Pseudo Controlled Stable Diffusion for Semi-Supervised and Cross-Domain Semantic Segmentation (Appendix)

Dong Zhao[1], Qi Zang[2], Shuang Wang[2], Nicu Sebe[1], Zhun Zhong[3] ✉
[1] Department of Information Engineering and Computer Science, University of Trento, Italy
[2] School of Artificial Intelligence, Xidian University, Shaanxi, China
[3] School of Computer Science and Information Engineering, Hefei University of Technology, China

1. Appendix A Related work

Semi-Supervised Semantic Segmentation (SSSS). Semi-supervised semantic segmentation (SSSS) utilizes a small amount of labeled data along with abundant unlabeled data to improve segmentation performance while reducing annotation costs. The standard approach relies on pseudo-labeling-based self-training, where models generate pseudo-labels for unlabeled data and use them for training. However, pseudo-labels often accumulate noise, leading to suboptimal results. To address this, various improvements have been proposed. Strong perturbation methods like ST++ [36] refine pseudo-labels through extensive data augmentation, while DARS [6] and USRN [5] mitigate errors in minority classes using distribution alignment and subclass clustering. Consistency regularization ensures stable predictions by enforcing agreement across different model variations, as seen in CPS [2] and CCT [21]. AEL [12] and UniMatch [34] introduce adaptive CutMix and dual-stream learning to improve data diversity and generalization. Contrastive learning has also been applied in methods like DCC [14] and U2PL [29], enhancing feature representation and reducing confirmation bias. Recently, Transformer-based SSSS has gained attention due to the ability of Vision Transformers (ViTs) to model long-range dependencies. SemiCVT [13] enforces class-wise consistency between CNNs and Transformers, while other works [15, 17] integrate ViTs into CPS-based frameworks, though often as auxiliary components.

Cross-Domain Semantic Segmentation (CDSS) transfers the source knowledge to the target mainly by alignment of both domains and self-training on the target. The alignment based CDSS explore various domain alignment strategies, *e.g.*, adversarial training [8, 27], statistical matching [28, 30], across diverse alignment spaces (*e.g.*, input [7, 26], feature [27] and output space [25]) to reduce statistical differences between the two domains. Self-training-based CDSS methods primarily employ pseudo-labeling techniques to address the issue of inadequate target adaptation. Related works introduce pseudo-label selection strategy [1, 16, 19, 22, 39, 41, 42], strong augmentations [11] and high-resolution consistency [10] to alleviate the issue of error accumulation.

Stable Diffusion for Semantic Segmentation. Inspired by the success of SD [24]), prior works explore its potential in semantic segmentation tasks by designing new perception models [4, 33, 40] and synthetic new training data [3, 20, 31, 35, 35]. Benefiting from large-scale text-to-image pre-training, modifying SD as a backbone will enjoy strong cross-domain transfer capabilities [4], but the diffusion process and large denoising architecture will bring huge inference overhead [40]. Synthetic data provides feasible ideas for data scarcity semantic segmentation tasks. Some advanced works attempt to extract semantic masks during image generation and even don't fine-tune the SD. For instance, Nguyen et al. [20] and Wu et al. [32] refine the text-to-image cross-attention maps and treat them as semantic masks. Wu et al [31] add a perception head on SD and fine-tune the added units using a few target samples to generate paired data. However, this mode is prone to producing out-of-domain data with simple semantics, which makes it difficult for the model to learn useful knowledge from it. [31]. Yang et al [35] fine-tuned SD on massive labeled data to generate target-style data, which limits its application. Different from these methods, we focus on how to fine-tune SD with massive pseudo-labeled data and force it to generate diverse target-style training data in data scarcity segmentation scenarios.

2. Appendix B Implementation Details

Fine-Tuning Stable Diffusion (SD). $\mathcal{M}$ is estimated by thresholding the top 50% confidence-ranked pixel of each class across the entire dataset, meaning that pixel pseudo-

✉ Corresponding author.

labels with confidence rankings in the top 50% are assigned a value of 1 in $\mathcal{M}$, otherwise 0. In all experimental tasks, we fine-tune SD for 10k iterations, with the weights of the text encoder CLIP [23] frozen. The fine-tuning image size is fixed to 512×512, consistent with the pre-training. For datasets with image sizes larger than 512, we randomly crop 512×512 image patches. The batch size is set to 4. We generally fine-tune SD for only one round.

