Decomposed Soft Prompt Guided Fusion Enhancing for Compositional Zero-Shot Learning – Supplementary Materials

The supplementary file provides more details for paper "Decomposed Soft Prompt Guided Fusion Enhancing for Compositional Zero-Shot Learning", including the following aspects:

1. Summary Statistics of Datasets.
2. Backbone Study.
3. Hyper-Parameters Analysis.
4. Pseudocode.
5. Qualitative Results.

1. Summary Statistics of Datasets

We analyse three datasets included MIT -States [2], UT-Zappos [5] and C-GQA [4] statistically and the summary statistics can be seen in Tab. 1. s and o denote the number of *state* and *object* concepts, and i represents the number of images. Also, c_s and c_u are the pair concepts of seen and unseen classes.

Table 1. Summary statistics of the datasets used in our experiments, including MIT-States, UT-Zappos and CGQA.

			Train		Validation			Test		
Dataset	s	o	c_s	i	c_s	c_u	i	c_s	c_u	i
MIT-States	115	245	1262	30338	300	300	10420	400	400	12995
UT-Zappos	16	12	83	22998	15	15	3214	18	18	2914
CGQA	453	870	6963	26920	1173	1368	7280	1022	1047	5098

2. Backbone Study

To evaluate the performance of DFSP with various backbone, we also retrain the model with same parameters ($\alpha = 0.01$, $\beta = 0.1$ and $K = 1$) and only replace the backbone with ViT-B/32, Vit-B/16 and ViT-L/14@336px due to the limitation of image encoder in DFSP which is based on transformer networks [1]. Metrics contists of *S*, *U*, *H* and *AUC* and the model (*DFSP(t2i)*) are tested on MIT-States and UT-Zappos with the settings of Closed-World (CW) and Open-World (OW). Meanwhile, we evaluate DFSP with 5 random seeds to report the standard error, which can be seen in Tab. 2.

From the results in Tab. 2, we can see that ViT-14/L and ViT-L/14@336px achieve state-of-the-art (SOTA) results both on Closed-World and Open-World. Meanwhile, all backbones work well on multiple metrics, especially on MIT-States, such as *AUC* %12.8 versus %5.5 on SCEN [3] with the setting of Closed-World.

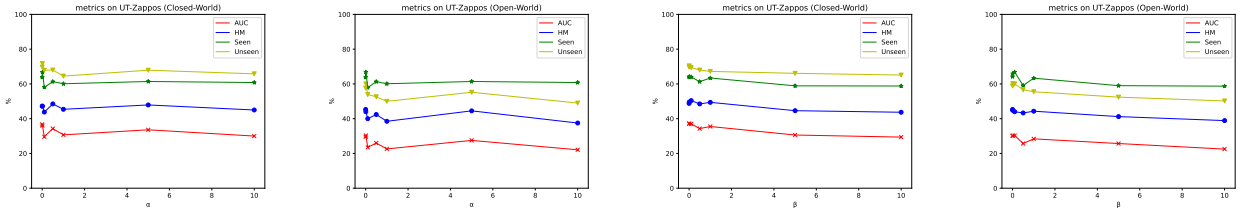
3. Hyper-Parameters Analysis

The loss function of DFSP is $\mathcal{L} = \mathcal{L}_{dfm} + \alpha\mathcal{L}_{st+obj} + \beta\mathcal{L}_{spm}$, which contains two hyper-parameters α and β. $\mathcal{L}_{dfm}$ is the final pair loss in DFM, $\mathcal{L}_{st+obj}$ is the decomposed features pair loss and $\mathcal{L}_{spm}$ is the pair loss before fusion in SPM. To evaluate the influence of them for DFSP(*t2i*), we show the hyper-parameters analysis in this section on MIT-States and UT-Zappos with the settings of Closed-World and Open-World. α and β are set to seven different orders of magnitude

Table 2. DFSP with different backbone results on MIT-States and UT-Zappos with the settings of Closed-World (CW) and Open-World (OW). We also report the standard error with 5 random seeds.

