Supplementary Material for Re-distributing Biased Pseudo Labels for Semi-supervised Semantic Segmentation: A Baseline Investigation

Outline

In this supplementary file, we first provide more results on Cityscapes in Sec. S1: parameter analysis in Sec. S1.1 to investigate the effects of labeling ratio and data augmentation magnitude; explanation of our consideration for data augmentation in the progressive strategy in Sec. S1.2; visualizations of pseudo labels generated by different methods in Sec. S1.3; the qualitative results of different methods in Sec. S1.4. Moreover, we provide additional experimental results for the semi-supervised settings on the ScanNet dataset [1] in Sec. S2 to further explore the effectiveness of the proposed method on indoor scene segmentation. Finally, in Sec. S3, we explore the importance of re-distributing biased pseudo labels by distribution aligning for unsupervised domain adaptation (UDA) on the GTA5 [7] → Cityscapes setting.

S1. More Results on Cityscapes

S1.1. Parameter Analysis

Labeling Ratio. Benefited from an improved teacher model, progressively enlarging labeling ratio α can help induce novel data while maintaining the quality of pseudo labels, and hence safely bootstrap the performance. Here, we present more experimental results and analysis on the Cityscapes split 1/8 at round k=2 to show the improvements from an enlarging labeling ratio.

α (%)	mIoU (%)
20	68.27 ± 0.12
30	68.54 ± 0.31
40	68.77 ± 0.10
50	**68.93 ± 0.16**
60	68.75 ± 0.04

Table S1. Parameter analysis for labeling ratio on the Cityscapes 1/8-split at round k=2 with random scaling between 0.25 and 1.0.

As shown in Table S1, if we directly apply iterative training without enlarging the labeling ratio, the performance gain is quite limited (68.01% → 68.27%). However, as we gradually enlarge the labeling ratio, a steady performance growth is observed with the largest improvement (68.01% → 68.93%) achieved at α=50%.

Moreover, we can observe the robustness of our progressive pseudo-labeling strategy from Table S1 that noticeable performance boost could be achieved in a relatively wide range (*i.e.* 40% ∼ 60%).

Data Augmentaion Magnitude. An orthogonal strategy is to progressively increase the magnitude of data augmentation. In our experiments, we focus on strengthening the random scaling factor on Cityscapes split 1/8 at round k=2. The range of random scaling intensity in round k=1 is [0.25, 1.0], which is regarded as the initial range. Afterward, we enlarge the initial range in the following self-training round, decreasing the lower bound 0.25 by $\beta_{\min}$ and increasing the upper bound 1.0 by $\beta_{\max}$.

$\beta_{\min}$	$\beta_{\max}$	mIoU (%)
0.4	0.0	68.77 ± 0.12
0.2	0.0	69.16 ± 0.03
0.0	0.0	68.93 ± 0.16
0.0	0.25	68.97 ± 0.44
0.0	**0.5**	69.52 ± 0.03
0.0	0.75	69.05 ± 0.06
0.2	0.5	**69.64 ± 0.01**

Table S2. Parameter analysis for random scaling magnitude on the Cityscapes split 1/8 at round k=2, with labeling ratio=50%.

As shown in Table S2, the best performance is achieved at $\beta_{min} = 0.2$ and $\beta_{max} = 0.5$. Notably, by enlarging the range of random scaling appropriately, a tangible performance gain is obtained (68.97%→ 69.64%).

Though enlarging data augmentation magnitude to different extent leads to various performance, we observe that we can harvest performance boost in a wide range of increased data augmentation magnitude as shown in Table S2, which proves the robustness of our progressive data augmentation strategy.

Data Augmentation	mIoU
None	70.24
Photometric Distortion	70.84
Random Rotation	70.26
Random Scaling	74.36

Table S3. Effectiveness of different data augmentation methods in semantic segmentation.

