DEEP ZERO-SHOT LEARNING FOR SCENE SKETCH

Yao Xie[+] *Peng Xu*[+] *Zhanyu Ma*[*]

Pattern Recognition and Intelligent System Lab., Beijing University of Posts and Telecommunications, China.

ABSTRACT

We introduce a novel problem of scene sketch zero-shot learning (SSZSL), which is a challenging task, since (i) different from photo, the gap between common semantic domain (*e.g.*, word vector) and sketch is too huge to exploit common semantic knowledge as the bridge for knowledge transfer, and (ii) compared with single-object sketch, more expressive feature representation for scene sketch is required to accommodate its high-level of abstraction and complexity. To overcome these challenges, we propose a deep embedding model for scene sketch zero-shot learning. In particular, we propose the augmented semantic vector to conduct domain alignment by fusing multi-modal semantic knowledge (*e.g.*, cartoon image, natural image, text description), and adopt attention-based network for scene sketch feature learning. Moreover, we propose a novel distance metric to improve the similarity measure during testing. Extensive experiments and ablation studies demonstrate the benefit of our sketch-specific design.

***Index Terms*—** Scene Sketch, Zero-Shot Learning, Deep Embedding Model.

1. INTRODUCTION

Sketch is abstract yet highly illustrative. With the increasing popularity of touch-screen devices, more and more free-hand sketches are used for human-computer interaction (*e.g.*, people can draw sketch as query to search specific shoe, chair, hand-bag [1]). The application conveniences of sketches have raised a flourish of sketch-related research, including recognition [2], sketch-based image retrieval [3, 1], sketch hashing [4], generation [5, 6], abstraction [7], *etc*. However, most of the existing works focus on *single-object sketches* (*e.g.*, apple, clock), leaving the *scene sketches* under-studied. Scene sketches are more abstract and complicated due to multiple objects and their interactions. In this paper, we propose a novel problem of **scene sketch zero-shot learning** (SSZSL), which is more challenging than scene sketch understanding [8, 9] and single-object sketch zero-shot learning [10].

Zero-shot learning methods rely on a labelled training set of *seen classes* and the knowledge about how an *unseen class* is semantically related to the seen classes. Seen and unseen classes are usually related in a high dimensional semantic knowledge domain, where the knowledge from seen classes can be transferred to unseen classes [11]. In previous computer vision works, word vector [12, 13], attribute vector [14, 15, 16, 17, 18] or text description [19] are widely studied as semantic knowledge, which is also used in the existing zero-shot learning methods engineered for photos. Different from photo, the gap between these semantic domains and sketch is too huge to exploit common semantic knowledge as the bridge for knowledge transfer. Moreover, we aim to solve the scene sketch zero-shot classification. Therefore, the main challenge to SSZSL is how to choose a reasonable semantic knowledge. In this work, we exploit a novel semantic knowledge, termed as *augmented semantic vector*, which can be obtained by fusing common semantic knowledge with the information from other modalities (*e.g.*, cartoon image, natural image, text description). Based on our augmented semantic vector, we propose a deep embedding model to solve scene sketch zero-shot learning, in which we adopt visual feature space of scene sketch as the embedding space to alleviate hubness problem [20]. Moreover, considering the high-level abstraction and complexity of scene sketch, we use attention-based technique to obtain discriminative feature representations for scene sketch. Most of the existing zero-shot learning (ZSL) methods use either Euclidean distance [20] or cosine similarity [21] as feature distance metric for testing. We define a new distance metric by combining Euclidean distance and cosine distance simultaneously, and achieved better evaluation results.

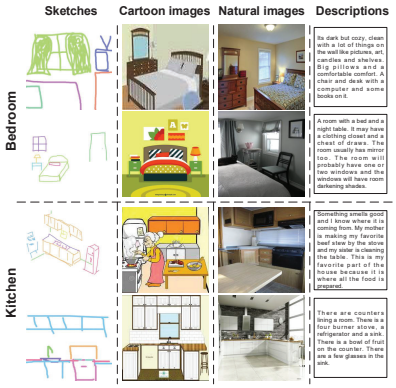

Fig. 1: Some samples from CMPlaces [22].