Data Synthesis. We set the diffusion iterations to 50 for SD. In Semi-Supervised Semantic Segmentation(SSSS), we use all structured pseudo-labels of unlabeled data, as well as semantic labels of a few labeled data as spatially controlled synthetic images. We generate 10k synthetic training data for Pascal-VOC, 5k for Cityscapes, 20k for ADE20K, and 10k for COCO. In Cross-Domain Semantic Segmentation(CDSS), we only use all structured pseudo-labels of unlabeled target domain data as a control to synthetic images. For all CDSS tasks, we generate 5k training data for Cityscapes. The spatial resolution of synthetic data for Pascal-VOC and COCO is 512×512, consistent with the labeled training data. For Cityscapes, due to its high resolution and variable scale, we first randomly re-scale the pseudo-labels within (0.5, 1), and then crop 512×512 patches with the proposed re-sampling strategy.

Details on Training Segmentation Models. For SSSS, we first train Unimatch [34] for 50 epochs, and then add our synthetic data to the labeled set to continue training. For CDSS, we first train the UDA methods HRDA [10] and DTST [38] for 20k and 10k iterations respectively, and treat our synthetic data as the labeled source data and re-train the UDA. During training, the synthetic data are resized to the same size as the real data.

3. Appendix C Detailed Analysis

More observations on generative hallucinations using pseudo-label. Fig. 1 shows more examples to demonstrate the generative hallucination issues that occur with Stable Diffusion under poorly structured pseudo-labels. Our proposed structured pseudo-labeling strategy effectively alleviates this issue, facilitating high-quality image generation.

Class IoU scores for cross-domain segmentation. Table 1 demonstrates that our method still exhibits significant advantages across various categories. Particularly, for some challenging adaptation categories such as Bike, Train, Truck, and Bus, our method significantly improves upon the baseline methods, further validating the effectiveness of our approach.

Image-level Class distribution of resampled synthetic images. Fig. 2 illustrates the class imbalance issue under semi-supervised learning. Certain minority classes such as trucks, buses, and trains are difficult for the model to learn sufficiently. Our method enriches their data distribution, greatly increasing the proportion of labels for these minority classes. This provides the model with the opportunity to fully adapt to them.

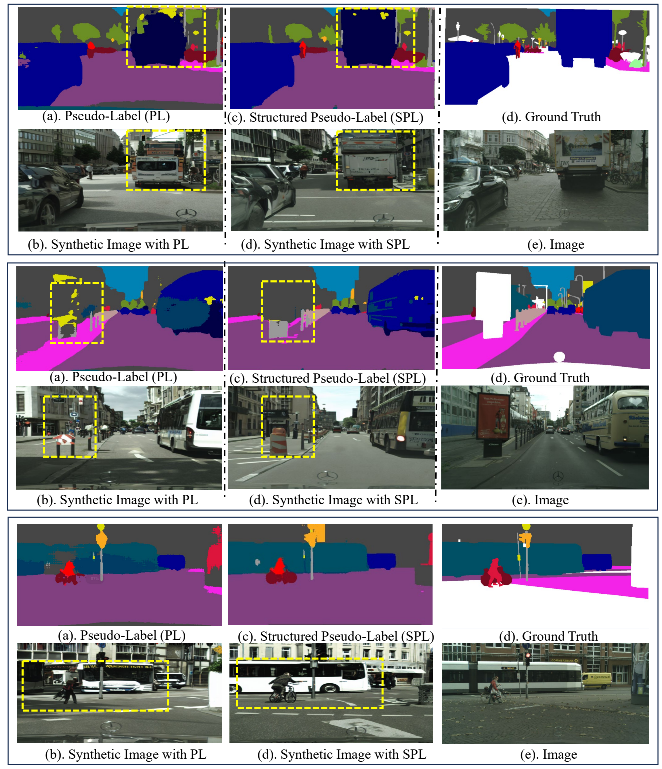


Figure 1. Visualization on generative hallucinations using pseudo-label. Using poorly structured pseudo-labels (a) as a control condition leads to the illusion in the yellow box of (b), but the structured ones (c) can alleviate this phenomenon as shown in (d).

