Supplementary Material for
VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

Wenjia Xu[1,7,8] Yongqin Xian[2] Jiuniu Wang[5,7,8] Bernt Schiele[3] Zeynep Akata[3,4,6]

[1] Beijing University of Posts and Telecommunications [2] ETH Zurich
[3] Max Planck Institute for Informatics [4] University of Tübingen [5] City University of Hong Kong
[6] Max Planck Institute for Intelligent Systems
[7] University of Chinese Academy of Sciences [8] Aerospace Information Research Institute, CAS

Contents

In this supplementary material, we provide:

A . User Study 2

In this section, we introduce the details for user evaluation, and present the user discovered semantics in AWA2 dataset.

B . Additional Qualitative Results 2

In this section, we present additional qualitative results for SUN and CUB datasets.

C . Ablation Study 2

In this section, we include the ablation study results on CUB and SUN dataset. We measure the influence of cluster number D_v and patch number N_t on the ZSL performance of our learnt `VGSE-SMO` embeddings.

D . SOTA results for `VGSE-WAvg` 3

In this section, we present the results of applying `VGSE-WAvg` embeddings on the state-of-the-art ZSL models.

E . Implementation Details 5

In this section, we first describe in detail how we generate image patches. then present the training details for our experiments.

F . Limitations and Broader Impact 6

In this section, we discuss the possible limitations and impact of our work.

Step1: observe a folder with 30 images

Step2: answer the following questions

There are 30 images coming from one cluster. Please observe the images and rate them with the following score:
(1: not at all 2: mostly not 3: yes 4: mostly yes 5: absolutely)

1. Do images in this cluster contain consistent visual property?
2. Do images in this cluster convey consistent semantic information?
3. Please name the semantics you observed from the clusters, if your answer to Q2 is true.

Figure A.1. The illustration of user study. Participants are required to observe a folder (a cluster containing 30 images), and rate the clusters according to the visual and semantic consistency, then name the semantics they observed in the clusters.

A. User Study

To evaluate if our VGSE embeddings convey consistent visual and semantic properties, we perform an user evaluation over the visual clusters. The illustration of user evaluation is shown in A.1. We randomly pick 50 clusters, each equipped with 30 images from the cluster center, and ask the users to observe the images and answer the following three questions. Q1: Do images in this cluster contain consistent visual property? Q2: Do images in this cluster convey consistent semantic information? Q3: Please name the semantics you observed from the clusters, if your answer to Q2 is true.

We rate both the clusters learnt by our model for AWA2 dataset, and the clusters learnt by k-means. For each experiment, we employed 5 annotators, i.e., postgraduate students (2 female) aged between 20 and 30 and majoring in computer science. In total, we collect 500 ratings for each experiment. We treat the ratings higher than 3 as a hit. The results reveal that in 88.5% and 87.0% cases, users think our clusters convey consistent visual and semantic information. While for k-means clusters, the results are 71.5% and 71.0%, respectively.

In addition, we display some of the clusters and their semantics named by users in Figure A.2. As shown in the figure, images in each clusters show consistent visual properties that are human understandable, e.g., local properties such as the *white fur*, *horns* and *stout legs*, and global properties such as *animals living near water* and *animals living near cage*.

B. Additional Qualitative Results

We show additional qualitative results for SUN and CUB datasets in Figure B.2 and Figure B.1. Images shown in each cluster represent the cluster center. We have the following observations. First, the image patches in each cluster convey consistent visual properties, e.g., the *slender bird legs* (row 1, column 1) and *white wing* (row 2, column 1) in Figure B.1; the *wheels* (row 1, column 2) and the *crowds* (row 2, column 3) in Figure B.2. Moreover, our clusters convey fine-grained semantics that may be neglected by human-annotated attributes, e.g., the *electrical screen* (row 2, column 2) in Figure B.2. Though some clusters are consist of background patches, they still convey semantic information that is category related. For instance, some birds in the CUB dataset may live near *Pine trees* and *Cypress trees* (row 2, column 3), and some may live near water (row 1, column 3) in Figure B.1. However, we can still observe some clusters with semantically different patches, i.e., the cluster of *grids* contains patches of window and fence (row 2, column 3 in Figure B.2)

C. Ablation Study

In this section, we include the ablation study results on CUB and SUN dataset. To measure the influence of the cluster number D_v on our semantic embeddings, we train the `PC` module with various D_v (results shown in Figure C.1a and Figure C.2a). The observation is similar to that of the AWA2 dataset. When the unseen semantic embeddings are predicted under an oracle setting (predicted from the unseen class images), various dimension D_v does not influence the classification accuracy on unseen classes (the orange curve). While under the ZSL setting where unseen semantic embeddings are predicted