The contributions in this paper can be summarized as fol-

[+]These authors contributed equally. [*] Corresponding author.

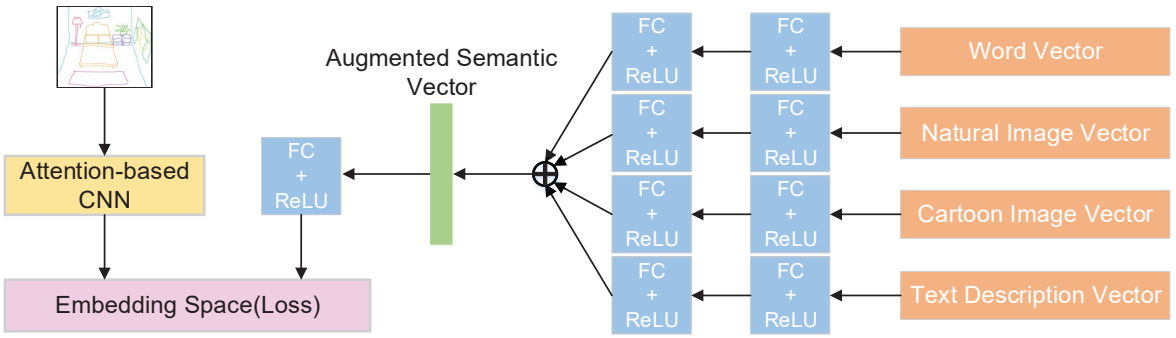

Fig. 2: The network architecture of our SSZSL model.

lows: (i) To the best of our knowledge, this is the first time that zero-shot learning setting is defined in scene sketch classification task. We propose this novel problem, and illustrate its intrinsic traits. (ii) We propose a deep embedding model for scene sketch zero-shot learning (SSZSL), which achieves a better performance than the state-of-the-art word vector based ZSL methods. Our superior performances effectively demonstrate the benefit of our sketch-specific design. (iii) Moreover, we define a new distance metric that performs better than conventional metric in SSZSL testing.

2. RELATED WORK

Sketch can be used for human-computer interaction, thus sketch recognition has been the important research topic in recent years. Most of previous works focus on single-object sketch classification or retrieval [23, 2, 24, 25, 26], leaving scene sketch classification under-studied. In particular, scene sketch zero-shot learning (SSZSL) has not been studied to date. To the best of our knowledge, only Ye *et al*. [8] have proposed a deep CNN model "Scene-Net" for scene sketch classification.

Existing ZSL methods engineered for scene photo mainly differ in what semantic knowledge are used: typically either word vector [12], attribute [14, 15, 27] or text description [19]. Sketch is different from photo. Due to the high-level abstraction, the domain gap between scene sketch and these common semantic knowledge is huge so that existing ZSL methods designed for photo fail to perform well on scene sketch. In this paper, considering the intrinsic traits of scene sketch, we propose a deep embedding model to solve SSZSL, in which we propose the augmented semantic vector as the semantic knowledge to conduct the knowledge transfer and domain alignment[12, 28].

3. LEARNING A MODEL FOR SCENE SKETCH ZERO-SHOT CLASSIFICATION

3.1. Problem Formulation

We now provide a formal denition of the scene sketch zero-shot learning (SSZSL). Let $\mathcal{S}_{tr} = \{(\mathbf{S}_i, \mathbf{x}_i, \mathbf{y}_i^u, t_i^u)\}_{i=1}^M$ denote a labelled training set of M training samples, where $\mathbf{x}_i \in \mathbb{R}^{J\times 1}$ is the visual feature vector of the i-th training scene sketch $\mathbf{S}_i$. $\mathbf{y}_i^u \in \mathbb{R}^{L\times 1}$ and $t_i^u \in \mathcal{T}_{tr}$ denote the semantic representation vector and class label of $\mathbf{S}_i$, which belongs to the u-th training class.

