SEMANTIC MANIFOLD ALIGNMENT IN VISUAL FEATURE SPACE FOR ZERO-SHOT LEARNING

Changsu Liao[1], *Li Su*[*1], *Wegang Zhang*[2], *Qingming Huang*[1]

[1]University of Chinese Academy of Sciences, China
[2]Harbin Institute of Technology, Weihai, China
liaochangsu16@mails.ucas.ac.cn, {suli, qmhuang}@ucas.ac.cn, wgzhang@hit.edu.cn

ABSTRACT

Zero-Shot Learning (ZSL) is getting more attention for its potential to solve a task without training examples, such as to recognize a category of unseen object in computer vision task. Most existing methods are suffered from hubness problem and semantic gap problem. In this paper, we propose a novel strategy based on Aligning Semantic Manifolds in Feature Space (ASMFS) to boost the performance of ZSL. Considering that the semantic representations must be predicted in the location of their corresponding visual instances, we adjust the predicted unseen semantic representations by the average of their K nearest neighbors (K-NN). The experimental results over two basic ZSL models and four public datasets demonstrate the universal enhancement performance of the proposed strategy. It significantly boosts the existing ZSL approaches with low over cost and outperforms eight state-of-the-art methods.

Index Terms— Zero-Shot Learning, Semantic Manifold, Visual Feature Space

1. INTRODUCTION

Humans can distinguish 30,000 basic object classes and many more subcategories [1]. With the knowledge humans have acquired from past experience, we can also create new categories only based on some abstract descriptions. Zero-Shot Learning emulates this human thought process, which has gained growing attention recently. For ZSL approaches, the key idea is to construct models with transfer ability to distinguish unseen object classes via an intermediate-level auxiliary information called semantic class prototype. There are typically some forms of semantic class prototypes, e.g. human-annotated attributes [2], word vectors extracted automatically from online text resources or textual descriptions. Most existing ZSL approaches learn a projection from image visual feature space χ to semantic embedding space $\mathcal{K}$ (visual-semantic mapping). In testing, unseen instances are projected into space $\mathcal{K}$ by the mapping function, and then their classes are predicted by a nearest neighbors (NN) search, i.e. they are annotated by their nearest unseen class prototypes in the embedding space.

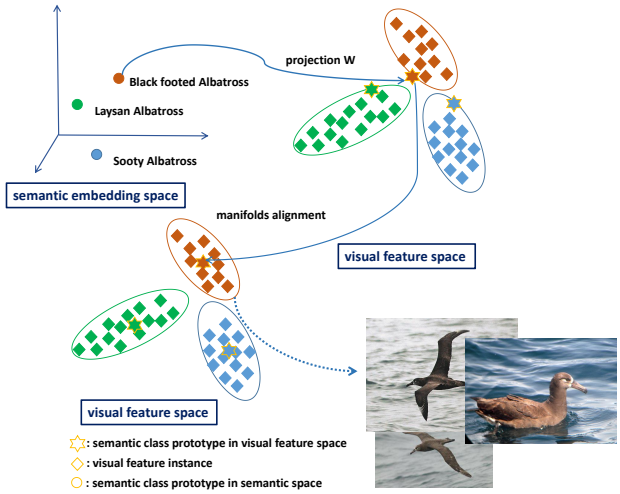


Fig. 1. Given the semantic class prototypes, the proposed method learns a semantic-visual projection. Then, manifolds of semantic class prototypes and visual feature instances can be aligned and better unseen semantic class prototypes can be obtained, which are applied for NN classification and will achieve better performance than the original method.

Many existing ZSL methods choose semantic space or construct a intermediate space as the embedding space that visual feature instances and semantic class prototypes can be compared. However, Since the NN search is applied to obtain the accuracy, and embedding space is of high dimension, the hubness problem is inevitable [3]. That means, a few number of instances in the dataset, or hubs, may occur as the nearest neighbor of many instances [4], resulting diminishing the utility of NN search. And since many state-of-the-art ZSL ap-

* Corresponding author (suli@ucas.ac.cn)

This work was supported in part by National Natural Science Foundation of China: 61472389, 61650202, 61332016, 61620106009 and U1636214 in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013.

978-1-5386-1737-3/18/$31.00

proches [5, 6] are based on ridge regression, it aggravates the problem after projecting visual feature instances into a space with lower dimansions. [4]

Another problem of ZSL is semantic gap between the visual feature and the semantic class prototypes. Specifically, the manifold of visual feature space is much distinct from that of their underlying semantic space. Thus, learning projection directly between χ and $\mathcal{K}$ can be a difficult task, and make the learned classifier become high computational complex, which is at the risk of over-fitting.

