Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels Supplementary Material

Yuchao Wang[1†] Haochen Wang[1†] Yujun Shen[2] Jingjing Fei[3]
Wei Li[3] Guoqiang Jin[3] Liwei Wu[3] Rui Zhao[1,3‡] Xinyi Le[1*]

[1]Shanghai Jiao Tong University [2]The Chinese University of Hong Kong [3]SenseTime Research

{44442222, wanghaochen0409, lexinyi}@sjtu.edu.cn shenyujun0302@gmail.com
{feijingjing1, liwei1, jinguoqiang, wuliwei, zhaorui}@sensetime.com

A. Overview

We organize the Supplementary Material as follows. Above all, more details for reproducing the results will be given in Sec. B. Then we will give more results on Cityscapes from two perspectives in Sec. C. We also provide an alternative of contrastive learning to prove our main insight does not only rely on contrastive learning in Sec. D. Besides, ablation studies on both PASCAL VOC 2012 and Cityscapes for more hyper-parameters are given in Sec. E. Finally, visualization on feature space gives a visual proof for the effectiveness of U^{2}PL in Sec. F.

B. More Details for Reproducibility

For Cityscapes [2], we utilize OHEM which is the same as previous methods [1, 4]. The temperature τ is set to 0.5 for both PASCAL VOC 2012 [3] and Cityscapes [2]. We use SGD optimizer for all experiments. For experiments in PASCAL VOC 2012 [3], the initial base learning rate is 0.001 and the weight decay is 0.0001. For experiments in Cityscapes [2], the initial base learning rate is 0.01 and the weight decay is 0.0005. In our experiments, we find if we train the model only with supervised loss for the initial a few epochs then apply U^{2}PL, it can achieve better performance. We define such epoch as the warm start epoch, and the corresponding warm start epochs for PASCAL VOC 2012 and Cityscapes are 1 and 20 respectively.

To prevent overfitting, we apply random cropping, random horizontal flipping, and random scaling with the range of $[0.5, 2.0]$ for both PASCAL VOC 2012 [3] and Cityscapes [2] following previous methods [1, 4, 9, 10]. Our memory queue is category-specific. For the background category, the length of the queue is set to be $50,000$. For foreground categories, the length of the queue is all $30,000$. All baselines *i.e.*, "SupOnly", "MT", and "CutMix" are re-implemented by ourselves, where the only difference between "MT" and "CutMix" is that the latter applies CutMix [8] augmentation for unlabeled images.

The hyper-parameters used in this work are listed in Tab. A1. Among them, M, N, δ_p are used for contrastive learning, for which we simply follow [6]. λ_c, η, τ are training-related, while α_0, r_l, r_h are additionally introduced by our U^{2}PL.

Table A1. **Summary of hyper-parameters** used in U^{2}PL.

Symbol	Description	Default Value
(M, N)	contrastive learning settings	(50, 256)
δ_p	confidence threshold of positive samples	0.3
(λ_c, η)	loss weights	(0.1, 1)
τ	loss temperature	0.5
α_0	initial proportion of unreliable pixels	20%
(r_l, r_h)	probability rank thresholds	(3, 20)