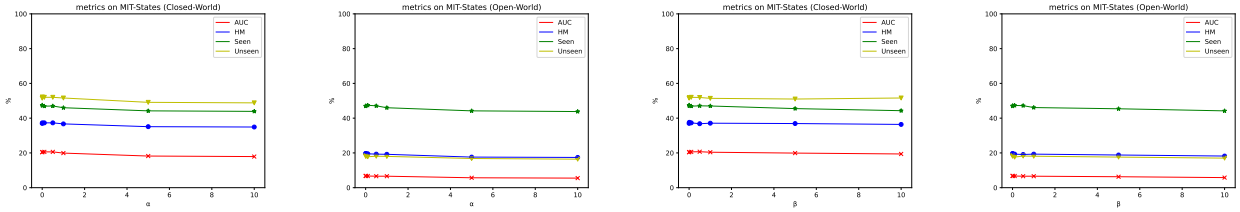
		MIT-States				UT-Zappos			
	Backbone	S	U	H	AUC	S	U	H	AUC
CW	ViT-B/32	36.7±0.25	43.4±0.37	29.4±0.18	13.2±0.69	55.8±1.66	59.5±2.25	38.5±2.72	23.3±1.32
	ViT-B/16	39.6±0.17	46.5±0.28	31.5±0.18	15.1±0.32	62.1±2.20	67.7±3.21	48.0±0.89	33.2±1.21
	ViT-L/14	46.8±0.54	52.2±0.17	37.4±0.35	20.6±0.28	65.2±1.80	70.7±1.51	50.4±2.14	36.8±0.72
	ViT-L/14@336px	45.6±0.21	50.8±0.38	36.9±0.57	19.9±0.32	64.6±1.39	70.6±1.35	48.3±1.89	36.9±0.65
OW	ViT-B/32	36.7±0.19	12.8±0.13	13.1±0.25	3.3±0.12	56.8±1.19	38.9±2.11	33.8±1.30	17.0±0.84
	ViT-B/16	39.6±0.11	15.2±0.16	15.4±0.09	4.4±0.14	62.9±1.30	48.4±1.88	42.9±2.02	25.4±0.93
	ViT-L/14	47.6±0.21	18.7±0.35	19.3±0.09	6.7±0.13	64.8±1.22	59.8±2.19	45.4±1.21	29.6±0.88
	ViT-L/14@336px	45.5±0.14	17.1±0.86	18.4±0.11	6.3±0.16	63.6±2.10	58.4±1.12	45.2±0.88	28.3±0.56

parameters, specifically $[0, 0.01, 0.1, 0.5, 1.0, 5.0, 10.0]$. The results are shown in Fig. 1 and Fig. 2. Since UT-Zappos consists of less *states* and *objects* in Tab. 1, the metrics consist of *S*, *U*, *H* and *AUC* are more sensitive to α and β, while metrics on MIT-States are going to be smoother. It can be concluded from the graph that the optimal selection range of α and β is $0 \sim 1$.



(a) Closed-World (α) (b) Open-World (α) (c) Closed-World (β) (d) Open-World (β)

Figure 1. Metrics on UT-Zappos with the settings of Closed-World and Open-World.



(a) Closed-World (α) (b) Open-World (α) (c) Closed-World (β) (d) Open-World (β)

Figure 2. Metrics on MIT-States with the settings of Closed-World and Open-World.

4. Pseudocode

We illustrate the core pseudocode in this section, including the decompose and recompose function. Given the pair_idx, we can obtain the att_idx and obj_idx, which could be utilized to decompose the state feature text_att and object feature text_obj. Additionally, the recompose function is reverse to decompose function and the code is shown as Code 1.

```python
def decompose(text_feature, pair_idx):
    t, l, c = text_feature.shape
    att_idx, obj_idx = pair_idx[:, 0].cpu().numpy(), pair_idx[:, 1].cpu().numpy()
    text_att = torch.zeros(t, self.attributes, c).cuda()
    text_obj = torch.zeros(t, self.classes, c).cuda()
    for i in range(self.attributes):
        text_att[:, i, :] = text_feature[:, np.where(att_idx==i)[0], :].mean(-2)
    for i in range(self.classes):
        text_obj[:, i, :] = text_feature[:, np.where(obj_idx==i)[0], :].mean(-2)
    text_feature_plus = torch.cat([text_att, text_obj], dim=1)
    return text_feature_plus


def recompose(text_feature_plus, pair_idx):
    t, l, c = text_feature.shape
    att_idx, obj_idx = pair_idx[:, 0].cpu().numpy(), pair_idx[:, 1].cpu().numpy()
    text_com_feature = torch.zeros(t, len(idx), c).cuda()
    text_com_feature = text_feature[:, att_idx, :] * text_feature[:, obj_idx + offset, :]
    return text_feature
```

Code 1: Decompose and Recompose

5. Qualitative Results

We report the top-1 qualitative results in the body of the paper. To better prove the effectiveness of DFSP, we show the top-3 qualitative results on MIT-States, UT-Zappos and CGQA in this section, which can be seen in Fig. 3. From the prediction results of top-3, it can be seen that even if top-1 has no successful cases, most of top-3 results can be predicted correctly. The compositions that the model has not seen can still be predicted correctly, which proves the generalization ability of the model to unseen concepts. Due to the abstract nature of state, it is more difficult to identify a state than an object, which can be seen more failure cases in Fig. 3 for states.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[2] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015. 1, 4