In our work, in order to tackle the two problems of ZSL simultaneously, we propose a universal strategy with low over cost that significantly boosts existing ridge regression based ZSL models. To be specific, we first utilize existing ZSL methods to learn the semantic-visual projection. Then we obtain unseen semantic class prototypes in visual feature space by the projection. In order to bridge the semantic gap between visual feature instances and semantic class prototypes, we align their manifolds to obtain better semantic class prototypes. Specifically, we adjust the acquired unseen semantic class prototypes with the average of their K-NN visual instances. Thus, the new semantic class prototypes are more representative of their corresponding visual feature cluster centers. Moreover, the NN search can be more efficient in visual feature space, since it is less affected by hubness phenomenon [4]. Fig.1 shows the conceptual diagram of the proposed framework.

To summarize, our main contributions are as follows:

- We propose a universal strategy with low over cost to tackle hubness phenomenon and semantic gap problem simultaneously. Extensive experiments demonstrate the promising enhancement performance of the proposed strategy applied in feature space.
- The proposed strategy is universal and efficient to boost existing ridge regression based ZSL models.
- The experimental results on four public datasets show that the basic ZSL model boosted by the proposed strategy outpeforms existing state-of-the-art ZSL methods.

2. RELATED WORKS

In this section, we focus on the following three aspects to compare the proposed strategy and related works.

The hubness problem The hubness problem is first defined by Radovanovic *et al.* [3] and they showed that it is a inherent problem occurs in high-dimensional space. There are some studies [4, 7] proved that the problem also reduces regression based ZSL methods performance. Among them, Shigeto *et al.* [4] proposed that ridge regression makes the hubness phenomenon worse, and suggested learning the projection from semantic space to visual feature space. Many existing models [5, 6, 8] followed this advice and achieved outstanding performances. In our work, we adjust semantic class prototypes in visual feature space to boost models that learn semantic-visual projection, thus, the proposed strategy is naturally benefited from the hubness phenomenon reduction advice.

Semantic gap Early works focus on learning a direct visual-semantic mapping [9, 10], which are suffered from the semantic gap problem. In recent years, there are a large amount of strategies to bridge the semantic gap. Most of them construct a common embedding space where visual feature instances and semantic class prototypes can be compared. Following this, Zhang *et al.* [11] proposed a probabilistic framework for learning a embedding space where visual and semantic embedding along with a class-independent similarity measure are learned simultaneously. However, most of them build mapping by inherent visual and semantic manifolds, while the proposed method aligns their manifolds by simple K-NN method to obtain better unseen semantic prototypes.

Transductive setting There is much information of unseen class instances, and methods utilizing unseen class instances during testing are called transductive ZSL. They obtain the manifold of target set to boost ZSL performances [12, 13, 10]. The proposed method is also transductive, and it only uses unseen class instances in the testing step to rebuild unseen semantic prototypes.

3. APPROACH

For ease of understanding, in this section, we first briefly introduce two simple regression based ZSL models, and then we describe how to apply the proposed strategy to boost these models.

3.1. Problem Formulation

Given a m-dimentional semantic embedding space $\mathcal{K}$ composed of b seen class semantic prototypes $S^s \in R^{b\times m}$ and d unseen class semantic prototypes $S^u \in R^{d\times m}$, and each semantic vector $s_i \in [S^s; S^u]$ corresponds to a unique object class $y_i \in [Y^s; Y^u]$. The input data matrix $X = [X^s; X^u] \in R^{n\times l}$ are composed of n feature vectors of l dimentions as its columns. And the corresponding semantic matrix of X becomes $S_x = [S_x^s; S_x^u] \in R^{n\times m}$. The task is to learn a projection (classifier) $f : X \rightarrow Y$ for the seen class Y^s that is disjoint from unseen class Y^u.

3.2. Two Basic Models

We introduce two ridge regression based methods, for which the proposed strategy will be applied.

Simple regression (SR) The simplest method to learn the semantic-visual projection is constructing a square loss with

norm-regularizer, i.e. the objective function is:

$$\min_{W} \|X^s - S_x^s W\|_F^2 + \eta\Omega(W) \quad (1)$$

where $\Omega(W)$ is a l_2-norm regularizer, and η is a trade-off parameter. $\|\cdot\|_F$ denotes Frobenius norm. The optimal projection W can be obtained directly and used in testing step.

