Supplementary for Domain-aware Visual Bias Eliminating for Generalized Zero-Shot Learning

Shaobo Min[1], Hantao Yao[2], Hongtao Xie[1]*, Chaoqun Wang[1], Zheng-Jun Zha[1], and Yongdong Zhang[1]
[1]University of Science and Technology of China
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
{mbobo,cq14}@mail.ustc.edu.cn, hantao.yao@nlpr.ia.ac.cn, {htxie,zhazj,zhyd73}@ustc.edu.cn

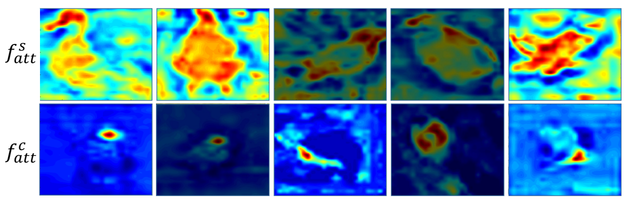

Figure 1. The learned attention maps from $f_{att}^s(\cdot)$ and $f_{att}^c(\cdot)$. For $f_{att}^c(\cdot)$, the feature map of maximum response in attention vector is selected for visualization.

1. Visualized Results of f_{att}^s and f_{att}^c

Both $f_{att}^s(\cdot)$ and $f_{att}^c(\cdot)$ give the improvements based on feature selections for the proposed DVBE. Some learned attention cases of $f_{att}^s(\cdot)$ and $f_{att}^c(\cdot)$ are given. Specifically, we visualize the inferred attention maps in Figure 1, where we find that the spatial attention $f_{att}^s(\cdot)$ focuses on localizing the foreground region, while the channel attention $f_{att}^c(\cdot)$ tends to localize local part regions. This proves the effectiveness of complementary feature selections from differnet attentions.

2. Definition for $\mathcal{L}_{cet}$

In Eq. (9) of the main text, $\mathcal{L}_{cet}$ is defined by:

$$\mathcal{L}_{cet}(f_v(\boldsymbol{x})) = -\sum_{\boldsymbol{x}\in\mathcal{X}_s} \log \frac{exp(W_{y*}^1 f_v(\boldsymbol{x}))}{\sum_{y\in\mathcal{Y}_s} exp(W_y^1 f_v(\boldsymbol{x}))}, \tag{1}$$

where W_y^1 is classifier weight for y, and $y*$ is the truth label.

3. Effects of Varying σ and γ

In Figure 2 (a), we evaluate the effects of different σ. As illustrated in the main text, σ is the hyper-parameter of adaptive margin λ:

$$\lambda = e^{-(p_y(\boldsymbol{x})-1)^2/\sigma^2}. \tag{2}$$

*Corresponding author.

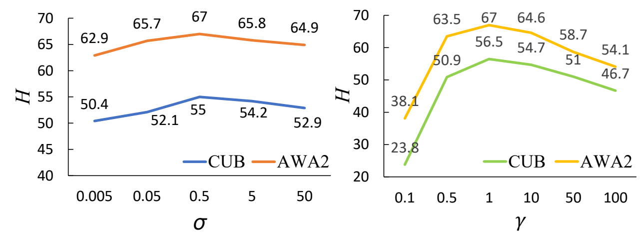

Figure 2. The effects of varying σ and γ on CUB and AWA2, respectively.

When σ becomes small, more samples will have small λ, which results in similar effects with the fixed margin case in [4]. Inversely, when σ becomes large, more samples have large λ (≈ 1), which approximates the standard Softmax. Therefore, as shown in Figure 2, we find that $\sigma = 0.5$ achieves a trade-off, which is suitable for all experimental datasets.

Since AMSE ($\mathcal{L}_{ams}$) and autoS2V ($\mathcal{L}_{s2v}+\mathcal{L}_{cet}$) are separately used with different outputs, they are not sensitive to loss weight. For autoS2V, we set $\mathcal{L}_{s2v}+\gamma\mathcal{L}_{cet}$, and varying γ. In Figure 2 (b), increasing γ from 0 to 0.5 will boost the performance obviously. The reason is that, $\mathcal{L}_{cet}$ can avoid both visual and semantic embeddings being $\mathbf{0}$ vector. When $\gamma > 10$, the visual embeddings cannot be well aligned with semantic labels, and the performance drops. When $\gamma \in [0.5, 10]$, the performance is stable.