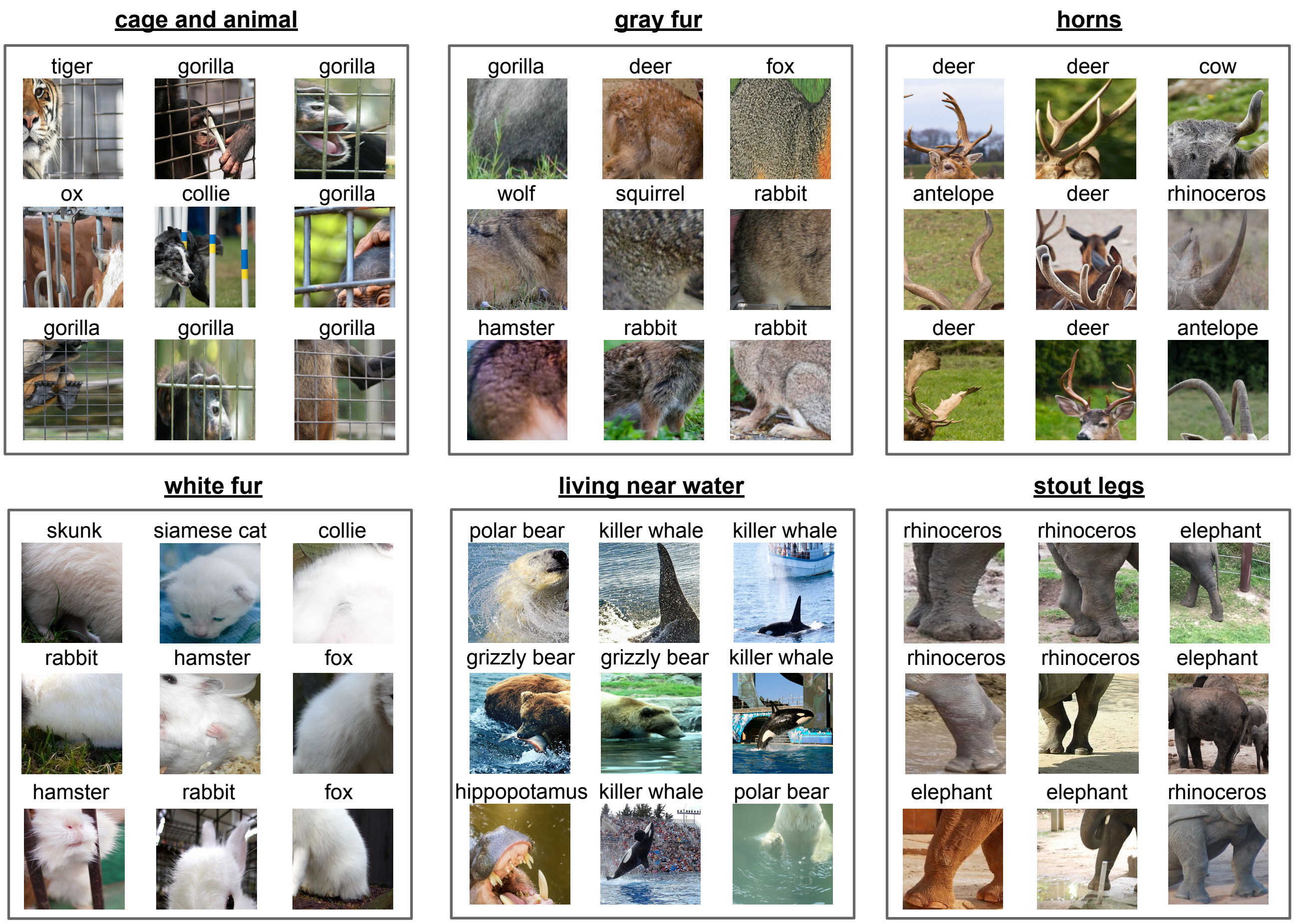


Figure A.2. Qualitative results for AWA2 dataset. Each box represents one cluster, with images from the cluster center. The name above each image is the category. The phrase above each cluster is the semantic named during user evaluation.

from class relations (`VGSE-SMO`), the cluster numbers influence the ZSL performance. Before the cluster number increases up to a breaking point ($D_v = 200$ for CUB dataset and $D_v = 300$ for SUN dataset), the ability of the semantic embeddings is also improved (from 24.4% to 26.3% on CUB and from 35.0% to 36.1% on SUN), since the learned clusters contain visually similar patches from different classes, which can model the visual relation between classes. However, increasing the number of clusters leads to small pure clusters (patches coming from one single category), resulting in poor generalization between seen and unseen classes.