SAE [5] is simple and it achieves a strong projection that preserve the information contained in the original visual features as much as possible , resulting in the learned classifier less susceptible to domain shift. Thus, the learning objective function is:

$$\min_{W} \|X^s - S_x^s W^T\|_F^2 + \lambda\|X^s W - S_x^s\|_F^2 \quad (2)$$

where W^T denotes the transpose of projection matrix W, and λ is a weighting coefficient that controls the importance of the first and second terms.

3.3. The Proposed Strategy

The inspiration of the proposed strategy is from clustering structure assumption that semantic prototypes must be predicted in the locations of their corresponding visual instances. The essence of the regression functions Eq (1-2) is predicting cluster centers of their corresponding visual instances. However, performance of a simple projection is suffered from the semantic gap problem. Thus, we adjust semantic class prototypes manifold to be aligned with visual feature instances manifold.

The proposed strategy is only applied in the testing step. First, we project unseen semantic class prototypes $s_i^u \in S^u$ from semantic embedding space to visual feature space to be $e_i^u \in E^u$ by the learned projection W. Then, for each $e_i^u \in E^u$, we search its K-NN unseen visual instances $NN_x^k(x_i^u)$ with cosine distance, and take place e_i^u with $\frac{1}{k}NN_x^k(x_i^u)$ as new semantic class prototypes $\hat{e_i^u} \in \hat{E^u}$. Finally, unseen instances are annotated by their nearest new unseen semantic class prototypes.

3.4. Comparison to Related Approaches

The proposed strategy is similar with DMap proposed by Li *et al.* [14], since it also utilized K-NN method to re-represent semantic class prototypes. However, they need to adjust additional seen semantic class prototypes manifold, and alternately optimize the visual-semantic mapping and the semantic representations. In our work, we only build new unseen semantic class prototypes, and since our adjustment is in visual feature space, K-NN method is naturally benefited from hubness phenomenon reduction.

Soravit *et al.* [8] proposed a method to obtain new semantic class prototypes, and it also predicts unseen instances cluster centers in visual feature space. However, they apply a non-linear function to learn the semantic-visual projection directly, which is still suffered from semantic gap problem.

We will detailedly describe the differences between the proposed strategy with the above two methods in Section 4.

Table 1. Benchmark datasets. We follow Kodirov *et al* [5] to use the same splits of all the datasets.

Dataset	instances	Attribute-D	seen/unseen class
AwA [2]	30475	85	40/10
CUB [15]	11788	312	150/50
SUN [16]	14340	102	707/10
ImNet-2 [17]	254000	1000	1000/360

4. EXPERIMENTS

4.1. Experimental setup

Datasets We evaluate the proposed method on four standard ZSL datasets, including three small-scale standard datasets: Animals with Attributes (AwA) [2], CUB-200-2011 Birds (CUB) [15], and SUN Attributes (SUN) [16]. And a large-scale dataset ILSVRC2012/ILSVRC2010(ImNet-2) [17]. As in [5], the 1000 classes of ILSVRC2012 are used as seen classes, and 360 classes, which are not included in ILSVRC2012,of ILSVRC2010 for unseen classes.Table 1 summarizes their key characteristics.

Semantic space We use attributes provided by datasets that 85, 312 and 102-dimensional attributes for AwA, CUB and SUN, respectively. And We use semantic word vector representation provided by Kodirov *et al.* [5] for the large-scale dataset. They trained a skip-gram test model on a corpus of 4.6M Wikipedia dataset to obtain the word vectors.

Visual features As for all small datasets, we use features extracted from GoogLeNet(1024 dimensions) [18] and VGG-19(4096 dimesions) [19].Those features are provided by Changpinyo *et al.* [20] and Zhang *et al.* [21], respectively. And the ImNet-2 features are provided by Kodirov *et al.* [5].

Parameter settings The proposed strategy only has one parameter, i.e. the number of K for K-NN method. And there is one parameter η of the simple regression and one parameter λ of SAE [5]. We follow Ye *et al.* [23] to obtain validation set from seen class data and select parameters on it. Then, we use the selected parameters on original source and target sets.

Evaludation metric We use multi-way classification accuracy as in previous work [5] for the three small datasets, while hit@K classification accuracy is used as in [24] for the large-scale dataset. We report hit@5 accuracy, that means, the test image is classified to be correct if it is among the top 5 labels.