Structured pseudo-labels (SPL) across object sizes. The class-wise results (grouped by large, medium, and small objects) show that SPL offers broad improvements: 1) Medium-structured classes (e.g., *Truck*, *Train*, *Bus*, *Wall*, *Rider*) benefit most, because they are highly sensitive to spatial structure. 2) Small objects (e.g., *Light* ↑1.0, *Sign* ↑2.1) also improve, confirming SPL's effect beyond large regions. 3) Hard classes like *Train* (28.4→36.2) and *Rider* (46.2→58.1) improve notably, showing gains are not limited to easily identifiable categories.

Compared to DatasetDM. To ensure fairness, we compare with DatasetDM under the same fine-tuning (FT) scope (denoising UNet only), using both labeled data and reliable pseudo-labels(PLs). In the table below, our method outperforms DatasetDM in both FID and mIoU. Notably, PLs-based FT yields minor improvements for DatasetDM, indicating that directly using PLs for SD tuning is non-trivial.

4. Appendix D Parameter analysis

Number of synthesized images. Fig. 3 indicates that, with an increase in the number of synthesized images, the perfor-

Method	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
Unsupervised domain adaptation: GTA → Cityscapes (Val.)- ViT-Mix																				
DAFormer [CVPR'22] [9]	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
HRDA [ECCV'22] [10]	96.4	74.4	91	61.6	51.5	57.1	63.9	69.3	91.3	48.4	94.2	79	52.9	93.9	84.1	85.7	75.9	63.9	67.5	73.8
Rtea [ICCV'23] [39]	97.1	75.2	92.6	63.5	51.8	58.2	66.5	71.2	91.1	49	96.8	81.5	54.2	94.2	84.8	86.6	75.7	62.2	66.7	74.7
MIC [CVPR'23] [11]	97.4	80.1	91.7	61.2	56.9	59.7	66.0	71.3	91.7	51.4	94.3	79.8	56.1	**94.6**	**85.4**	**90.3**	80.4	**64.5**	68.5	75.9
Pseudo-SD [Ours]	**97.6**	**81.8**	**92.2**	**62.3**	**59.9**	**60.5**	**66.0**	**73.9**	**92.2**	**53.2**	**95.1**	**80.4**	**58.9**	93.8	84.8	89.9	**83.3**	61.9	**73.9**	**77.0**
Unsupervised domain adaptation: Synthia → Cityscapes (Val.) - ViT-Mix																				
DAFormer [CVPR'22] [9]	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	–	89.8	73.2	48.2	87.2	–	53.2	–	53.9	61.7	60.9
HRDA [ECCV'22] [10]	85.2	47.7	88.8	49.5	4.8	57.2	65.7	60.9	85.3	–	92.9	79.4	52.8	89.0	–	64.7	–	63.9	64.9	65.8
Rtea [ICCV'23] [39]	87.8	49.0	90.3	50.3	5.5	58.6	66.0	61.4	86.8	–	93.1	79.5	53.1	89.5	–	65.1	–	63.7	64.6	66.5
MIC [CVPR'23] [11]	86.6	50.5	89.3	47.9	7.8	59.4	66.7	63.4	87.1	–	94.6	81.0	58.9	90.1	–	61.9	–	**67.1**	64.3	67.3
Pseudo-SD [Ours]	**87.6**	**51.8**	**90.7**	**48.4**	**9.3**	**61.3**	**66.7**	**63.9**	**88.7**	–	**95.9**	**82.0**	**60.5**	**91.6**	–	**63.1**	–	66.7	**68.7**	**68.6**
Source-free Unsupervised domain adaptation:: GTA → Cityscapes (Val.) - ResNet 101																				
DTST [CVPR'23] [38]	93.5	57.6	84.7	36.5	25.2	33.4	44.7	36.7	86.8	42.8	81.3	62.3	37.2	88.1	48.7	50.6	**35.5**	48.3	59.1	55.4
CROTS [IJCV'24] [18]	92.0	52.4	85.9	37.3	35.8	34.6	42.2	38.4	86.9	45.6	91.1	65.1	36.1	87.3	41.6	51.1	0.0	41.4	56.2	53.7
Cross-match [ICCV'23] [37]	94.5	**65.5**	87.4	45.7	42.6	42.3	46.7	**54.5**	**88.3**	48.0	84.7	66.0	33.4	89.9	53.5	56.8	0.0	46.9	49.4	57.7
Pseudo-SD [Ours]	**94.1**	58.6	**87.9**	**48.7**	**44.3**	**43.4**	**47.4**	54.1	87.9	**53.1**	**93.2**	**69.7**	**48.4**	**92.5**	**63.7**	**65.6**	**37.5**	**54.6**	**65.3**	**63.6**
Source-free Unsupervised domain adaptation:: Synthia → Cityscapes (Val.) - ResNet 101																				
DTST [CVPR'23] [38]	88.9	45.8	83.3	13.7	0.8	32.7	31.6	20.8	85.7	–	82.5	64.4	27.8	88.1	–	50.9	–	37.6	57.3	50.7
CROTS [IJCV'24] [18]	89.4	41.6	82.7	15.1	1.2	34.7	33.7	25.7	83.7	–	87.9	66.6	34.6	85.4	–	45.9	–	43.5	49.6	51.3
Cross-match [ICCV'23] [37]	**91.5**	**55.5**	85.4	34.4	8.3	40.8	40.0	44.4	86.6	–	84.3	62.4	22.0	88.3	–	**60.0**	–	40.6	45.6	55.6
Pseudo-SD [Ours]	88.6	47.5	**85.6**	**39.6**	**29.7**	**44.0**	**43.1**	**50.0**	**86.8**	**–**	**90.8**	**62.6**	**44.5**	**88.3**	**–**	57.9	**–**	**52.8**	**63.3**	**60.6**