The influence of the patch numbers are shown in Figure C.1b, which reveals two observations. First, with the patch number increase from 1 (single image clustering) to 9, the ZSL performance increases as well, since the image patches used for semantic embedding learning contain semantic object parts and thus result in better knowledge transfer between seen and unseen classes. However, for a large N_t, the patches might be too tiny to contain consistent semantic, thus resulting in performance dropping, e.g., the ZSL accuracy on CUB drops from 26.1% ($N_t = 9$) to 23.9% ($N_t = 128$). We also compare the patches generated by watershed segmentation proposal with using 3×3 grid patches. By comparing 3×3 grid with the patches generated by watershed segmentation proposal ($N_t = 9$), we found that using watershed as the region proposal results in accuracy boost (1.9% on CUB and 1.4% on SUN) compared to the regular grid patch, since the former patches tend to cover more complete object parts rather than random cropped regions.

D. SOTA results for `VGSE-WAvg`

In this section, we extend Table 1 in the main paper with the SOTA results of `VGSE-WAvg` embeddings. As shown in Table D.1, the `VGSE-WAvg` embeddings outperform the `w2v` embeddings by a large margin. In particular, when coupled with

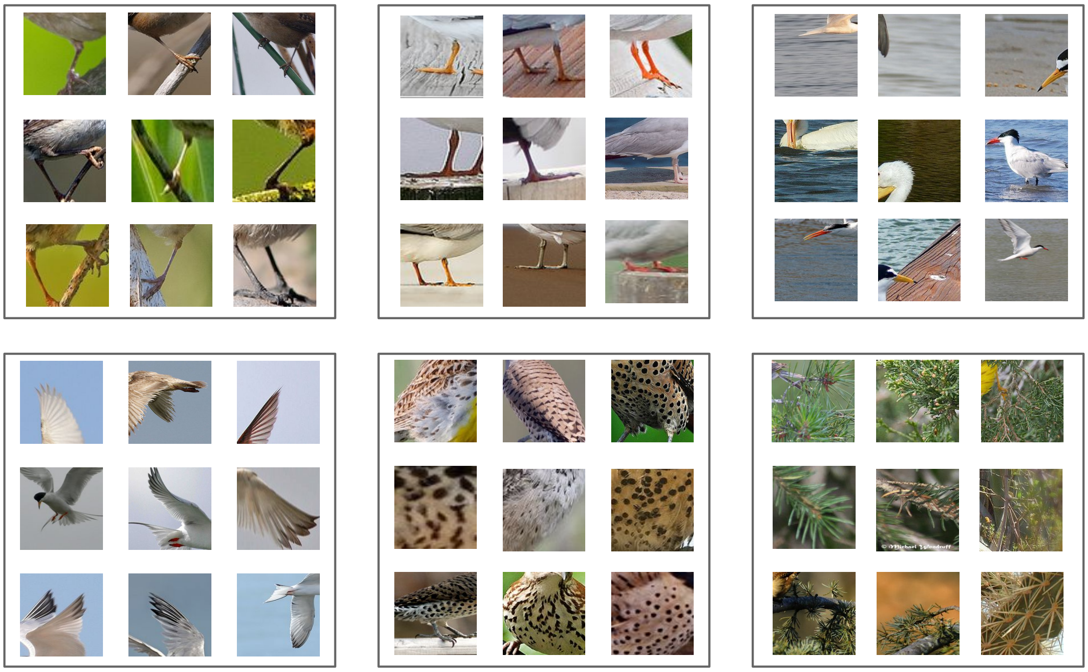


Figure B.1. Qualitative results for CUB dataset. Each box represents one cluster, with images from the cluster center.