Implementation details We apply the proposed strategy on two regression based methods, i.e. ASMFS-SR and

Table 2. ZSL accuracy (%) achieved by the proposed methods and state-of-the-art approaches on four benchmarks.

small-scale dataset							large-scale dataset	
methods	S	T/I	F	AwA	CUB	SUN	S	ImNet-2
SAE [5]	A	I	F_G	84.7	61.4	91.5	W	27.2
Zhang [6]	A/W	I	N_G	86.7/78.8	58.3/53.5	-/-	W	25.7
SS-Voc [22]	-	I	F_V	-	-	-	W	16.8
EXEM($SynC^{struct}$) [8]	A	I	F_G	77.2	-	-		
DSRL-LP [23]	A	T	F_V	87.2	57.1	85.4		
DMap [14]	A	T	F_G/F_V	-/85.7	**61.8**/-	-		
TSTD [13]	A	T	F_V	90.3	58.2	-/-		
TASTE [12]	A	T	F_V	89.7	54.3	-		
ASMFS-SR	A	T	F_G/F_V	88.9/89.6	57.8/51.0	93.0/83.5	W	27.7/-
ASMFS-SAE	A	T	F_G/F_V	**93.3**/91.0	61.2/57.6	**94.0**/84.5	W	**30.3**/-

Table 3. Accuracy (%) achieved by the two basic methods and the results boosted by the proposed strategy.

method	AwA	CUB	SUN	ImNet-2
SR/ASMFS-SR	76.5/88.9	52.7/57.8	87.0/93.0	26.1/27.7
SAE/ASMFS-SAE	82.2/93.3	53.2/61.2	90.5/94.0	27.2/30.3
Average	11.8	7.1	4.8	2.4

ASMFS-SAE, which are based on SR and SAE, respectively. As for ASMFS-SR we normalize X^s and X^u by zero-mean normalization, and normalize S^s and S^u by min-max normalization. While for ASMFS-SAE, since we found SAE is sensitive to data normalization, and there is no clear strategy to choose correct normalizations for different datasets, we apply code provided by Kodirov *et al.* [5] on AwA and ImNet-2, and implement code on CUB and SUN by ourselves.

4.2. Zero-Shot Classification Results

4.2.1. Compared with the state-of-the-art

We compared the proposed methods with seven existing ZSL models, including four inductive methods and four transductive methods on the small-scale datasets. As for the large-scale dataset, three methods are used to compare with the proposed methods. Most of them are published in the past two years and clearly represent the state-of-the-art. From the results in Table 2, where 'S' means semantic embedding, 'A' illustrates attribute, and 'W' means word vector. '-' means that no reported results are available, '+' means 'and', and '/' means 'or'. 'T' or 'I' denotes transductive or inductive methods. F_G, F_V denotes features extracted from GoogLeNet [18] and VGG19 [19], respectively. N_G indicates neural network based methods. ASMFS-SR and ASMFS-SAE means the simple regression method and SAE boosted by the proposed strategy, respectively. ASMFS-SAE achieves best performances on AwA, SUN and ImNet-2, and obtains outstanding result on CUB with GoogleNet features. Even simplest regression based method ASMFS-SR achieves outstanding performances on all datasets, including the large-scale dataset.

DMap [14] is similar as the proposed method, but it is much more complex than the proposed method as we mentioned in Section 3.4. And DMap reached 61.8% on CUB, while ASMFS-SAE achieves similar performance (61.2%) with GoogLeNet features even if it is simpler . On AwA, DMap utilized features extracted from VGG19. In such settings, ASMFS-SAE outperforms DMap by 5.3%. We argue that the primary cause of the limited performance of DMap is the hubness problem, since it learns a visual-semantic projection based on a ridge regression method.

EXEM($SynC^{struct}$) [8] directly predicts unseen classes centers in visual feature space so that it is a hard task to bridge the semantic-visual projection due to the semantic gap problem. Unlike it, the proposed method aligns the two manifolds to resist the problem so that it outperforms EXEM($SynC^{struct}$) by 16.1% on AwA with GoogLeNet features.