Given a new test sketch $\mathbf{S}_j$ with its feature visual vector $\mathbf{x}_j$, the goal of SSZSL is to predict a class label t_j^v by learning a classifier $f : \mathbf{x}_j \rightarrow t_j^v$, where $t_j^v \in \mathcal{T}_{te}$ is the class label of the j-th test instance $\mathbf{S}_j$ belonging to v-th test class. The training (seen) classes and test (unseen) classes are disjoint, *i.e.*, $\mathcal{T}_{tr} \cap \mathcal{T}_{te} = \varnothing$. Note that each class label t^u or t^v is associated with a pre-defined semantic representation $\mathbf{y}^u$ or $\mathbf{y}^v$.

3.2. Deep Embedding Model

As is shown in Fig. 2, there are two branches in our model. The first branch is the visual feature extraction branch for scene sketch, composed of a attention-based network. It takes a scene sketch $\mathbf{S}_i$ as input and outputs a visual feature vector $\mathbf{x}_i$. To alleviate hubness problem, we will use this J-dimensional visual feature space as embedding space, where both the scene sketch and its corresponding semantic representation vector will be embedded. The second branch is the semantic embedding branch, which will embed the semantic knowledge into our embedding space to conduct the domain alignment. Considering the huge domain gap between common semantic domain (*e.g.*, word vector, text description) and high-level abstract scene sketch, we propose the augmented semantic vector for SSZSL. In particular, our semantic embedding branch takes semantic representations from different modalities as input and it is implemented by three fully connected (FC) + Rectied Linear Unit (ReLU) layers with ℓ_2 parameter regularization, where the first two FC+ReLU layers are used to obtain our augmented semantic vector (see Sec. 3.3) and it will be embedded to the embedding space by the third FC+ReLU layer.

Then the outputs of two branches are connected by a mean square error (MSE) loss which aims to minimize the difference between visual feature vector $\mathbf{x}_i$ and its embedded semantic vector in the visual feature space. Our loss function is

$$J(\mathbf{W}) = \frac{1}{M}\sum_{i=0}^{M} \|\mathbf{x}_i - \Phi(\mathbf{y}_i^u : \mathbf{W})\|_2^2 + \lambda\|\mathbf{W}\|_2^2, \quad (1)$$

where $\mathbf{W}$ contains the weights of the semantic embedding branch. λ is the hyperparameter to weight the strength of the ℓ_2 parameter regularization loss against the MSE loss. $\Phi(\cdot)$ denotes the feature extraction by our semantic branch.

After training, to predict label t_j^v, we can calculate the distance between the embedded semantic vectors and $\mathbf{x}_j$ as

$$t_j^v \leftarrow v = \arg\min_v D(\Phi(\mathbf{y}^v), \mathbf{x}_j), \tag{2}$$

where D is a distance metric function defined by us (see Sec. 3.4), and $\mathbf{y}^v$ is the embedded semantic vector of the v-th unseen class.

3.3. Augmented Semantic Vector for Scene Sketch

In this work, we exploit a novel semantic knowledge for SSZSL, termed as augmented semantic vector. As illustrated in Fig. 2, our semantic embedding branch takes in semantic representations from different modalities (*i.e.*, word vector (W), natural image vector (I), cartoon image vector (C), text description vector (T)), and after the first two FC+ReLU layers, it outputs the angmented semantic vector $\tilde{\mathbf{y}}_i^u$ by element-wise addition and non-linear fusing

$$\begin{aligned}\tilde{\mathbf{y}}_i^u =& \Phi^W((\mathbf{y}_i^u)^W) + \Phi^C((\mathbf{y}_i^u)^C) + \\ & \Phi^I((\mathbf{y}_i^u)^I) + \Phi^T((\mathbf{y}_i^u)^T),\end{aligned} \tag{3}$$