[3] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. Siamese contrastive embedding network for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9335, 2022. 1

[4] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962, 2021. 1, 4

[5] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 192–199, 2014. 1, 4

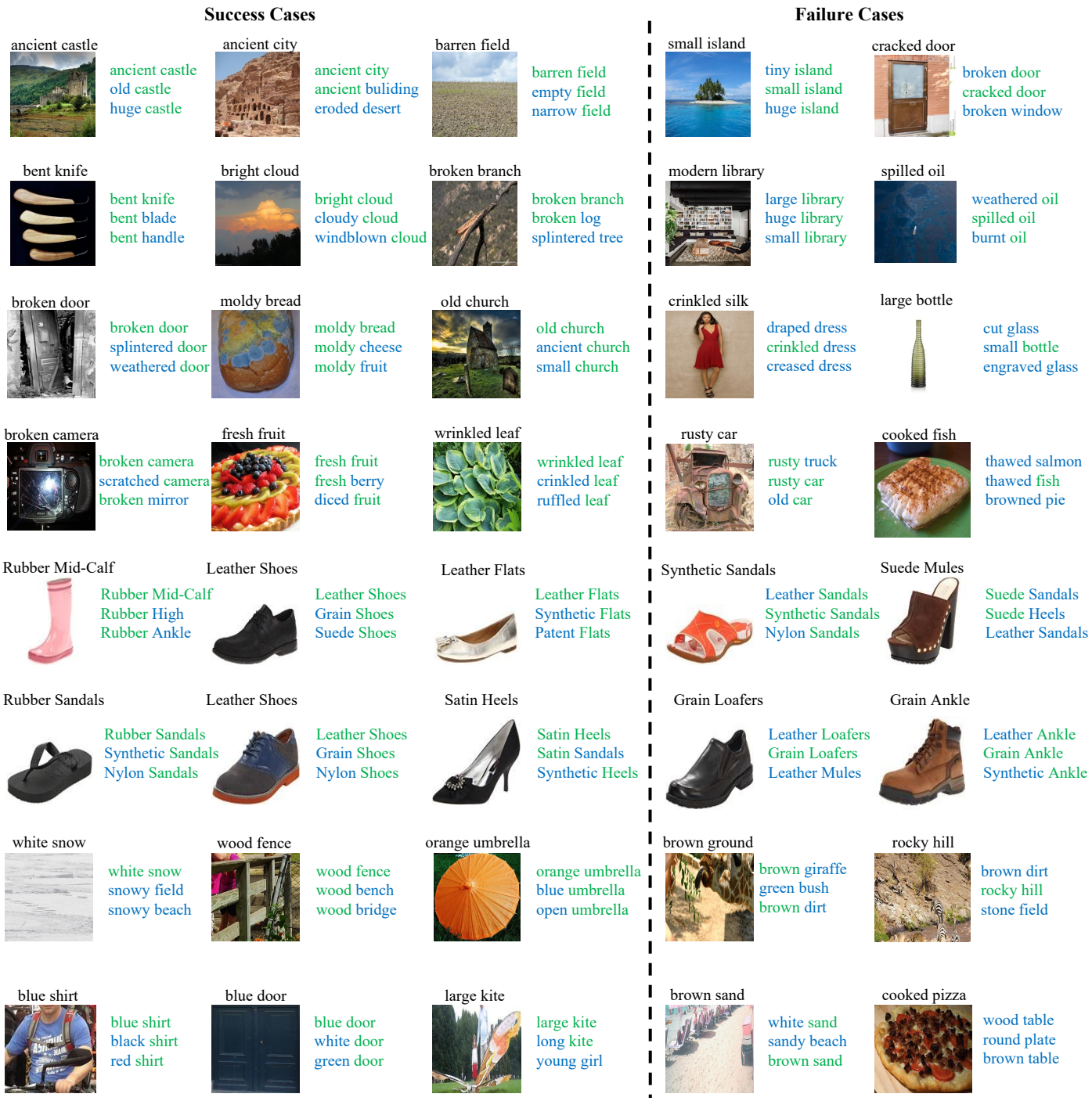


Figure 3. Qualitative top-3 results on MIT -States [2], UT-Zappos [5] and C-GQA [4]. blue denotes the wrong prediction and green represents the right case. The three columns on the left are success cases, and the two columns on the right are wrong cases for the top-1 prediction.