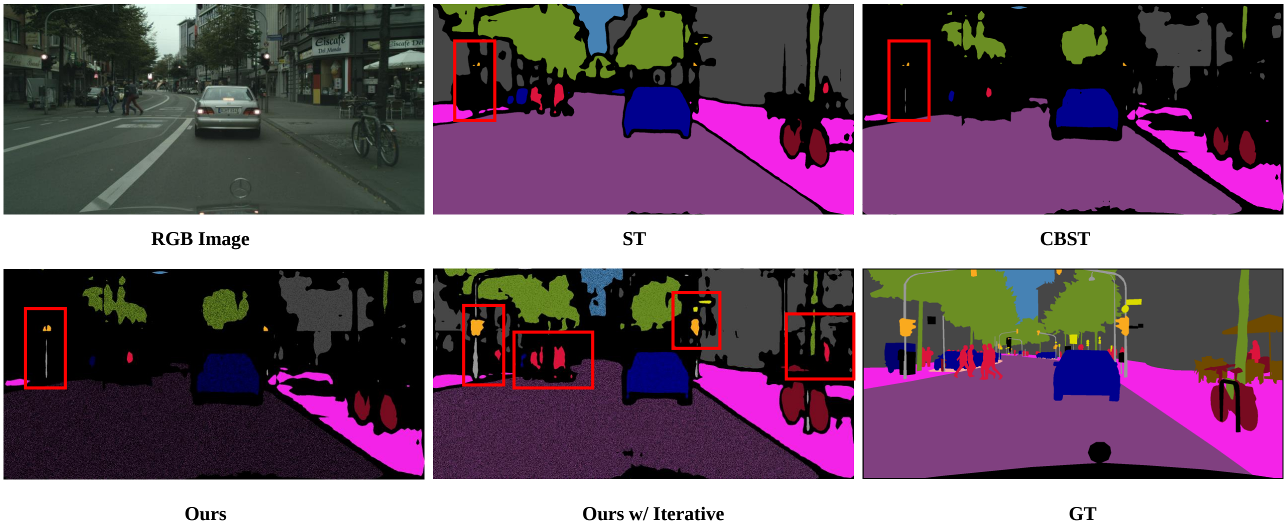

Figure S1. Visualization of pseudo labels generated by different methods. We provide the pseudo labels of ST, CBST and ours at round k=1 (*i.e.* Ours) as well as ours at round k=2 (*i.e.* Ours w/ Iterative), together with the RGB image and the corresponding ground-truth. Black areas indicate the ignored region.