where Φ^W, Φ^C, Φ^I, and Φ^T denote the mapping of the first two FC layers of our semantic branch (See Fig. 2). $(\mathbf{y}_i^u)^W$, $(\mathbf{y}_i^u)^C$, $(\mathbf{y}_i^u)^I$, and $(\mathbf{y}_i^u)^T$ denote the representation vectors from four modalities. To obtain natural image vector, we train a deep classifier network based on natural images and calculate a vector representation for each sample class by averaging the deep features by category. For cartoon image vector and text description vector, the similar process is used. For word vector, we adopt the word2vec model (based on model library in Gensim), which was trained with over $8,000,000$ text documents from Wiki-pedia, to represent each class (including seen and unseen classes).

3.4. Distance Metric

Euclidean distance is a frequently-used distance metric, however it will fail for the case shown in Fig. 3. $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ represent the embedded semantic representation vectors corresponding to two categories. $\boldsymbol{v}$ is a visual feature vector belonging to $\boldsymbol{c}_1$, where $\|\boldsymbol{v}-\boldsymbol{c}_1\|^2 > \|\boldsymbol{v}-\boldsymbol{c}_2\|^2$ and $\theta_1 < \theta_2$. If only using Euclidean distance, $\boldsymbol{v}$ will be classified to $\boldsymbol{c}_2$. However, if only use Cosine distance, $\boldsymbol{v}$ will be classified to $\boldsymbol{c}_1$. Therefore we propose a new distance metric termed as Euclidean Cosine (EC) distance to alleviate this problem. Our EC distance metric is defined as follows:

$$\begin{cases} D(\boldsymbol{\alpha},\boldsymbol{\beta}) = (1-\eta cos\langle\boldsymbol{\alpha},\boldsymbol{\beta}\rangle)\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|_2^2 \\ cos\langle\boldsymbol{\alpha},\boldsymbol{\beta}\rangle = \frac{\boldsymbol{\alpha}\cdot\boldsymbol{\beta}}{\|\boldsymbol{\alpha}\|\cdot\|\boldsymbol{\beta}\|} \end{cases}, \tag{4}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors, and $\eta(0 \leq \eta \leq 1)$ is a weighting coefcient that controls the importance of cosine distance. Assuming that η is 0.9, in Fig. 3, we can obtain $D(\boldsymbol{v},\boldsymbol{c_1}) < D(\boldsymbol{v},\boldsymbol{c_2})$, where $D(\boldsymbol{v},\boldsymbol{c_1})$ and $D(\boldsymbol{v},\boldsymbol{c_2})$ are approximately 0.2758 and 0.3636, respectively.

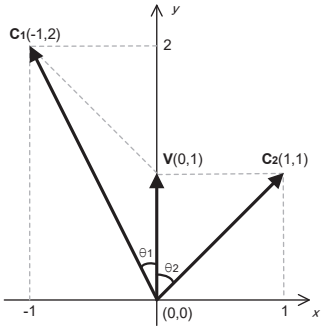

Fig. 3: Illustration for distance metric.

4. EXPERIMENT

4.1. Dataset

In the following experiments, CMPlaces [22] servers as the benchmark for our SSZSL experiment, which contains hundreds of natural scene categories across five modalities, including natural images, sketches, cartoons images, text descriptions, and spatial text images[1]. Some samples of CMPlaces are illustrated in Fig. 1, and we can observe that the scene sketches are abstract and have rich information. All the following zero-shot experiments are performed for scene sketch, while natural images ($\mathbf{I}$), cartoons images ($\mathbf{C}$) and text descriptions ($\mathbf{T}$) are used to obtain our augmented semantic vector. Ignoring some classes that are too messy to recognize, we selected approximate $15,000$ scene sketches from 174 classes. In particular, 144 classes were used as seen classes and the remaining 30 classes were used as unseen classes.

4.2. Experimental Settings

All our experiments are implemented in PyTorch, and on a single GEFORCE GTX 1080 Ti GPU card.