4. Extension To Zero-Shot Semantic Segmentation

Since the proposed Domain-aware Visual Bias Eliminating (DVBE) network is a robust framework to biased recognition problem, it can be extended to more challenging zero-shot semantic segmentation [2]. The detailed architecture is shown in Figure 3. Different from classification, semantic segmentation needs to recognize all pixels separately in an image under Generalized Zero-Shot Learning (GZSL)

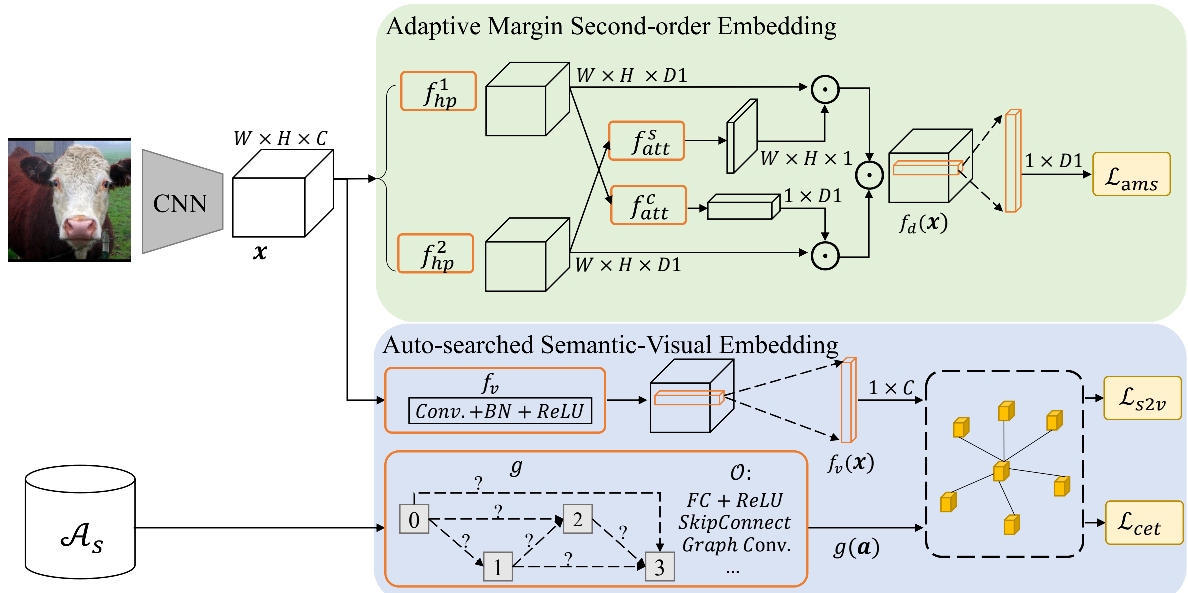

Figure 3. The extension of DVBE to zero-shot semantic segmentation task. The classifier in the AMSE is omitted for simplifying.