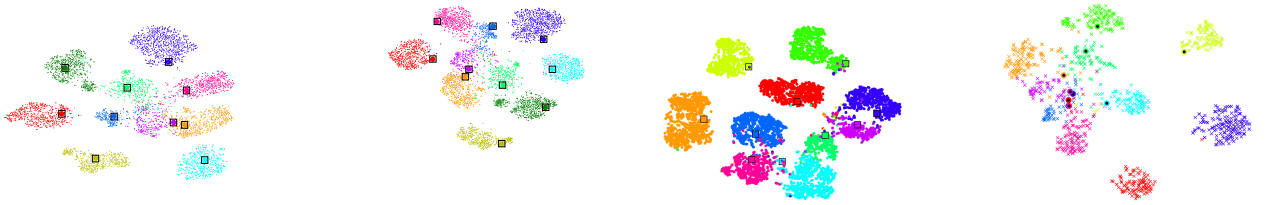
In summary, benefited from hubness phenomenon reduction and manifold alignment, even simple models boosted by the proposed strategy can obtain outstanding performances. Because of the better performance of methods with GoogleNet features, in the following we will give analysis results with GoogleNet features defaultly.

4.2.2. Enhancement ability

Table 3 illustrates the performance improvements over the two basic methods. These results demonstrate that in all cases, the proposed strategy can significantly boost their per-

Table 4. Accuracy (%) achieved by SR/ASMFS-SR in semantic embedding space (SS) and visual feature space (VS) on AwA and CUB, respectively.

method	SS			VS		
	AwA	CUB	Average	AwA	CUB	Average
SR/ASMFS-SR	60.3/66.1	35.5/37.3	3.8	76.5/88.9	52.7/57.8	8.8



(a)ASMFS-SAE (b) SAE [5] (c)Zhang[29] (d)EXEM($SynC^{struct}$)[28]

Fig. 2. Visualisation of the distribution of the 10 unseen class images on AwA using t-SNE [25]. Different classes as well as their corresponding new class representations (squares in (a), (b) and (c), and circels in (d)) are shown in different colours. (c) and (d) are copied from [6] and [8], respectively.

formances. To be specific, the average performance improvement is 11.8% on AwA. And there are also significant improvements on other datasets. These results validate the enhancement ability of the proposed strategy, even if it is based on a simple regression.

In addition, while SAE [5] reported 61.4% on CUB, we obtain only 53.2%. This could be explained by the different normalizations, since there are no details of their implementation on CUB. Nevertheless, the proposed strategy still boosts its performance by 8.0%.

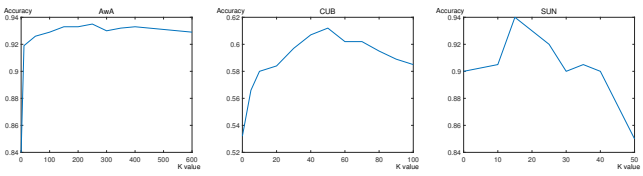
4.2.3. Importance of visual feature space

In order to validate applying the proposed strategy in visual feature space is better than its in semantic embedding space, we also implement the proposed strategy in semantic embedding space on AwA and CUB. SR only learns the semantic-visual projection, so we rebuild the simple regression and its objective function becomes:

$$\min_{W} \|X^s W - S_x^s\|_F^2 + \eta\Omega(W) \tag{3}$$

where all the variables have the same meaning as Eq (1). Since SAE is sensitive to data normalization, in order to avoid the effect of man-made poor data normalization, we only provide results of SR.

From the results of Table 4, the proposed strategy applied in visual feature space has better performances than its in semantic embedding space. There are two reasons for these results, the first one is that SR is ridge regression based method and learning the visual-semantic projection can make the hubness problem worse [4]. The other reason is that proposed strategy is based on K-NN method, thus it is also affected by hubness phenomenon and it is naturally benefited by semantic-visual projection. Thus, the proposed strategy improves the performances (average in 8.8%) in visual feature space more than their (average in 3.8%) in semantic embedding space.



(a)AwA (b)CUB (c)SUN

Fig. 3. Accuracy of different K value on three small datasets.

4.2.4. Selection of optimal K

For ASMFS-SAE, we tested $K = \{k|k = 5n, n = [1, ..., 120]\}$ on the three small-scale datasets. Fig.3 shows some prominent results with some K values on the three small datasets. The optimal K is data dependent, and the proposed method achieves outstanding performances when K is setted to be the average of sample numbers of each train set class(AwA: 618, CUB: 60, SUN: 20). Thus, it can be a guidance to choose the optimal K.

4.2.5. Further analysis

The proposed strategy makes semantic prototypes closer to clustering centers. Therefore, the new semantic class proto-

types can well represent cluster centers. Fig.2((a), (c) and (d)) illustrate our new semantic prototypes are closer to their corresponding class centers than the other two methods (Soravit *et al* [8] and Zhang *et al* [6]) which represent state-of-the-art. In Fig. 2(d), most of the new class prototypes are not close to their cluster centers. In Fig 2(c), only one new class prototype of the blue cluster in the middle is close to the cluster center.

