Supplementary Material
Enhanced Soft Label for Semi-Supervised Semantic Segmentation

Jie Ma[1] Chuan Wang[1] Yang Liu[1] Liang Lin[1] Guanbin Li[1,2*]
[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2] Research Institute, Sun Yat-sen University, Shenzhen, China
majie25@mail2.sysu.edu.cn, cwang.hkucs@gmail.com, liuy856@mail.sysu.edu.cn,
linliang@ieee.org, liguanbin@mail.sysu.edu.cn

1. More Experimental results

1.1. Additional Quantitative Result

Table. 1 further compares our ESL with other available state-of-the-art methods based on ResNet-50 [3] backbone. As can be observed, our ESL also achieves the best performance under all partition protocols on both PASCAL VOC [2] and Cityscapes [1]. Besides, our ESL achieves consistent performance gains with ResNet-50 and ResNet-101, demonstrating its effectiveness to the different backbones.

Considering the seemingly saturated performance on the Pascal and Cityscapes, it is practical to evaluate on the more challenging dataset. Therefore, in Table. 2, we further list the comparison results on the COCO [5] dataset, which is a large-scale dataset for semantic segmentation, containing 118k training images and 5k validation images. As seen, our ESL significantly surpasses all available methods.

1.2. Hyperparameter Analysis of λ_1 and λ_2

Table. 3 summarizes the influence of hyperparameter λ_1 and λ_2, which controls the weight of $\mathcal{L}_{\mathrm{SCE}}$ and $\mathcal{L}_{\mathrm{CTR}}$, respectively. As can be seen, the model is not sensitive to the coefficients, and achieves the best performance at $\lambda_1 = 0.2$, $\lambda_2 = 0.1$.

1.3. Additional Qualitative Result

Fig. 1 provides more part grouping results under full partition protocol training set on *classic* Pascal VOC [2] and $1/2$ partition protocol training set on Cityscapes [1]. As shown in the figure, for some structured categories like *person*, the class region can be grouped into several meaningful parts, e.g., *person-head*, *person-body*, *person-leg* and etc. As for other less-structured categories like *tvmontior*, the division process is mainly based on spatial location. Thanks to the designed unsupervised object-part grouping mechanism, we can conduct more faithful pixel-to-part contrastive learning.

Fig. 2 shows more comparison results on both Pascal VOC [2] and Cityscapes [1]. Benefiting from our dynamic soft label and pixel-to-part contrastive learning, the ESL can handle more complex scenarios and generally provides more accurate segmentation results than state-of-the-art methods U^{2}PL [9] and PSMT [6].

*Corresponding author

Method	1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)	Full (1464)
Sup Baseline	44.03	52.26	61.65	66.72	71.43
PseudoSeg [12]	54.89	61.88	64.85	70.42	71.00
PC^{2}Seg [11]	56.90	64.63	67.61	70.90	72.26
ESL	**61.74**	**69.50**	**72.63**	**74.69**	**77.11**

(a) mIoU on *classic* Pascal VOC

Method	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
Sup Baseline	63.72	68.49	72.46	75.14
MT [8]	66.77	70.78	73.22	75.41
CCT[7]	65.22	70.87	73.43	74.75
GCT [4]	64.05	70.47	73.45	75.20
ST++ [10]	72.60	74.40	75.40	-
PSMT [6]	72.83	75.70	76.43	77.88
ESL	**73.41**	**75.86**	**76.80**	**78.02**

(b) mIoU on *blender* Pascal VOC

Method	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
Sup Baseline	63.34	68.73	74.14	76.62
MT [8]	66.14	72.03	74.47	77.43
CCT [7]	66.35	72.46	75.68	76.78
GCT [4]	65.81	71.33	75.30	77.09
PSMT [6]	-	75.76	76.92	77.64
ESL	**71.07**	**76.25**	**77.58**	**78.92**

(c) mIoU on Cityscapes

Table 1. Comparison with existing methods on *classic* (a) and *blender* (b) PASCAL VOC and Cityscapes (c) validation set based on ResNet-50 backbone under various partition protocols. In (a)(b)(c), "Sup Baseline" represents supervised training without unlabeled data, and "-" means the corresponding method doesn't report the result. The best values are marked in bold.

Method	Backbone	1/128 (925)	1/64 (1849)	1/32 (3697)
Sup Baseline	Xception-65	33.60	37.80	42.24
PseudoSeg [12]	Xception-65	39.11	41.75	43.64
PC^{2}Seg [11]	Xception-65	40.12	43.67	46.05
ESL	Xception-65	**44.37**	**47.14**	**49.25**

Table 2. The comparison results on the COCO dataset. The best values are marked in bold.

λ_1	0.01	0.05	0.1	**0.2**	0.5
mIoU	81.37	81.42	81.50	**81.77**	81.46
λ_2	0.01	0.05	**0.1**	0.2	0.5
mIoU	81.44	81.52	**81.77**	81.34	81.18

Table 3. Ablation Study on λ_1 and λ_2 under full partition protocol on *classic* Pascal VOC [2].

