Rethinking Prior Information Generation with CLIP for Few-Shot Segmentation

Jin Wang[1] Bingfeng Zhang[1*] Jian Pang[1] Honglong Chen[1] Weifeng Liu[1*]
[1]China University of Petroleum (East China)
{wangjin, jianpang}@s.upc.edu.cn, {bingfeng.zhang, wfliu, chenhl}@upc.edu.cn

1. Ablation Study on High-order Matrix

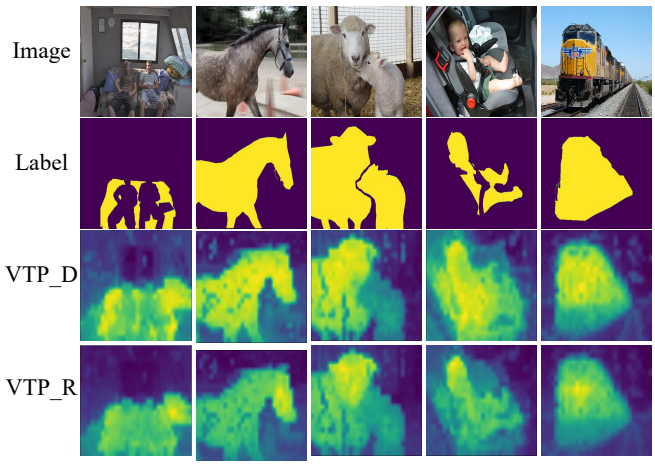

Figure 1. Qualitative results of the refinement of our designed high-order matrix. Each row from top to bottom represents the query images, query labels, refined VTP by initial matrix, and refined VTP by our proposed high-order matrix, respectively.

Table 1. Ablation study about our proposed high-order matrix refinement on the PASCAL-5^i, "D" represents the initial matrix, R represents the high-order matrix we designed.

D	R	mIoU (%)	FB-IoU (%)
✓		75.10	86.44
	✓	76.40	87.57

In order to verify the effectiveness of the higher-order matrix, we conduct ablation experiments on the effect of the higher-order matrix. As shown in Table 1, when utilizing our proposed high-order matrix to refine the prior information, it gets a better performance with 76.40 mIoU % which is 1.3 % higher than utilizing the original attention matrix, and the results of the two matrix fine-tuning methods are visualized in Figure. 1. It can be found that our proposed higher-order matrix refinement can indeed extract finer-grained prior information based on structure information and strengthen the suppression of chaotic background areas. Note that the experiments in this section follow the same setting as the experiments in the text.

*Corresponding author.

Table 2. Ablation study about text prompts on PASCAL-5^i, t_f represents the target prompt and t_b represents the non-target prompt.

t_f/t_b	mIoU (%)	FB-IoU (%)
"a photo of {*target class*}" "a photo without {*target class*}"	**76.40**	**87.57**
"a photo of {*target class*}" "not a photo of {*target class*}"	71.32	83.78
"a clean origami of {*target class*}" "a clean origami different from {*target class*}"	73.97	85.64

2. Ablation Study on Text Prompts

Different textual prompts directly decide the model's understanding of image visual content, which further affects the quality of the generated prior information, based on this, we conduct ablation experiments on the content of the text prompts, as shown in Table 2, the highest performance of 76.40% was achieved when using the text prompts we designed.

3. More Visualizations

To further demonstrate the effect of our model, we show more visualizations on the COCO-20^i [3], Figure. 2 compares more prior masks, and Figure. 3 provides a further visualization of the final experimental results.

4. CLIP Capacity Concern: Fair Comparisons

Not only CLIP pretraining, but also ImageNet pretraining may involve the tested novel classes. How to extract accurate features from the pretraining model and transfer them to few-shot segmentation is a more important issue than the size or capacity of the pretraining model. Besides, the CLIP model has been used in the few-shot task, *e.g.*, CLIPSeg [2], TIP-Apapter [5].

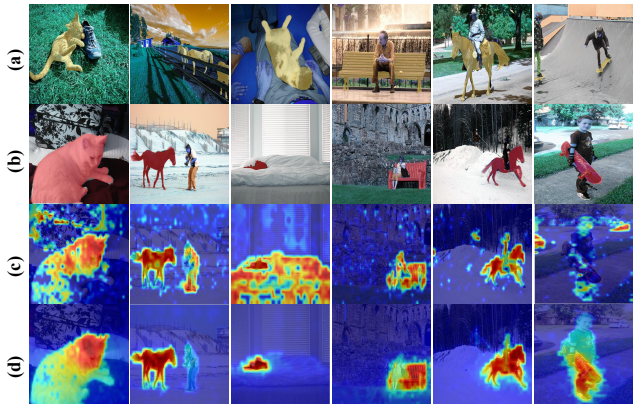

Figure 2. Comparison of more prior information from COCO-20^i [3]. (a) Support images with ground-truth masks; (b) Query images with ground-truth masks; (c) Prior information from previous approaches generated based on the frozen ImageNet [1] weights.

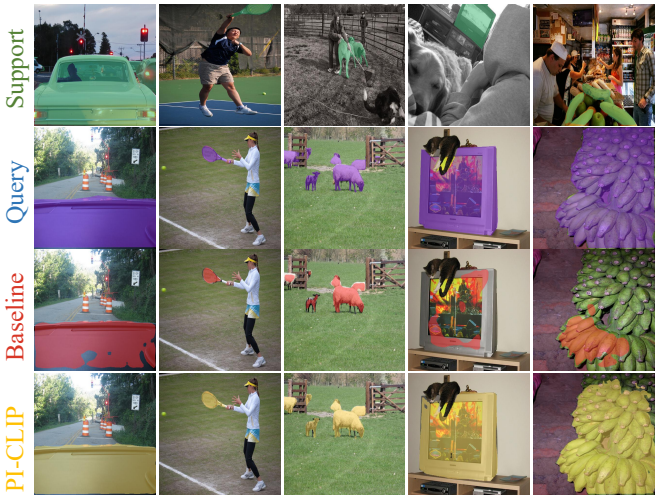

Figure 3. Qualitative results of the proposed PI-CLIP and baseline (HDMNet [4]) approach under 1-shot setting from COCO-20^i [3]. Each row from top to bottom represents the support images with ground-truth (GT) masks (green), query images with GT masks (purple), baseline results (red), and our results (yellow), respectively.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[2] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 1

[3] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 622–631, 2019. 1, 2

[4] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23641–23651, 2023. 2

[5] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer, 2022. 1