Appendix of "MixReorg: Cross-Modal Mixed Patch Reorganization is a Good Mask Learner for Open-World Semantic Segmentation"

Kaixin Cai[1*] Pengzhen Ren[1*] Yi Zhu[2] Hang Xu[2] Jianzhuang Liu[2]
Changlin Li[3] Guangrun Wang[4] Xiaodan Liang[1,5,6 †]
[1]Shenzhen Campus of Sun Yat-sen University [2]Huawei Noah's Ark Lab [3]University of Technology Sydney
[4]University of Oxford [5]MBZUAI [6]DarkMatter AI Research
caikx7@mail2.sysu.edu.cn, pzhren@foxmail.com, {zhuyi36, xu.hang, liu.jianzhuang}@huawei.com,
{changlinli.ai, wanggrun, xdliang328}@gmail.com

A. Training and Inference Detail

During training, we train MixReorg 16× Nvidia V100 GPUs. And we use the Adam optimizer with an initial learning rate 0.0016 and a weight decay of 0.05. Follow GroupViT, we train MixReorg for 30 epochs with 5 epochs contains linear warm-up.

During inference, following GroupViT, MixReorg gets masks from the attention maps of group tokens and predicts the foreground classes by the softmax-normalized-similarity between the embedding of the outputted image segments and the text segmentation labels while predicting the background class by thresholding the similarity. We resize each input image to a resolution 448×448. We set the thresholds on PASCAL VOC 2012, PASCAL Context, and COCO to 0.95, 0.35, and 0.95, respectively.

*Equal contribution.
†Corresponding author.

B. Additional Qualitative Results

In Figure 1, we show more semantic segmentation visualization results. It can be seen that MixReorg has better quality masks through mixed image mask prediction, which can provide fine-grained alignment information and guidance of grouping. In addition, MixReorg has a stronger ability for semantic alignment.

C. Comparison of Computing Cost

As shown in Table 1, we compare the number of parameters, training time, and inference time between MixReorg and GroupViT. The results showed that MixReorg does not have a significant advantage in terms of the number of parameters. Although MixReorg takes more time for training due to mixed images, their testing time is comparable.

Method	param. (M)	Training Time(h)	Inference Time(s/image)
GroupViT	28.7	54	0.14
MixReorg	30.5	60	0.16

Table 1: Comparison of computing cost.

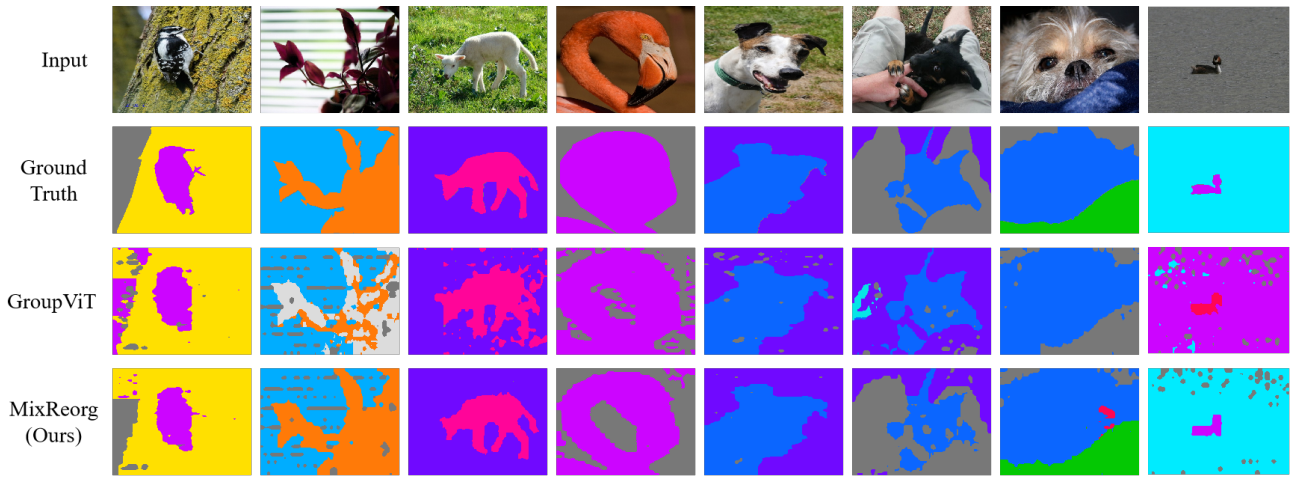

Figure 1: Comparison of semantic segmentation results on PASCAL Context.