2. Limitation and Future Work

Although the method proposed in this paper shows very superior performance, for the sake of simplicity, we uniformly preset the number of prototypes of each category to a fixed value in the pixel-to-subregion comparative learning module. Since objects of different categories have diverse structures, they are suitable for different parsing and segmentation. We believe that incorporating category-adaptive cluster number can further improve the performance of the model and obtain a more reasonable presentation of the internal results of the model.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *computer vision and pattern recognition*, 2016.

[2] Mark Everingham, S. M. Eslami, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 2015.

[3] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[4] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 429–445. Springer, 2020.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Figure 1. Qualitative results of subregion division with subclass prototype number $K = 5$. The different color represents different subregions. Top: Pascal VOC [2]. Bottom: Cityscapes [1].

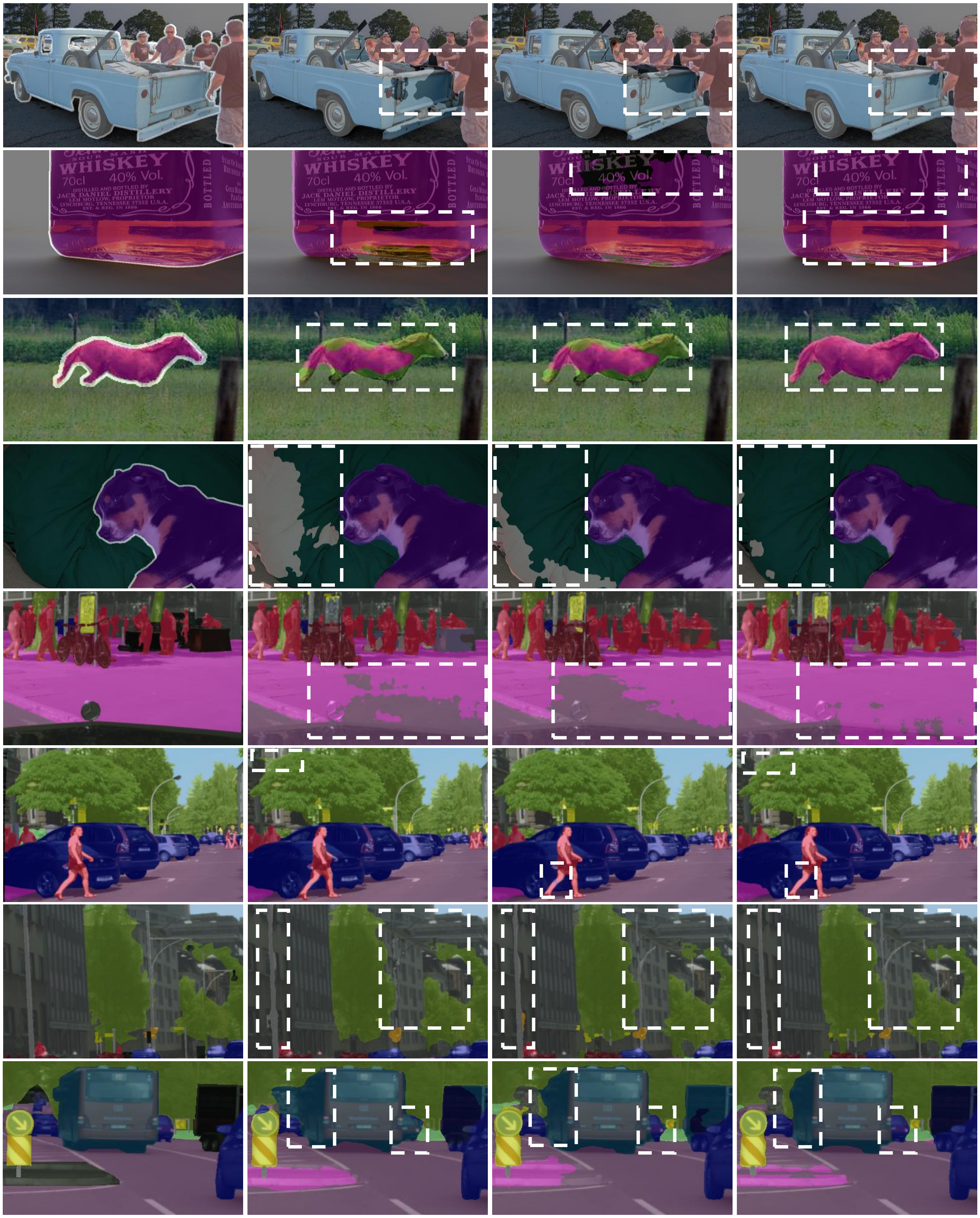
[6] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4258–4267, 2022.

[7] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

[8] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[9] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022.

[10] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Con-*



(a) Ground Truth **(b) PSMT** **(c) U^{2}PL** **(b) ESL**

Figure 2. Qualitative results of U^{2}PL [9], PSMT [6] and our ESL. Top: Pascal VOC [2]. Bottom: Cityscapes [1].

ference on Computer Vision and Pattern Recognition, pages 4268–4277, 2022.

[11] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021.

[12] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations*.