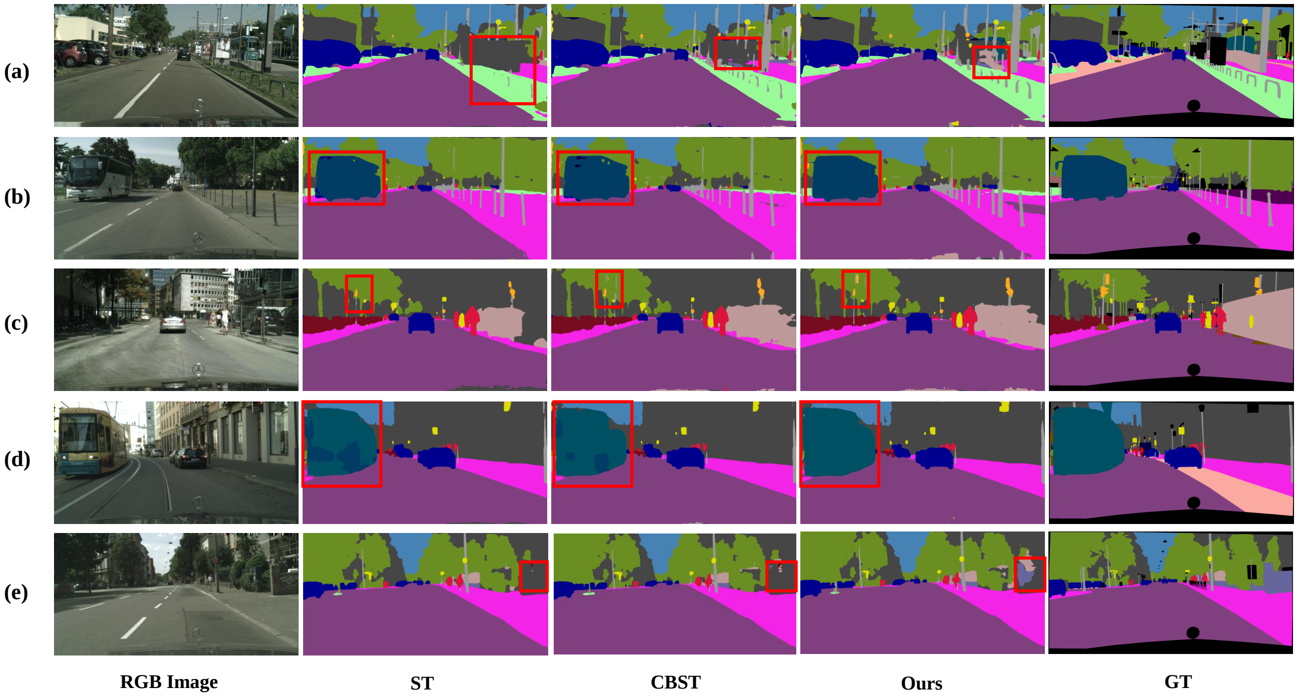

Figure S2. Qualitative results of different semi-supervised methods on Cityscapes at split 1/8 at round k=1. Along with the RGB image and its corresponding ground-truth, we provide the results of ST, CBST and our method respectively.

S1.2. Data augmentation in the progressive strategy

Here, we explain why only random scaling is considered in the progressive strategy. Our previous empirical experiments showed that random scaling is the most useful data augmentation method for semantic segmentation. To be specific, we have conducted experiments on the Cityscapes dataset with different data augmentation methods. Specifically, we trained a PSPNet50 using all 2975 fine-annotated training images with a crop size of 361×361 (half-resolution training). For data augmentation methods, we consider photometric distortion (brightness, contrast, saturation, and hue), random rotation, and random scaling following common setups in previous work [56]. We employed one of the three data augmentation methods or none of them, respectively, and report the performance on

Method	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	refriger	shower curt	toilet	sink	bathtub	others	mIoU	Gain
Baseline	75.1	81.3	47.0	63.5	56.7	**43.8**	59.3	48.6	44.4	55.6	28.3	34.5	45.4	44.8	47.3	53.4	76.6	52.5	62.8	36.1	52.8	0.0
ST	76.9	82.3	48.5	66.4	57.1	42.3	60.0	51.5	46.1	54.3	33.6	36.1	47.9	48.0	53.7	55.6	79.3	53.6	69.2	36.9	55.0	2.2
CBST	77.2	**83.5**	49.1	66.7	**57.7**	41.9	62.5	**52.5**	46.1	**57.1**	34.2	**37.0**	50.4	46.4	54.3	50.7	80.3	54.8	68.9	37.2	55.4	2.6
DARS (Ours)	**77.5**	83.4	**50.2**	**67.0**	57.5	**43.8**	**63.0**	52.3	**46.6**	56.7	**35.2**	35.3	**50.6**	**49.4**	**54.7**	**58.3**	**80.6**	**54.9**	**69.3**	**38.4**	**56.2**	**3.4**

Table S4. Comparisons of different semi-supervised approaches on the $1/8$ split of the ScanNet dataset at round k=1 with labeling ratio α=50%. The tail classes are highlighted in blue. We make the best performance result bold for each class. Single scale test is adopted for all methods here.

Method	Split	mIoU (%)			
		Baseline	Result	Oracle	Gain
DARS (Ours)	1/8	52.84	56.58	61.69	3.74
	1/4	56.61	58.35		1.74

Table S5. Our results on ScanNet with iterative training on split 1/8 and 1/4.

the validation set. As shown in Table S3, random scaling could bring significant performance boosts, whereas photometric distortion or random rotation could only bring limited gains. We will add this analysis to the supplementary material upon publication.