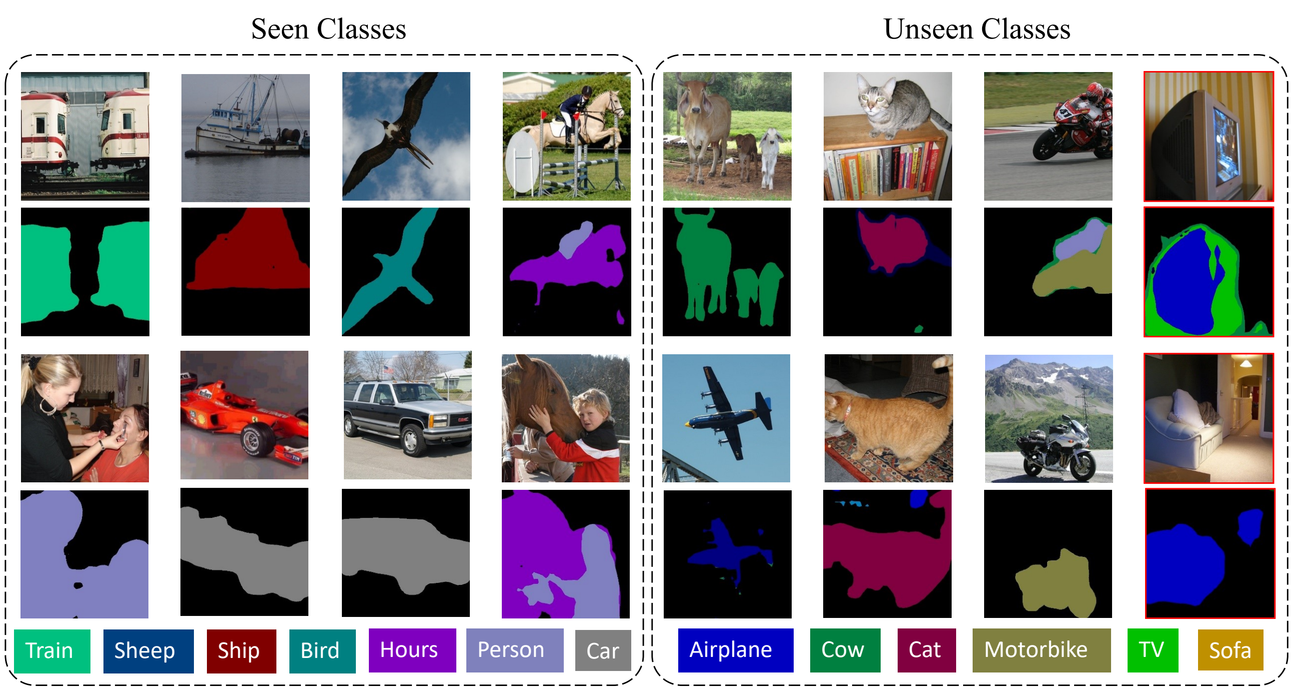

Figure 4. Some results for zero-shot semantic segmentation on Pascal VOC.

manner. Thus, the original AMSE of Eq. (5) in the main text should be modified, because $\otimes$ aggregates spatial feature vectors into a global one. To this end, we follow [10] by replacing $\{f_{rd}^1, f_{rd}^2\}$ and $\otimes$ in Eq. (5) with the high-dimension projections and Hadamard product $\odot$, which can approximate second-order interaction at each pixel by:

$$f_d(\boldsymbol{x}) = [f_{att}^s(\boldsymbol{x}_2) \odot \boldsymbol{x}_1] \odot [f_{att}^c(\boldsymbol{x}_1) \odot \boldsymbol{x}_2], \tag{3}$$

where $\boldsymbol{x}_1 = f_{hp}^1(\boldsymbol{x})$ and $\boldsymbol{x}_2 = f_{hp}^2(\boldsymbol{x})$ are two convolution layers to project $\boldsymbol{x}$ into different high-dimensional space $R^{N \times D_1}$, where $D_1 \gg C$. Here, $f_d(\boldsymbol{x}) \in R^{N \times D_1}$. $D1$ is set to 8192 in the experiment. Finally, the feature vector