Visual Feature Extraction Branch We use ImageNet pretrained SE-Resnet-50 [32] as our attention-based CNN model to conduct visual feature extracting for scene sketches, and only modify its fully connected layers to fine-tune it on the scene sketches from 144 seen classes. The output dimension of our visual feature extraction branch is 2048. We use SGD optimizer with a initial learning rate lr of 0.001, momentum of 0.9 and a mini-batch size of 16. We decrease lr by 0.9 every epoch and terminate the optimization after 30 epochs. All input sketches or images are resized to $3 \times 224 \times 224$. The loss weighting factor λ in Eq. 1 and hyperparameter η in Eq. 4 are experimentally set to 0.0005 and 0.9, respectively.

Semantic Embedding Branch In our semantic embedding branch, the outputs of three FC layers are set to 512, 1024, and 2048, respectively. Adam is used to optimise our semantic embedding branch with a initial learning rate of 0.0001 and a mini-batch size of 256.

[1] Spatial text images are not available in current online CMPlaces dataset.

Table 1: SSZSL (Top1/Top5) accuracy (%) comparison with state-of-the-art ZSL models on CMPlaces.

Model	Default Metric					EC Distance Metric				
	W	C	I	T	W+C+I+T	W	C	I	T	W+C+I+T
DEM (V → S) [20]	23.5/39.9	30.9/70.5	26.8/61.4	38.2/77.2	38.8/78.6	30.2/58.3	34.8/72.8	27.3/62.3	39.3/77.9	40.4/79.3
DEM (S → V) [20]	23.9/54.4	41.7/79.6	36.9/78.3	41.3/82.1	47.7/84.1	30.9/67.5	44.6/81.4	37.7/80.7	42.5/82.6	48.1/84.7
SAE [29]	25.9/58.9	34.1/72.5	27.5/71.1	33.1/76.7	45.5/84.1	30.7/66.2	34.5/73.8	29.7/73.6	35.3/77.1	48.7/86.2
f-CLSWGAN [30]	18.1/44.1	39.9/73.3	32.5/70.5	37.3/70.9	48.7/83.4	-	-	-	-	-
RELATION NET [31]	30.6/66.1	40.3/76.4	35.8/74.5	40.5/79.1	47.6/82.5	-	-	-	-	-
Ours	30.4/61.5	45.2/80.5	38.9/81.6	43.4/82.9	52.1/85.7	**35.6/70.2**	**46.9/81.8**	**40.1/82.8**	**45.2/83.7**	**54.0/86.9**

Table 2: Ablation study for our proposed model: SSZSL (Top1/Top5) accuracy (%) on CMPlaces.

Semantic Representation	EC Distance		Euclidean Distance	
	S → V	**V → S**	**S → V**	**V → S**
W	35.6/70.2	33.6/61.3	30.4/61.5	25.1/41.9
C	**46.9**/81.8	36.5/74.6	45.2/80.5	32.6/71.2
I	40.1/82.8	29.4/63.9	38.9/81.6	28.9/63.2
T	45.2/**83.7**	42.5/78.9	43.4/82.9	41.4/77.9
C+T	**51.1/86.3**	44.8/81.0	49.6/84.6	42.7/78.5
I+T	49.4/85.6	44.6/81.1	48.3/84.9	40.6/80.3
I+C	47.9/84.6	36.7/78.0	46.1/83.2	36.5/77.8
W+C	47.5/82.4	41.8/75.6	44.6/80.9	34.5/68.7
W+T	46.7/84.1	44.4/79.7	45.1/82.7	40.4/74.7
W+I	45.1/83.4	31.9/68.4	43.3/80.9	29.7/64.2
C+I+T	**52.8/86.4**	43.2/81.4	51.7/85.6	42.8/80.3
W+C+I	49.0/84.6	38.0/78.8	48.6/83.2	36.6/77.8
W+C+T	51.7/86.2	44.7/80.4	50.5/85.2	41.0/77.5
W+I+T	50.4/86.1	42.3/79.7	49.6/84.9	41.6/78.9
W+C+I+T	**54.0/86.9**	43.6/81.4	52.1/85.7	42.4/79.8

Competitors. As aforementioned state, SSZSL is a novel problem, thus there is no existing methods can perform as our baselines. Therefore, we have to compare with state-of-the-art ZSL methods, including DEM [20], SAE [29], f-CLSWGAN [30], and RELATION NET [31].