Moreover, applying too strong magnitudes for data augmentation methods like brightness and rotation might influence data distribution. Hence, we only consider random scaling in the progressive strategy. However, we believe other data augmentation methods like mixup could also be incorporated into our progressive strategy to further boost performance and we hope our idea of progressively increasing data augmentation magnitude for iterative training could benefit future research.

S1.3. Visualization of Pseudo Labels

To provide more information about our approach, we visualize the pseudo labels generated by our method as well as conducting comparisons with methods such as ST and CBST on the Cityscapes dataset.

As shown in Fig. S1, pseudo labels form ST and CBST are often overwhelmed by the majority classes like **road** and **vegetation**. And the tail class objects are often ignored in their pseudo labels, such as the **pole** and **traffic light** in the red box. As a result, the label distribution of their pseudo labels is extremely biased towards the dominant classes.

In contrast, with our distribution alignment and random sampling strategy to deal with the confidence overlapping phenomenon, the percentage of dominant classes are reduced and the pseudo labels are re-distributed to cover a large spatial area. Besides, our method successfully pseudo-labels the tail classes such as the **pole** and **traffic light** in the red box at round k=1 (see Ours in Figure S1).

Further, when we enlarge the labeling ratio to 50% at round k=2, the quality of our pseudo labeled data is further enhanced. More tail class objects are pseudo-labeled and incorporated into our pseudo labels, as shown by the red boxes of Ours (w/ Iterative) in Figure S1.

S1.4. Qualitative Results

In this section, we provide qualitative results of the semi-supervised semantic segmentation methods on the Cityscapes dataset. Concretely, we compare our results with ST and CBST methods at round k=1.

As shown in Fig. S2, previous methods mainly have two failure modes in segmenting tail classes: (1) they tend to leave out some tail classes like **fence**, **traffic light** and **wall** (*e.g.* in the red box areas, the **fence** is missing in (a), one **traffic light** is lost in (c), and the **wall** is completely unrecognized in (e)); (2) they suffer from the confusion with similar classes and mistake tail class object as other classes. For instance, in (b), part of the **bus** is mistaken as **vegetation**, **truck** or **car**, and in (d), some part of the **train** is misclassified as **bus**.

Thanks to our distribution alignment and sampling strategy to calibrate the bias, our method can alleviate the above two issues and thus outperforms ST and CBST on tail classes significantly. As shown in Fig. S2, our method can successfully segment the tail class objects as in (c) and recognize most tail class areas (*e.g.* the **fence** in (a) and the **wall** area in (e)). Moreover, our method significantly improves the model's ability to handle the confusion between similar classes and give consistent and correct predictions as in (b) and (d).

S2. Additional Experiments on ScanNet

To further demonstrate the transferability and broad applicability of our method, we evaluate it on the indoor scene dataset, ScanNet [1]. To be noted, we do not tune the hyperparameters on the ScanNet dataset to show the generality of our method.

Dataset. ScanNet is an RGB-D dataset collected from 1,513 indoor scenes. For the 2D semantic segmentation task, ScanNet contains 19,466 RGB images for training and 5,436 images for validation with a resolution of 1296×968. In our semi-supervised setting, 1/8 (*i.e.* 1/8-split) and 1/4 (*i.e.* 1/4-split) of the images are randomly chosen from the training set to serve as the labeled set. Pixel-level annotations for the following 21 object classes are provided: *wall, floor, cabinet, bed, chair, sofa, table, door, window, bookshelf, picture, counter, desk, curtain, refrigerator, shower curtain, toilet, sink, bathtub, other furniture, and void (the ignore category).*

Implementation Details We follow the same experimental setup as the Cityscapes dataset, except that the number of epochs is set to 20 for each training round and a crop size of 481×481 is adopted. Also, since the variance in our experiments is rather small as shown in this supplementary file, we only run one experiment for each setting on ScanNet to save the computational cost.