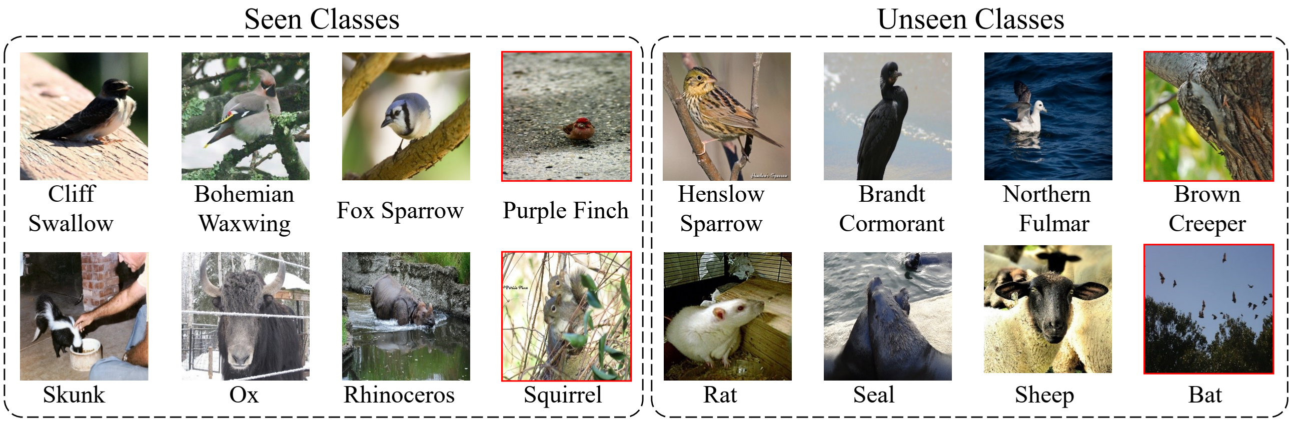

Figure 5. Some results for zero-shot recognition on CUB and AWA2. The red boxes indicate the incorrectly recognized cases.

at each pixel will be recognized by the classifier to generate a segmentation map.

5. More Visualized Results of DVBE

Some predicted samples by DVBE for zero-shot semantic segmentations and classification are given in Figure 4 and Figure 5. Specifically, for zero-shot semantic segmentation, the unseen classes include: "Airplane", "Cow", "Cat", "Motorbike", "TV", and "Sofa". From the results, it can be seen that the first four unseen classes can be well segmented, but the segmentations for "TV" and "Sofa" are dissatisfied. One possible reason is that the shape and appearance characteristics for the first four classes can be well described by word2vec [7], but the semantic descriptions for "TV" and "Sofa" are not good enough. In summary, the proposed DVBE is an effective framework for zero-shot learning with good generalization to both classification and semantic segmentation.

6. Improvement of ASME for Seen Class Recognition

ASME can significantly improve the feature discrimination, thus we further evaluate its improvement for seen class recognition. In Table 1, we use the baseline visual feature and discriminative f_d of AMSE to respectively recognize seen class samples, under a standard recognition setting. The training loss is standard Softmax, and the domain of testing sample is known in advance. From the results, AMSE can significantly improve the visual feature discrimination and obtain obvious gains on four datasets.

7. Conventional ZSL Results

We also give the results of DVBE under conventional zero-shot setting (CZSL) in Table 2. Under CZSL, only the unseen domain samples are evaluated. Since the sample domain is known before recognition in CZSL, the improvement of DVBE in CZSL is less obvious than that in GZSL. As we automatically search the optimal architecture for semantic-visual alignment, DVBE outperforms most of the previous methods.

Table 1. Improvement of AMSE for seen class prediction.

Methods	CUB	AWA2	aPY	SUN
Baseline	86.1	93.2	75.7	45.8
AMSE	90.2	96.1	78.3	53.6

Table 2. Conventional zeros-shot learning. The evaluation metric is MCA_u.

Methods	CUB	SUN	AWA2	aPY
FGN[9]	61.5	62.1	-	-
SE-ZSL[8]	59.6	63.4	69.2	-
PSR-ZSL[1]	56.0	61.4	63.8	38.4
CDL[5]	54.5	63.6	-	**43.0**
SP-AEN[3]	55.4	59.2	58.5	24.1
LDF[6]	70.4	-	-	-
DVBE	**74.3**	**65.7**	**71.7**	41.7

References

[1] Yashas Annadani and Soma Biswas. Preserving semantic relations for zero-shot learning. In *CVPR*, pages 7603–7612, 2018.

[2] Maxime Bucher, Tuan-Hung VU, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *NeurIPS*, pages 468–479, 2019.

[3] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR*, pages 1043–1052, 2018.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019.

[5] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Learning class prototypes via structure alignment for zero-shot recognition. In *ECCV*, pages 118–134, 2018.

[6] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Discriminative learning of latent features for zero-shot recognition. In *CVPR*, pages 7463–7471, 2018.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, pages 3111–3119, 2013.

[8] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, 2018.

[9] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018.

[10] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. Hierarchical bilinear pooling for fine-grained visual recognition. In *ECCV*, pages 574–589, 2018.