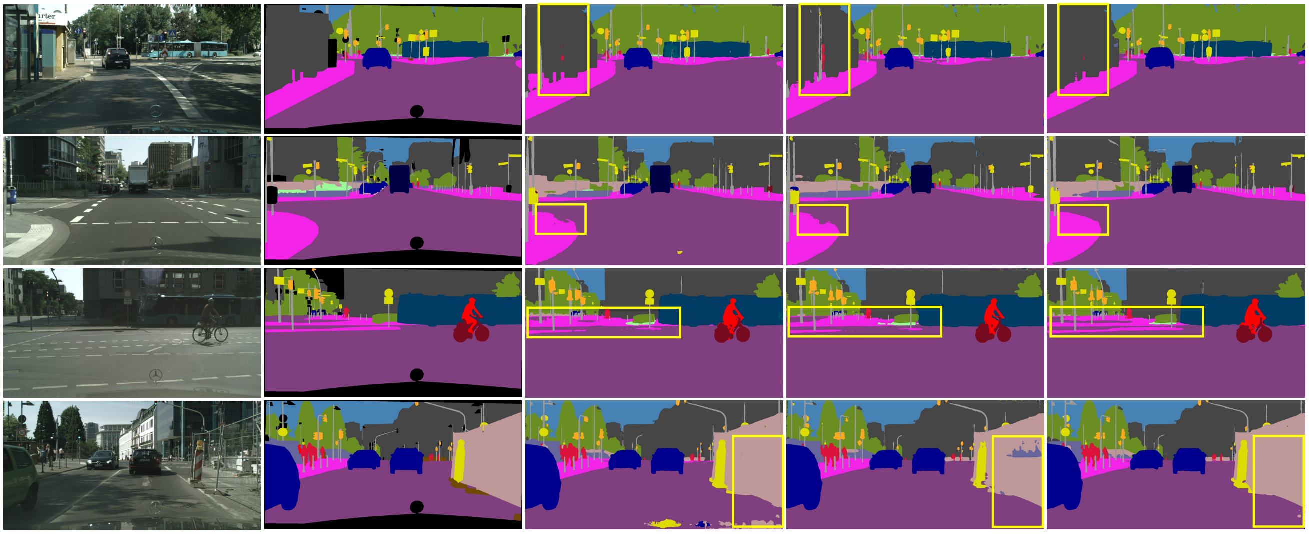
C. More Results on Cityscapes

Quantitative Results. Tab. A2 demonstrates the mIoU results on Cityscapes `val` set. "Unreliable" outperforms other options, proving using unreliable pseudo-labels does help. U^{2}PL fully mines the information of all pixels.

Qualitative Results. Fig. A1 shows the results of different methods on the Cityscapes `val` set. Benefiting by using unreliable pseudo-labels, U^{2}PL outperforms other methods. Note that using contrastive learning without filtering those unreliable pixels, sometimes does harm to the model (see the 1-st row and the 4-th row in Fig. A1), leading to worse results than those when the model is trained only by labeled data. Such visual difference proves that our method finally makes the reliability of unreliable prediction labels stronger.

D. Alternative of Contrastive Learning

Our proposed U^{2}PL is not limited by contrastive learning. Binary classification is also a sufficient way to use unreliable pseudo-labels, *i.e.*, using binary cross-entropy



(a) Image (b) Ground-Truth (c) Supervised Only (d) Plain $\mathcal{L}_c$ (e) U^{2}PL

Figure A1. Qualitative results on **Cityscapes** `val` set. All models are trained under the $1/2$ partition protocol, which contains $1,488$ labeled images and $1,487$ unlabeled images. (a) Input images. (b) Hand-annotated labels for the corresponding image. (c) *Only* labeled images are used for training. (d) The vanilla contrastive learning framework, where all pixels are used as negative samples without entropy filtering. (e) Predictions from our U^{2}PL. Yellow rectangles highlight the promotion by adequately using unreliable pseudo-labels.

Table A2. **Ablation study on using pseudo pixels with different reliability**, which is measured by the entropy of pixel-wise prediction. "Unreliable" denotes selecting negative candidates from pixels with top 20% highest entropy scores. "Reliable" denotes the bottom 20% counterpart. "All" denotes sampling regardless of entropy. We prove this effectiveness under $1/2$ and $1/4$ partition protocol on Cityscapes `val` set.

	Unreliable	Reliable	All
1/2 (1488)	**79.05**	77.19	76.96
1/4 (744)	**76.47**	75.16	74.51

loss (BCE) $\mathcal{L}_b$ other than contrastive loss. For i-th anchor $\mathbf{z}_{ci}$ belongs to class c, we simply use its negative samples $\{\mathbf{z}_{cij}^-\}_{j=1}^N$ and positive sample $\mathbf{z}_c^+$ to compute the BCE loss:

$$\mathcal{L}_b = -\frac{1}{C \times M \times N} \sum_{c=0}^{C-1} \sum_{i=1}^{M} \sum_{j=1}^{N} \log \left[\frac{e^{\langle \mathbf{z}_{ci}, \mathbf{z}_c^+ \rangle / \tau}}{e^{\langle \mathbf{z}_{ci}, \mathbf{z}_c^+ \rangle / \tau} + e^{\langle \mathbf{z}_{ci}, \mathbf{z}_{cij}^- \rangle / \tau}} \right], \tag{1}$$

where C, M, and N are the total number of classes, anchor pixels, and negative samples, respectively. $\langle \cdot, \cdot \rangle$ is the cosine similarity of two features, and τ represents the temperature.