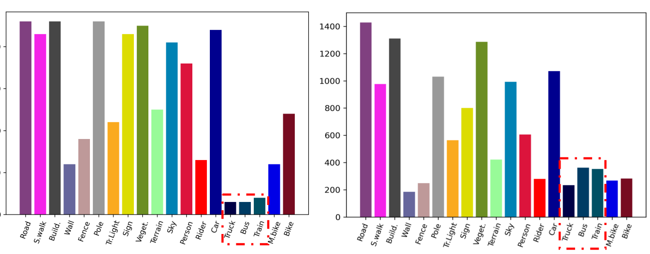
Table 1. Cross-Domain Semantic Segmentation performance (IoU in %).

Method	Road	SW	BD.	Veg.	Sky	Pers.	mIoU^{L6}	Car	Truck	Bus	Train	Wall	Fence	Terrain	Rider	mIoU^{M8}	M.bike	Bike	Pole	Light	Sign	mIoU^{S5}	mIoU
Unimatch	**97.4**	**79.3**	**90.5**	**91.3**	**94.1**	77.6	**88.3**	92.2	28.9	33.6	28.4	36.7	48.1	57.9	54.3	48.7	53.3	72.7	56.6	65.9	73.9	64.5	65.4
Ours w/o SPL	97.1	77.7	90.2	90.7	94.0	75.7	87.5	92.1	41.8	73.1	33.5	35.9	48.7	58.4	50.8	55.5	56.7	72.8	56.0	65.6	72.5	64.7	68.1
Ours w SPL	97.0	78.8	90.4	90.5	94.0	**77.7**	88.1	**93.5**	**50.1**	**80.5**	**36.2**	**38.2**	**48.3**	**59.8**	**57.7**	**58.1**	**60.5**	**74.7**	**57.3**	**66.6**	**74.2**	**66.6**	**69.8**

Table 2. The impact of using Structured Pseudo-Labels (SPL) on performance across different object sizes.

	SSSS: Pascal (1/115)		SSSS:Cityscapes (1/64)		CDSS: G→C	
	mIoU (↑)	FID (↓)	mIoU (↑)	FID (↓)	mIoU (↑)	FID (↓)
Base (Unimatch/HRDA)	75.2	-	65.4	-	75.9	-
DatasetDM w/o FT	74.6	44.2	60.6	87.4	71.8	87.4
+ Labeled FT	75.4	34.1	62.3	49.1	72.2	73.1
+ Labeled and unlabeled FT	75.7	30.6	63.7	35.7	75.0	47.7
Ours	**78.9**	**15.6**	**69.8**	**21.5**	**77.0**	**21.5**

Table 3. Comparison with DatasetDM under the same fine-tuning (FT) scope (denoising UNet only), using both labeled data and reliable pseudo-labels (PLs). In the table below, our method outperforms DatasetDM in both FID and mIoU.



