of Computing Sciences at the University of East Anglia, Norwich, UK. Previously, he was a professor (2014-2016) with Northumbria University, a senior lecturer (2009-2014) with the University of Sheffield and a senior scientist (2005- 2009) with Philips Research, The Netherlands. His research interests include computer vision, image/video processing and machine learning. He is an associate editor of IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems and several other journals. He is a Fellow of the British Computer Society and the Institution of Engineering and Technology. He is a senior member of the IEEE.

Xuelong Li (M'02–SM'07–F'12) is a Full Professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China

0162-8828 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.