Tab. A3 and Tab. A4 are results of using unreliable pseudo-labels based on binary classification on Cityscapes [2] and PASCAL VOC 2012 [3] `val` set respectively. From Tab. A3 and Tab. A4, we can tell that our U^{2}PL is not restricted by contrastive learning, a basic binary classification also does help. On Cityscapes `val` set,

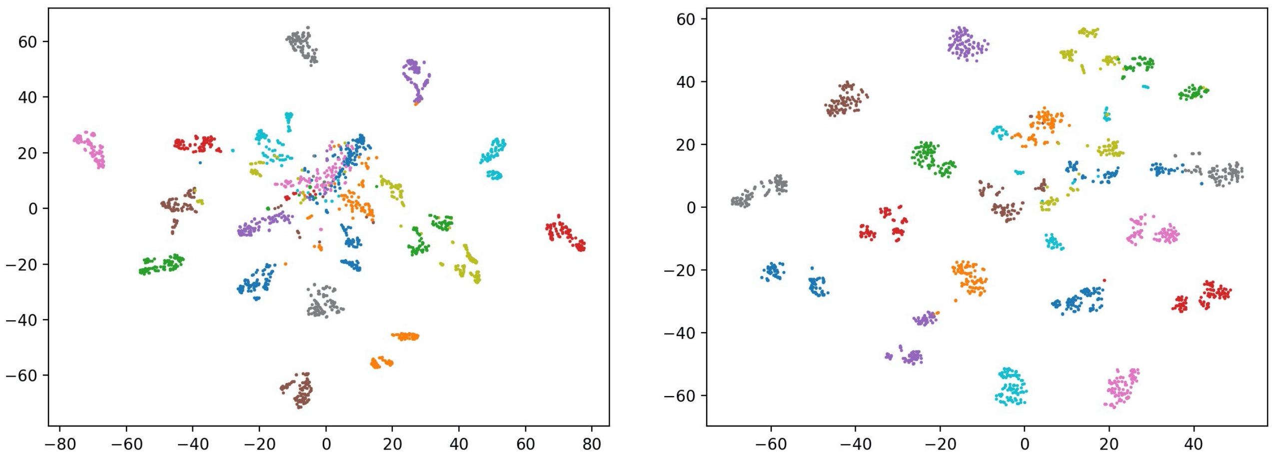
Table A3. Using unreliable pseudo-labels based on binary classification on **Cityscapes** `val` set under different partition protocols.

Method	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
SupOnly	65.74	72.53	74.43	77.83
MT [7]	69.03	72.06	74.20	78.15
U^{2}PL (w/ $\mathcal{L}_c$)	**70.30**	**74.37**	**76.47**	**79.05**
U^{2}PL (w/ $\mathcal{L}_b$)	69.87	72.93	75.91	78.36

Table A4. Using unreliable pseudo-labels based on binary classification on **PASCAL VOC 2012** `val` set under different splits.