4.3. Results and Discussion

First of all, we compare our proposed model with the state-of-the-art ZSL models on CMPlaces dataset, as illustrated in Tab. 1. For a fair comparison, we use Resnet-50 as their visual feature extractor. These selected competitors have performed well using word vector (W) as semantic input [20, 29, 30, 31], thus we also evaluate them based on word vector. Moreover, in order to demonstrate the importance of semantic knowledge for SSZSL, other semantic knowledge (*i.e.*, natural image vector (I), cartoon image vector (C), text description vector (T)), and their fused vector (W+C+I+T) are also evaluated as semantic input. In our experiments, we find that our proposed distance metric outperforms both of Euclidean distance and cosine distance for SSZSL testing, and cosine distance performs really poor. Therefore, if our proposed metric can be applied to these ZSL baselines in Tab. 1, they are also calculated on our proposed metric. In Tab. 1, we can observe that: (i) Our proposed model outperforms all the competitors by a large margin based on different semantic knowledge. This is benefit from that we choose sketch feature space as a reasonable embedding space and our attention-based network achieves better feature representation. (ii) When using the augmented semantic vector based on four modalities, our model obtains a obvious performance improvement, and our augmented semantic vector also improves the performances of all the selected baselines. This demonstrates the superiority of our sketch-specific augmented semantic vector.

The results of our ablation study are reported in Tab. 2. As aforementioned state, our semantic embedding branch is scalable that can be adaptive to combinations of semantic vectors from different modalities, thus we evaluate our proposed model based on these combinations. In Tab. 2, we can make following observations: (i) For SSZSL, our augmented semantic vector performs better than single-modal semantic vector. (ii) Choosing a reasonable embedding space is important. Using sketch visual feature space as embedding space ($S \rightarrow V$) obtains better performance than using semantic space as embedding space ($V \rightarrow S$). This interesting phenomenon can be explained by the hubness issue discussed in [20]. (iii) Our proposed EC distance outperforms Euclidean distance for zero-shot testing on CMPlaces. (iv) When using single-modal semantic vector for our model, cartoon image vector outperforms word vector by a clear margin (46.9% vs. 35.6%), since the domain gap between cartoon images and sketch is smaller.

5. CONCLUSION

In this paper, we introduce a novel problem of scene sketch zero-shot learning (SSZSL). We have proposed a deep embedding model for scene sketch zero-shot learning. The model differs from existing ZSL model in that we propose the augmented semantic vector to conduct domain alignment by fusing multi-modal semantic knowledge, and adopt attention-based network for scene sketch feature learning. What's more, a new distance metric is used instead of Euclidean distance. Experimental results on CMPlaces validated the effectiveness of the proposed method.

Acknowledgement This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 61773701, in part by the Beijing Nova Program No. Z171100001117049, in part by the Beijing Nova Program Interdisciplinary Cooperation Project No. Z181100006218137, and in part by BUPT Excellent PhD Student Foundation CX2017307 and BUPT-SICE Excellent Graduate Student Innovation Foundation (2016).

References

[1] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *ICCV*, 2017.

[2] Rosália G Schneider and Tinne Tuytelaars, "Sketch classification and classification-driven analysis using fisher vectors," *ACM Transactions on Graphics*, 2014.

[3] Peng Xu, Qiyue Yin, Yonggang Qi, Yi-Zhe Song, Zhanyu Ma, Liang Wang, and Jun Guo, "Instance-level coupled subspace learning for fine-grained sketch-based image retrieval," in *ECCV workshops*, 2016.