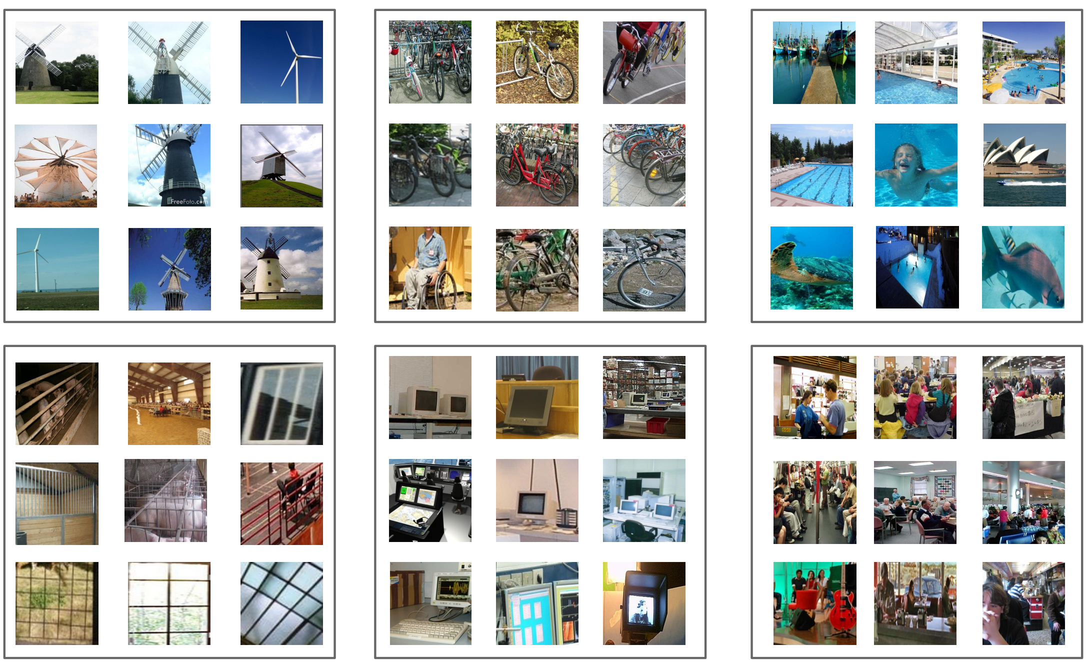


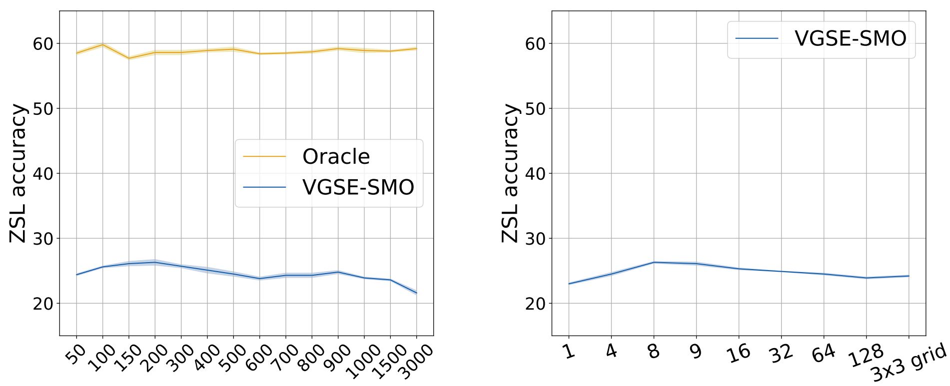
Figure B.2. Qualitative results for SUN dataset. Each box represents one cluster, with images from the cluster center.

APN, our VGSE-WAvg boosts the ZSL performance of w2v from 59.6% to 63.7% on AWA2 dataset and from 23.6% to 35.8% on SUN dataset. We compare our two class relation functions VGSE-WAvg and VGSE-SMO in Table D.1. The results demonstrate that VGSE-WAvg works on par with VGSE-SMO on SUN and CUB datasets, i.e., when coupled with f-VAEGAN-D2, VGSE-WAvg achieves 34.8% on CUB comparing to VGSE-SMO with 35.0%. While on AWA2 dataset, VGSE-SMO yields slightly better ZSL performance than VGSE-WAvg. In particular, when coupled with GEM-ZSL, VGSE-SMO (with 58.0%)

improves over VGSE−WAvg (with 53.3%) by 4.7%. The results indicate that predicting the unseen semantic embeddings with the weighted average of a few seen classes semantic embeddings (VGSE−WAvg) is working well for fine-grained datasets such as SUN and CUB, since the visual discrepancy between classes is small. However, for coarse-grained dataset AWA2, the class relation function considering all the seen classes embeddings (VGSE−SMO) works better.