Main Results. We compare the proposed DARS method with the single thresholding method [16, 15, 5, 14, 8] (ST) and the class balance thresholding method [18, 2] (CBST) considering on the 1/8-split setting at round k=1 without iterative training. As shown in Table S4, the proposed simple DARS method achieves 56.2% mIoU on the validation set, surpassing ST and CBST method, which reiterates the superiority of our proposed method. To be noted, our method introduces little computational cost in comparison with the compared approaches.

Further, we report the final results of the proposed method with iterative training at split 1/8 and 1/4 in Table S5. Notably, our method achieves 58.35% in terms of mIoU with only 1/4 labeled data, which is very close to the fully-supervised results of 61.69%.

Analysis We notice that the performance gain achieved by self-training is relatively small on the ScanNet dataset in comparison with the Cityscapes dataset. We mainly attribute this to the difference between indoor and urban scenes. While urban scenes usually have similar structures (*e.g*. road is always at the bottom and the sky at the top), indoor scenes tend to have large variance and complex spatial relationships which impose obstacles for pseudo-labeling that relies on models trained with only a small set of labeled data. Exploring the 3D structure for semi-supervised learning in indoor scene parsing have the potential to address these difficulties which will be our future work. We believe our method could also be incorporated into other methods to further boost the performance for semi-supervised indoor scene parsing. Also, we barely finetune the hyperparameters like labeling ratio and data augmentation magnitude to save time and computational costs since our main purpose for experiments on ScanNet is to show the broad applicability of our method with superiority to previous self-training methods.

S3. Unsupervised Domain Adaptation Setting

In this section, we further conduct experiments on the more challenging unsupervised domain adaptation setting, in order to confirm our major insight about the importance of semantic-level distribution alignment in pseudo-labeling.

While we do not have the labeled set for the target domain to obtain the true label distribution, for comparison fairness with other methods, we could not perform DARS for generating unbiased pseudo labels. Instead, we use this setting to study the relationship between the extent of distribution mismatch in pseudo labels (*i.e*. KL divergence with target label distribution) and the performance boost.

We compare the following pseudo-labeling methods:

- ST: the single confidence thresholding method like [11, 16], regraded as the self-training baseline method;
- CBST: the class balanced confidence thresholding method [2, 18], which actually uses confidence to estimate the target label distribution;
- DARS (SD): DARS using the source label distribution as the target label distribution;
- DARS (TD): DARS using the target label distribution counted on the validation set of the target domain.

Dataset. We follow [4, 10, 12] to consider the popular synthetic-to-real adaptation task: GTA5 → Cityscapes. The GTA5 dataset [7] provides 24,966 images with pixel-wise labels. We use the 19 classes of GTA5 in common with the Cityscapes for adaptation. Moreover, we take advantage of image translation and use images translated by CyCADA [3] in GTA5 for training.

Implementation Details We also follow the same experimental setup as the Cityscapes dataset, except that the number of epochs is set to 10 for the pre-training round on GTA5 and a crop size of 713×713 is adopted.

Main Results. As shown in Table S7, the smaller the KL divergence between the distribution of pseudo labels and target labels is, the better performance is achieved, which highly validates our motivation to re-distribute biased pseudo labels. Moreover, it is noteworthy that when the pseudo labels achieve perfect distribution alignment with true distribution (*e.g*. DARS (TD)), it could achieve much more performance gain than other pseudo-labeling (*e.g*. 4% mIoU higher than DARS (SD), 2.27% higher than CBST in a single round).

Further, we report the results with iterative training of DARS (TD) in comparison with previous works in Table S6. We claim that the comparison is not fair since DARS (TD) utilizes the target label distribution from the validation set, and we show the encouraging and superior performance (*i.e*. 55.0% mIoU) of it only to highlight the importance of distribution aligning in pseudo-labeling for unsupervised domain adaptation settings, hoping to inspire more works in this direction.