Method	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
SupOnly	67.87	71.55	75.80	77.13
MT [7]	70.51	71.53	73.02	76.58
U^{2}PL (w/ $\mathcal{L}_c$)	**77.21**	**79.01**	79.30	**80.50**
U^{2}PL (w/ $\mathcal{L}_b$)	75.36	76.62	**79.64**	79.80

U^{2}PL with $\mathcal{L}_b$ can outperforms supervised only baseline by $+3.77\%$, $+0.40\%$, $+1.48\%$, and $+0.53\%$ under $1/16$, $1/8$, $1/4$, and $1/2$ partial protocols. U^{2}PL with $\mathcal{L}_b$ can outperforms supervised only baseline by $+7.49\%$, $+5.07\%$, $+3.84\%$, and $+2.67\%$ under $1/16$, $1/8$, $1/4$, and $1/2$ partial protocols on PASCAL VOC 2012 `val` set. Note that under the $1/4$ partition protocol of *blender* PASCAL VOC 2012, the bianry classification based U^{2}PL (w/ $\mathcal{L}_b$) outperforms the contrastive learning based U^{2}PL (w/ $\mathcal{L}_c$) by $+0.34\%$, which proves that contrastive learning is not the only efficient way of using unreliable pseudo-labels.



(a) Supervised Only

(b) U^{2}PL

Figure A2. **Visualization of the feature spaces** learned by our U^{2}PL and its supervised counterpart, using t-SNE [5]. The training set is the 1/4 partition protocol (2646) in *blender* VOC PASCAL 2012 Dataset.

Table A5. **Ablation study on base learning rate** under 1/4 partition protocol (2646) in *blender* VOC PASCAL 2012 Dataset.

lr_{base}	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
mIoU	3.49	77.82	**79.30**	74.58	65.69

Table A6. **Ablation study on temperature** under 1/4 partition protocol (2646) in *blender* VOC PASCAL 2012 Dataset.

τ	10	1	0.5	0.1	0.01
mIoU	78.88	78.91	**79.30**	79.22	78.78

Table A7. **Ablation study on PRT** on Cityscapes `val` set.

r_l	1	1	3	3	10
r_h	3	20	10	20	20
1/8 (372)	71.41	72.08	72.60	**74.37**	72.24
1/4 (744)	76.27	76.04	76.01	**76.47**	76.18

Table A8. **Ablation study on** α_0 on Cityscapes `val` set.

α_0	40%	30%	20%	10%
1/8 (372)	72.07	72.93	**74.37**	71.63
1/4 (744)	75.20	76.08	**76.47**	76.40

E. More Ablation Studies

E.1. More Hyper-parameters on VOC

Base Learning Rate. The impact of the base learning rate is shown in Tab. A5. Results are based on U^{2}PL on *blender* VOC PASCAL 2012 Dataset. We find that 0.001 outperforms other alternatives.

Temperature. Tab. A6 gives a study on the effect of temperature τ. Temperature τ plays an important role to adjust the importance to hard samples When $\tau = 0.5$, our U^{2}PL achieves best results. Too large or too small of τ will have an adverse effect on overall performance.

E.2. Ablation Studies on Cityscapes

Probability Rank Threshold. Tab. A7 provides a verification that such balance promotes the performance. $r_l = 3$ and $r_h = 20$ outperform other options by a large margin.

Initial Reliable-Unreliable Partition. Tab. A8 studies the impact of different α_0. When $\alpha_0 = 20\%$, the model achieves the best performance.

F. Visualization on Feature Space

To have a better understanding of U^{2}PL, we give an illustration on visualization of feature space. Two t-SNE [5] plots are given respectively on the supervised only method and U^{2}PL.

We can observe from Fig. A2 that decision boundaries of features generated by the supervised only method are quite confusing, while U^{2}PL has much more clear ones. This explains why U^{2}PL works from a feature point of view.

References

[1] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1

[2] 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. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1, 2

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 1, 2

[4] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. In *Adv. Neural Inform. Process. Syst.*, 2021. 1

[5] Geoffrey Hinton Laurens Van der Maaten. Visualizing data using t-sne. In *JMLR*, 2008. 3

[6] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew J Davison. Bootstrapping semantic segmentation with regional contrast. *arXiv preprint arXiv:2104.04465*, 2021. 1

[7] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Adv. Neural Inform. Process. Syst.*, 2017. 2

[8] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Int. Conf. Comput. Vis.*, 2019. 1

[9] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Int. Conf. Comput. Vis.*, 2021. 1

[10] 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 *Int. Conf. Learn. Represent.*, 2020. 1