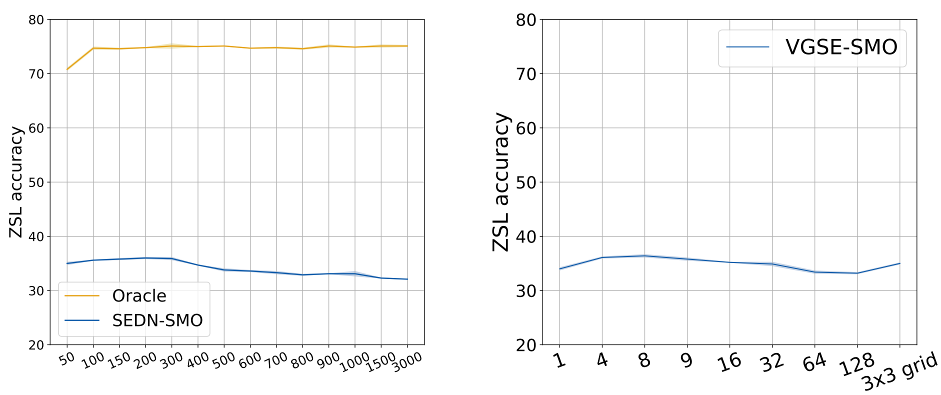
E. Implementation Details

Image regions. To discover the clusters of image patches, we crop the image x_n into N_t patches $\{x_{nt}\}_{t=1}^{N_t}$. Previous works [9, 10] obtain $1,000$ regions for each image with Selective Search [11], resulting in large amount of overlapped patches. To avoid that, we crop the segments generated by unsupervised compact watershed segmentation algorithm [7] into image patches. In detail, for each image x_n, we find the smallest bounding box that fully covers each segment and crop x into N_t rectangular patches $\{x_{nt}\}_{t=1}^{N_t}$ that cover different parts of the image. In our experiment, the patch number N_t is set to 9, and the tiny patches with $w < W/20$ or $h < H/20$ are removed, where w and W represents the width of the patch x_{nt} and the original image x_n respectively; h and H represents the height of the patch x_{nt} and the original image x_n.



(a) cluster number D_v (b) patch number N_t

Figure C.1. Ablation study on CUB dataset. (a) Influence of the cluster number $D_v = 50, \ldots, 3000$. In the oracle setting, we feed unseen classes images to the PC module to predict unseen semantic embeddings. (b) Influence of the patch number N_t we used per image with the watershed segmentation for obtaining our VGSE−SMO class embeddings. $N_t = 1$ uses the whole image (no patches). "3×3 grid" crops the image into 9 square patches. Both plots report ZSL accuracy with SJE model trained on CUB dataset (mean and std over 5 runs).



(a) cluster number D_v (b) patch number N_t

Figure C.2. Ablation study on SUN dataset. (a) Influence of the cluster number $D_v = 50, \ldots, 3000$. In the oracle setting, we feed unseen classes images to the PC module to predict unseen semantic embeddings. (b) Influence of the patch number N_t we used per image with the watershed segmentation for obtaining our VGSE−SMO class embeddings. $N_t = 1$ uses the whole image (no patches). "3×3 grid" crops the image into 9 square patches. Both plots report ZSL accuracy with SJE model trained on SUN dataset (mean and std over 5 runs).