In Fig. 2(a) of the proposed method, there are at least three new class prototypes close to their cluster centers: Dark green cluster, light green cluster, light blue cluster, respectively. Fig.2((a) and (b)) shows that the method boosted by the proposed strategy obtains better class centers. Most of the new class prototypes are closer to their cluster centers after the enhancement of the proposed method. It once again validate the proposed strategy has significant effectiveness that it obtains better semantic class prototypes.

5. CONCLUSION

In this paper, we proposed an universal strategy with low over cost that significantly boosts existing ridge regression based ZSL models. The key idea is that using a K-NN method to align manifolds of visual feature instances and semantic class prototypes in visual feature space. Extensive experiments show that the method boosted by the proposed strategy outperforms existing state-of-the-art methods and validate the importance of applying the proposed strategy in visual feature space.

6. REFERENCES

[1] Irving Biederman, "Recognition-by-components: a theory of human image understanding," *Psychological review*, vol. 94, no. 2, pp. 115, 1987.

[2] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, "Attribute-based classification for zero-shot visual object categorization," *TPAMI*, vol. 36, no. 3, pp. 453–465, 2014.

[3] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *JMLR*, vol. 11, no. 9, pp. 2487–2531, 2010.

[4] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto, "Ridge regression, hubness, and zero-shot learning," in *ECML PKDD*, 2015, pp. 135–151.

[5] Elyor Kodirov, Tao Xiang, and Shaogang Gong, "Semantic autoencoder for zero-shot learning," in *CVPR*, 2017, pp. 4447–4456.

[6] Li Zhang, Tao Xiang, and Shaogang Gong, "Learning a deep embedding model for zero-shot learning," in *CVPR*, 2017, pp. 3010–3019.

[7] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni, "Improving zero-shot learning by mitigating the hubness problem," in *ICLR workshop*, 2014, pp. 135–151.

[8] Soravit Changpinyo, Wei-Lun Chao, and Fei Sha, "Predicting visual exemplars of unseen classes for zero-shot learning," *arXiv preprint arXiv:1605.08151*, 2016.

[9] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele, "Evaluation of output embeddings for fine-grained image classification," in *CVPR*, 2015, pp. 2927–2936.

[10] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong, "Unsupervised domain adaptation for zero-shot learning," in *ICCV*, 2015, pp. 2452–2460.

[11] Ziming Zhang and Venkatesh Saligrama, "Zero-shot learning via joint latent similarity embedding," in *CVPR*, 2016, pp. 6034–6042.

[12] Yunlong Yu, Zhong Ji, Xi Li, Jichang Guo, Zhongfei Zhang, Haibin Ling, and Fei Wu, "Transductive zero-shot learning with a self-training dictionary approach," *arXiv preprint arXiv:1703.08893*, 2017.

[13] Yunlong Yu, Zhong Ji, Jichang Guo, and Yanwei Pang, "Transductive zero-shot learning with adaptive structural embedding," *arXiv preprint arXiv:1703.08897*, 2017.

[14] Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, and Yueting Zhuang, "Zero-shot recognition using dual visual-semantic mapping paths," in *CVPR*, 2017, pp. 5207–5215.

[15] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie, "The caltech-ucsd birds-200-2011 dataset," *California Institute of Technology*, 2011.

[16] Genevieve Patterson, Chen Xu, Hang Su, and James Hays, "The sun attribute database: Beyond categories for deeper scene understanding," *IJCV*, vol. 108, no. 1-2, pp. 59–81, 2014.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha, "Synthesized classifiers for zero-shot learning," in *CVPR*, 2016, pp. 5327–5336.

[21] Ziming Zhang and Venkatesh Saligrama, "Zero-shot learning via semantic similarity embedding," in *CVPR*, 2015, pp. 4166–4174.

[22] Yanwei Fu and Leonid Sigal, "Semi-supervised vocabulary-informed learning," in *CVPR*, 2016, pp. 5337–5346.

[23] Meng Ye and Yuhong Guo, "Zero-shot classification with discriminative semantic representation learning," in *CVPR*, 2017, pp. 7140–7148.

[24] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al., "Devise: A deep visual-semantic embedding model," in *NIPS*, 2013, pp. 2121–2129.

[25] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *JMLR*, vol. 9, no. 11, pp. 2579–2605, 2008.