Supplementary for Progressive Semantic-Visual Mutual Adaption for Generalized Zero-Shot Learning

Man Liu[1,2], Feng Li[3], Chunjie Zhang[1,2], Yunchao Wei[1,2], Huihui Bai[1,2]*, Yao Zhao[1,2]
[1]Institute of Information Science, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Advanced Information Science and Network, Beijing, China
[3]Hefei University of Technology, Hefei, China
{manliu, cjzhang, yunchao.wei, hhbai, yzhao}@bjtu.edu.cn fengli@hfut.edu.cn

1. Results on ResNet101

In this part, we replace the visual backbone of PSVMA with ResNet101, and compare it with other ResNet101-based methods [4, 5, 7–9] under the setting of input size 224×224. Since the feature map size varies between different layers of ResNet101, we only deploy the DSVTM in the final visual layer (*i.e.,* $Z = 1$) to establish visual-semantic mutual adaption. As shown in Tab. 4, our PSVMA outperforms the best previous method by respectively 0.6%, 2.3%, and 3.0% on CUB, SUN, and AwA2 datasets, respectively. The SOTA performance on different backbones (ViT and ResNet101) demonstrates the effectiveness and generalization of our PSVMA.

Table 4. Results of GZSL on three public benchmarks using ResNet101 backbone with the input size 224×224. The best and second-best results are marked in red and blue, respectively.

Methods	CUB			SUN			AwA2		
	U	S	H	U	S	H	U	S	H
PREN [9]	32.5	55.8	43.1	35.4	27.2	30.8	32.4	88.6	47.4
LFGAA [5]	43.4	79.6	56.2	20.8	34.9	26.1	50.0	90.3	64.4
AREN [7]	63.2	69.0	66.0	40.3	32.3	35.9	54.7	79.1	64.7
DAZLE [4]	56.7	59.6	58.1	52.3	24.3	33.2	60.3	75.7	67.1
APN [8]	65.3	69.3	67.2	41.9	34.0	37.6	56.5	78.0	65.5
PSVMA (Ours)	66.2	69.4	67.8	47.0	34.6	39.9	62.8	79.5	70.1

2. Conventional ZSL Results

Table 5. Results of conventional ZSL. The best and second-best results are marked in red and blue, respectively.

Methods	CUB	SUN	AwA2
AREN [7]	71.8	60.0	67.9
APN [8]	72.0	61.6	68.4
GEM-ZSL [6]	77.8	62.8	67.3
TransZero [1]	76.8	65.6	70.1
MSDN [2]	76.1	65.8	70.1
DUET [3]	72.3	64.4	69.9
PSVMA (Ours)	78.2	72.4	77.8

*Corresponding author

Table 6. Effect of detailed components in ACA and PMA.

IASA	ACA		SRIA	PMA		CUB			AwA2		
	communication	activation		mixing	activation	U	S	H	U	S	H
			✓			63.5	71.11	67.1	63.2	75.6	69.3
			✓	✓		68.3	69.8	69.1	64.0	77.1	69.9
			✓		✓	68.1	68.9	68.5	63.3	76.7	69.4
			✓	✓	✓	70.0	70.0	70.0	65.0	77.3	70.6
✓			✓	✓	✓	70.0	72.8	71.3	71.1	78.1	74.5
✓	✓		✓	✓	✓	70.4	74.9	72.6	72.0	78.4	75.1
✓		✓	✓	✓	✓	70.4	73.0	71.7	71.6	78.7	75.0
✓	✓	✓	✓	✓	✓	70.1	77.8	**73.8**	73.6	77.3	**75.4**

Table 7. Effect of N_h on CUB and AwA2 datasets.

N_h	CUB			AwA2		
	U	S	H	U	S	H
392	68.3	77.7	72.7	70.5	77.1	73.7
512	70.1	77.8	**73.8**	73.6	77.3	**75.4**
588	69.4	76.9	72.9	68.8	78.8	74.0
1024	68.5	77.1	72.6	67.8	79.4	73.1

We provide the comparison results with recent methods under the conventional ZSL setting in Tab. 5. The top-1 accuracy of unseen classes is used as the evaluation metric. Our PSVMA achieves the best accuracy of 78.2%, 72.4%, and 77.8% on CUB, SUN and AwA2 datasets, respectively. This shows that PSVMA distills the transferable and discriminative representations for distinguishing unseen classes.

3. Additional Ablations

Effect of detailed components in ACA and PMA. We further evaluate the components in ACA and PMA, *i.e.,* attribute communication and activation, and the patch mixing and activation, respectively. The results are shown Tab. 6. We see that the ACA benefits from the communication and activation operations with the H gains of 1.3% and 0.4% on CUB, and 0.6% and 0.5% on AwA2, respectively. When both communication and activation operations are used simultaneously in ACA, our method achieves better results. Similarly, both patch mixing and activation operations in the PMA improve the recognition performance of the model.

Effect of N_h in PMA. N_h is the dimension of expanded patches in expansion layer $f_e(\cdot)$ that enlarges the length of visual patches from N_v. Here, we analyze the effectiveness of N_h as shown in Tab. 7. When $N_h = 512$, the best performance is obtained. This proves that the large expansion dimension prevents valid information from being filtered out by subsequent filtering layers. Nevertheless, a larger expansion dimension can introduce information redundancy and hamper the performance of recognition. Thus, we set $N_h = 512$ for the best results.

4. The progressive learning process.

Our PSVMA applies progressive learning to achieve accurate visual-semantic interaction and produce the unambiguous visual representation, which helps to improve the transferability for GZSL. In addition to the illustration in Fig. 2, here, we provide pseudocode (see Algorithm 1) to further explain progressive learning process for PSVMA.

Algorithm 1 Proposed PSVMA Method

Input : $F^l, S, \mathcal{A}$
Output : Predicted labels $\tilde{y}$
while $l \leq L$ **do**
  while r $\leq$ R **do**
    $S^{l,r} \leftarrow \text{IMSE}(S, F^l)$ `// instance-centric prototypes`
  end while
  $\hat{F}^l \leftarrow \text{SMID}(S^{l,R}, F^l)$ `// unambiguous visual representations`
  if $l \neq L$ **then**
    $F^{l+1} = Layer_{l+1}(\hat{F}^l)$ `//` $Layer$ `is the layer of ViT backbone`
  end if
end while
$\tilde{y} \leftarrow$ eqn (13,14,18) `// using` $\hat{F}^L$, $\mathcal{A}$

References

[1] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. Transzero: Attribute-guided transformer for zero-shot learning. In *AAAI*, 2022. 1

[2] Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao, and Xinge You. Msdn: Mutually semantic distillation network for zero-shot learning. In *CVPR*, 2022. 1

[3] Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, Yin Fang, Jeff Z Pan, Wenting Song, and Huajun Chen. Duet: Cross-modal semantic grounding for contrastive zero-shot learning. *arXiv preprint arXiv:2207.01328*, 2022. 1

[4] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, 2020. 1

[5] Yang Liu, Jishun Guo, Deng Cai, and Xiaofei He. Attribute attention for semantic disambiguation in zero-shot learning. In *ICCV*, 2019. 1

[6] 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. 1

[7] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *CVPR*, 2019. 1

[8] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020. 1

[9] Meng Ye and Yuhong Guo. Progressive ensemble networks for zero-shot recognition. In *CVPR*, 2019. 1