			Zero-Shot Learning			Generalized Zero-Shot Learning								
			AWA2	CUB	SUN	AWA2			CUB			SUN		
	ZSL Model	**Semantic Embeddings**	**T1**	**T1**	**T1**	**u**	**s**	**H**	**u**	**s**	**H**	**u**	**s**	**H**
Generative	`CADA-VAE` [8]	`w2v` [6]	49.0	22.5	37.8	38.6	60.1	47.0	16.3	39.7	23.1	26.0	28.2	27.0
		`VGSE-WAvg` (Ours)	51.0	24.6	40.4	44.8	55.8	49.7	17.3	38.8	23.9	29.0	28.9	28.9
		`VGSE-SMO` (Ours)	52.7	24.8	40.3	46.9	61.6	53.9	18.3	44.5	25.9	**29.4**	29.6	29.5
	`f-VAEGAN-D2` [14]	`w2v` [6]	58.4	32.7	39.6	46.7	59.0	52.2	23.0	44.5	30.3	25.9	33.3	29.1
		`VGSE-WAvg` (Ours)	60.2	34.8	40.6	48.9	59.3	53.6	24.0	45.3	31.4	24.6	36.1	29.3
		`VGSE-SMO` (Ours)	61.3	**35.0**	**41.1**	45.7	66.7	54.2	**24.1**	45.7	**31.5**	25.5	**35.7**	**29.8**
Non-Generative	`SJE` [1]	`w2v` [6]	53.7	14.4	26.3	39.7	65.3	48.8	13.2	28.6	18.0	19.8	18.6	19.2
		`VGSE-WAvg` (Ours)	57.7	25.8	35.3	47.8	62.9	54.3	16.7	43.5	24.1	26.8	25.6	26.2
		`VGSE-SMO` (Ours)	62.4	26.1	35.8	46.8	72.3	56.8	16.4	44.7	24.3	28.7	25.2	26.8
	`GEM-ZSL` [5]	`w2v` [6]	50.2	25.7	-	40.1	80.0	53.4	11.2	**48.8**	18.2	-	-	-
		`VGSE-WAvg` (Ours)	53.3	27.5	-	41.4	77.6	54.0	13.3	42.0	20.2	-	-	-
		`VGSE-SMO` (Ours)	58.0	29.1	-	49.1	78.2	60.3	13.1	43.0	20.0	-	-	-
	`APN` [15]	`w2v` [6]	59.6	22.7	23.6	41.8	75.0	53.7	17.6	29.4	22.1	16.3	15.3	15.8
		`VGSE-WAvg` (Ours)	63.7	28.5	35.8	47.7	83.5	60.7	21.7	45.5	29.3	22.0	31.6	26.0
		`VGSE-SMO` (Ours)	**64.0**	28.9	38.1	**51.2**	**81.8**	**63.0**	21.9	45.5	29.5	24.1	31.8	27.4

Table D.1. Comparing our `VGSE-SMO`, `VGSE-WAvg`, with the `w2v` semantic embedding over state-of-the-art ZSL models. In ZSL, we measure Top-1 accuracy (**T1**) on unseen classes, in GZSL on seen/unseen (**s/u**) classes and their harmonic mean (**H**). Feature Generating Methods, i.e., `f-VAEGAN-D2`, and `CADA-VAE` synthesizing training samples, and `SJE`, `APN`, `GEM-ZSL` using only real image features.

Training details. In the patch clustering (PC) module, we learn seen-semantic embeddings with train set (seen classes) proposed by [13]. We adopt ResNet50 [3] pretrained on ImageNet1K [2] as the backbone. We use ADAM optimizer [4] by setting weight decay of 10^{-4} and learning rate of 10^{-4}. The cluster number D_v is set as 150 for three datasets. We set λ as 5 following [12]. The unseen-class embeddings are predicted in the class relation (CR) module without seeing unseen images. For the Weighted Average module, we set η as 5 for all datasets, and use 5 neighbors for all datasets. For the similarity matrix optimization, we set α as -1 for AWA2 and CUB, and as 0 for SUN. All hyperparameters are selected over the validation set. We set λ to 5 following [12]; β and γ to 1 for all datasets.

F. Limitations and Broader Impact

The generalization ability of our `VGSE` class embeddings depends to a great extent on the external knowledge used to model the seen and unseen class relations. External knowledge that can well capture the visual relation between classes will lead to higher ZSL accuracy. This motivates us to discover better external knowledge that captures both the semantic and visual relation between classes in future work. Broadly speaking, the prediction accuracy of current zero-shot learning models is still lower than models trained with both seen and unseen classes. To this end, ZSL models might not be applicable to situations that require high confidence and precision, e.g., medical auxiliary diagnosis and self-driving cars.

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 6

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[5] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. Goal-oriented gaze estimation for zero-shot learning. In *CVPR*, 2021. 6

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *NeurIPS*, 2013. 6

[7] Peer Neubert and Peter Protzel. Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. In *ICPR*. IEEE, 2014. 5

[8] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, 2019. 6

[9] Ronan Sicre, Yannis Avrithis, Ewa Kijak, and Frédéric Jurie. Unsupervised part learning for visual recognition. In *CVPR*, 2017. 5
[10] Ronan Sicre, Julien Rabin, Yannis Avrithis, Teddy Furon, Frédéric Jurie, and Ewa Kijak. Automatic discovery of discriminative parts as a quadratic assignment problem. In *ICCV Workshops*, 2017. 5
[11] Koen EA Van de Sande, Jasper RR Uijlings, Theo Gevers, and Arnold WM Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011. 5
[12] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *ECCV*, 2020. 6
[13] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *T-PAMI*, 2019. 6
[14] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, 2019. 6
[15] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *NeurIPS*, 2020. 6