[4] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, Zhanyu Ma, and Jun Guo, "Sketchmate: Deep hashing for million-scale human sketch retrieval," in *CVPR*, 2018.

[5] Conghui Hu, Da Li, Yi-Zhe Song, and Timothy M Hospedales, "Now you see me: Deep face hallucination for unviewed sketches," in *BMCV*, 2016.

[6] Wengling Chen and James Hays, "Sketchygan: Towards diverse and realistic sketch to image synthesis," in *CVPR*, 2018.

[7] Umar Riaz Muhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales, "Learning deep sketch abstraction," in *CVPR*, 2018.

[8] Yuxiang Ye, Yijuan Lu, and Hao Jiang, "Human's scene sketch understanding," in *ICMR*, 2016.

[9] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang, "Sketchyscene: Richly-annotated scene sketches," in *ECCV*, 2018.

[10] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal, "A zero-shot framework for sketch based image retrieval," in *ECCV*, 2018.

[11] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong, "Zero-shot object recognition by semantic manifold distance," in *CVPR*, 2015.

[12] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng, "Zero-shot learning through cross-modal transfer," in *NIPS*, 2013.

[13] Kun Liu, Wu Liu, Huadong Ma, Wenbing Huang, and Xiongxiong Dong, "Generalized zero-shot learning for action recognition with web-scale video data," *WWWJ*, 2017.

[14] Vittorio Ferrari and Andrew Zisserman, "Learning visual attributes," in *NIPS*, 2008.

[15] Devi Parikh and Kristen Grauman, "Relative attributes," in *ICCV*, 2011.

[16] Qi Dong, Shaogang Gong, and Xiatian Zhu, "Multi-task curriculum transfer deep learning of clothing attributes," in *WACV*, 2017.

[17] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li, "Attribute recognition by joint recurrent learning of context and correlation," in *ICCV*, 2017.

[18] Minxian Li, Xiatian Zhu, and Shaogang Gong, "Unsupervised person re-identification by deep learning tracklet association," in *ECCV*.

[19] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele, "Learning deep representations of fine-grained visual descriptions," in *CVPR*, 2016.

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

[21] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal, "A zero-shot framework for sketch based image retrieval," in *ECCV*, 2018.

[22] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba, "Learning aligned cross-modal representations from weakly aligned data," in *CVPR*, 2016.

[23] Mathias Eitz, James Hays, and Marc Alexa, "How do humans sketch objects?," *ACM Transactions on Graphics*, 2012.

[24] Peng Xu, Ke Li, Zhanyu Ma, Yi-Zhe Song, Liang Wang, and Jun Guo, "Cross-modal subspace learning for sketch-based image retrieval: A comparative study," in *IC-NIDC*, 2016.

[25] Yue Zhong, Honggang Zhang, Jun Guo, and Yi-Zhe Song, "Directional element hog for sketch recognition," in *IC-NIDC*, 2018.

[26] Peng Xu, Qiyue Yin, Yongye Huang, Yi-Zhe Song, Zhanyu Ma, Liang Wang, Tao Xiang, W Bastiaan Kleijn, and Jun Guo, "Cross-modal subspace learning for fine-grained sketch-based image retrieval," *Neurocomputing*, 2017.

[27] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama, "Generalized zero-shot recognition based on visually semantic embedding," *arXiv preprint arXiv:1811.07993*, 2018.

[28] Xiaoxu Li, Liyun Yu, Dongliang Chang, Zhanyu Ma, and Jie Cao, "Dual cross-entropy loss for small-sample fine-grained vehicle classification," *IEEE Transactions on Vehicular Technology*, 2019.

[29] Elyor Kodirov, Tao Xiang, and Shaogang Gong, "Semantic autoencoder for zero-shot learning," in *CVPR*, 2017.

[30] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata, "Feature generating networks for zero-shot learning," in *CVPR*, 2018.

[31] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018.

[32] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.