Method	Backbone	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
CyCADA [3]	VGG-16	85.2	37.2	76.5	21.8	15.0	23.8	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
ASN [9]	ResNet-101	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
CLAN [6]	ResNet-101	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
ADVENT [10]	ResNet-101	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
CAG-UDA [13]	ResNet-101	90.4	**51.6**	83.8	34.2	27.8	38.4	25.3	**48.4**	85.4	38.2	78.1	58.6	**34.6**	84.7	21.9	42.7	**41.1**	29.3	37.2	50.2
RPT [14]	ResNet-101	89.7	44.8	**86.4**	**44.2**	**30.6**	41.4	**51.7**	33.0	**87.8**	**39.4**	**86.3**	65.6	24.5	**89.0**	36.2	46.8	17.6	**39.1**	**58.3**	53.2
CBST [18]	WideResNet-38	89.6	**58.9**	78.5	33.0	22.3	41.4	48.2	39.2	83.6	24.3	65.4	49.3	20.2	83.3	**39.0**	**48.6**	12.5	20.3	35.3	47.0
PyCDA [5]	WideResNet-38	**92.3**	49.2	84.4	33.4	30.2	33.3	37.1	35.2	86.5	36.9	77.3	63.3	30.5	86.6	34.5	40.7	7.9	17.6	35.5	48.0
CRST [17]	WideResNet-38	**91.7**	45.1	80.9	29.0	23.4	**43.8**	47.1	40.9	84.0	20.0	60.6	64.0	31.9	85.8	**39.5**	**48.7**	**25.0**	38.0	47.0	49.8
CyCADA* [3]	ResNet-50	85.6	37.6	81.7	34.3	20.2	35.8	41.4	31.7	85.2	37.8	74.0	**66.5**	24.5	83.5	24.6	19.0	0.0	30.1	26.7	44.2
DARS (TD)	ResNet-50	90.6	50.8	**88.0**	**43.4**	**33.9**	**48.3**	**53.4**	**50.2**	**87.0**	**46.2**	**80.9**	**71.6**	**34.4**	**87.3**	33.1	40.6	6.6	**45.6**	**53.4**	**55.0**

Table S6. Adaptation results from GTA5 → Cityscapes. The tail classes are highlighted in blue. We make the top-2 performance results bold for each class. CyCADA*: we re-implements CyCADA on our PSPNet-50 framework.

Method	D_{KL}	mIoU (%)	Gain (%)
Baseline	/	43.43 ± 0.01	0.0
ST	0.0727	48.04 ± 0.69	+4.61
CBST	0.0196	48.74 ± 0.16	+5.31
DARS (SD)	0.1558	47.01 ± 0.42	+3.58
DARS (TD)	0.0006	**51.01 ± 0.05**	**+7.58**

Table S7. GTA5 → Cityscapes results at round k=1. D_{KL} indicates the KL divergence between the distribution of pseudo labels and Cityscapes real labels.

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1, 3

[2] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum. *arXiv preprint arXiv:2004.08514*, 2020. 4

[3] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 4, 5

[4] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. 4

[5] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6758–6767, 2019. 4, 5

[6] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 5

[7] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 1, 4

[8] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 4

[9] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, pages 7472–7481, 2018. 5

[10] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 4, 5

[11] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 4

[12] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. In *AAAI*, pages 12613–12620, 2020. 4

[13] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 435–445, 2019. 5

[14] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Dong Liu, and Tao Mei. Transferring and regularizing prediction for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2020. 4, 5

[15] Yi Zhu, Zhongyue Zhang, Chongruo Wu, Zhi Zhang, Tong He, Hang Zhang, R Manmatha, Mu Li, and Alexander Smola. Improving semantic segmentation via self-training. *arXiv preprint arXiv:2004.14960*, 2020. 4

[16] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*, 2020. 4

[17] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019. 5

[18] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 4, 5