CLIMS: Cross Language Image Matching for Weakly Supervised Semantic Segmentation

Jinheng Xie, Xianxu Hou, Kai Ye, Linlin Shen
School of Computer Science & Software Engineering, Shenzhen University, China
Shenzhen Institute of Artificial Intelligence of Robotics of Society, Shenzhen, China
Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, China
National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, China
{xiejinheng,yekai}2020@email.szu.edu.cn, hxianxu@gmail.com, llshen@szu.edu.cn

A. Appendix

A.1. Visual Results

Figure 1 presents examples of predicted masks of PuzzleCAM and our CLIMS for the class of train. It can be observed in the second column that, PuzzleCAM [2] usually falsely discriminates the railroad as train, which leads to an overestimation of the target region of train. By contrast, our CLIMS successfully segment the correct regions of train, without the involvement of railroad.

We also present examples of masks predicted by PuzzleCAM and our CLIMS for the class of boat and person in Figure 2. It can be obviously seen that, compared with PuzzleCAM, less false positives of water and more complete regions of person are predicted in the semantic masks of our CLIMS.

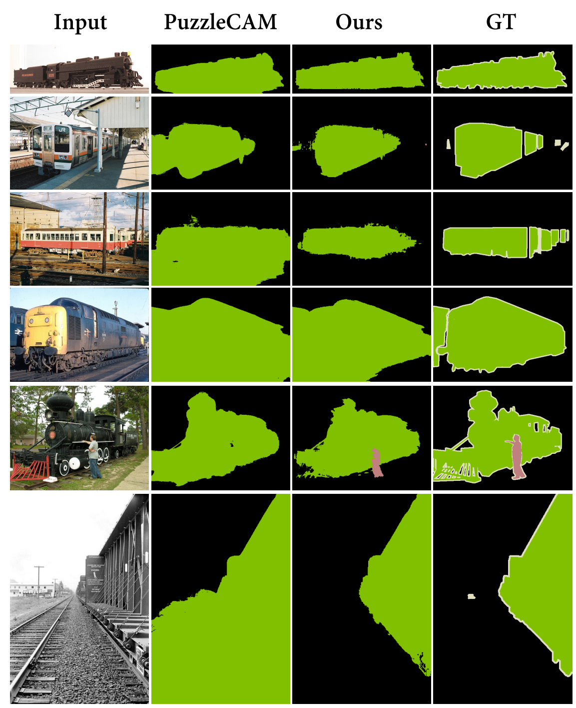

Figure 1. Examples of predicted semantic masks for the class of train on PASCAL VOC2012 *val* set. Best viewed in color.

A.2. Semi-supervised Semantic Segmentation

The generated pseudo ground-truth masks can be further used in the semi-supervised semantic setting. We try to use the generated masks to improve the performance of the semi-supervised method CCT [5]. Table 1 presents the performance comparison with those recent semi-supervised methods. It can be observed that with our masks, CCT can obtain a significant improvement of 3.7% mIoU on the PASCAL VOC2012 *val* set.

Table 1. Comparison of semi-supervised semantic segmentation methods on PASCAL VOC2012 *val* and *test* set.

Method	Training set		*val*	*test*
	Pixel-level	Image-level		
WSSL [6]	1.5K	9.1K	64.6	66.2
GAIN [4]	1.5K	9.1K	60.5	-
MDC [9]	1.5K	9.1K	65.7	67.6
DSRG [1]	1.5K	9.1K	64.3	-
Souly *et al.* [8]	1.5K	9.1K	65.8	-
FickleNet [3]	1.5K	9.1K	65.8	-
CCT [5]	1.5K	9.1K	73.2	-
CCT + CLIMS	1.5K	9.1K	**76.9**	**75.8**

A.3. Limitation

CLIMS leverages the power of CLIP [7] to recognize those diverse backgrounds. However, as the text descriptions of objects in the segmentation dataset, i.e., PASCAL VOC2012, may be different from that used to train CLIP, many objects like person cannot be identified well using CLIP. In this work, we must finetune the CLIP model using the text descriptions of PASCAL VOC to mitigate the above issue.

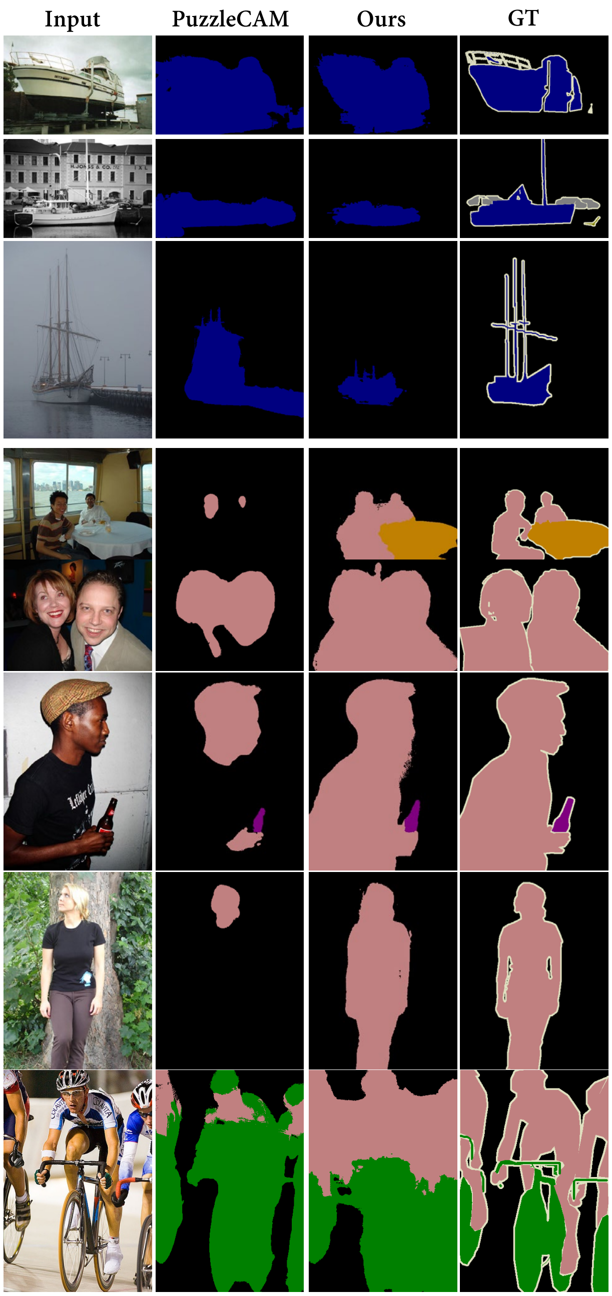

Figure 2. Examples of predicted semantic masks for the class of boat and person on PASCAL VOC2012 *val* set. Best viewed in color.

References

[1] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, pages 7014–7023, 2018. 1

[2] Sanghyun Jo and In-Jae Yu. Puzzle-cam: Improved localization via matching partial and full features. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 639–643, 2021. 1

[3] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*, pages 5267–5276, 2019. 1

[4] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *CVPR*, 2018. 1

[5] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, pages 12671–12681, 2020. 1

[6] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, pages 1742–1750, 2015. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1

[8] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *ICCV*, pages 5689–5697, 2017. 1

[9] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *CVPR*, pages 7268–7277, 2018. 1