(a) Image-level Class distribution of 64 labeled data

(b) Image-level Class distribution of 2k synthetic data

Figure 2. Comparison of image-level label distributions for synthetic and real images in semi-supervised cityscapes (64 labeled data) task.

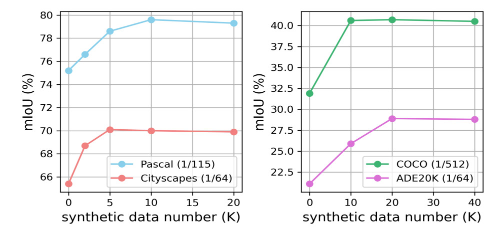


Figure 3. The impact of the number of synthesis images

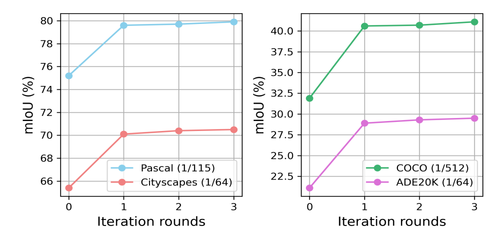


Figure 4. The impact of iteration rounds for finetuning.

mance of semi-supervised learning continuously improves. Once it reaches a certain threshold, further increases have a diminishing impact on performance stability.

Iteration rounds. Fig. 4 shows that our method can be iterated multiple rounds to enhance the quality of synthesis and benefit the segmentation model. Although multiple rounds yield marginal improvement, it significantly increases training time. We ultimately opted for a single-round iteration.

5. Appendix E More visualizations

Visualization of synthesize images only using pseudo-text-prompts As shown in Fig. 5, the layout of synthesized images is confusing because using category names as text prompts does not include spatial relationships. This makes it difficult to synthesize high-quality images on urban street scenes with complex layouts.

Synthesize images using source-domain semantic mask as conditions. Fig. 6 demonstrates that using semantic masks from the source domain as spatial conditions can still

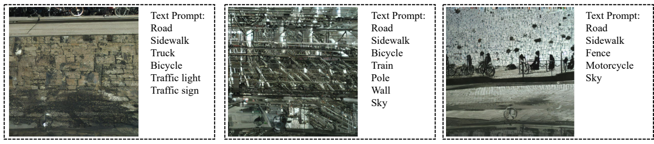


Figure 5. Failure cases of synthetic images using only pseudo-text-prompts. Images synthesized this way exhibit chaotic layouts and noisy textures.

generate high-quality images. However, due to differences in perspective, object resolution, and spatial layout between the source and target domains, using these synthesized data only brings about marginal improvements in target performance. For example, data from the source domain GTA5, generated using a synthetic engine, presents a wide field of view, whereas the target domain Cityscapes, captured by in-car cameras, exhibits a narrower field of view.

Figure 6. Synthesize images using source-domain semantic masks in GTA $\rightarrow$ Cityscapes unsupervised domain adaptation task. Despite its visual resemblance to the target domain, there is a significant domain shift in spatial layout.

More visualization about filtering out distorted connected regions Fig. 7 shows our method's ability to remove heavily distorted regions in synthesized images while preserving some challenging recognition capabilities, thereby aiding in enhancing the model's performance.

More visualization of synthetic images Fig. 8 demonstrates that our method can synthesize a wide variety of images for minority classes, as reflected in diverse layouts and scene variations. This diversity is advantageous for improving the model's generalization boundaries.

More indoor scene visualizations. Fig. 9 shows that our method still performs well in indoor scenes.

6. Appendix F Limitations

Extra overhead. Since our method requires fine-tuning Stable Diffusion and then generating new samples for training, we admit that our method introduces extra overhead but it can significantly improve the performance, *e.g.*, +5% of mIoU on ADE20K (1/64), Cityscape(1/64), and COCO(1/512). Commonly, our method requires twice the

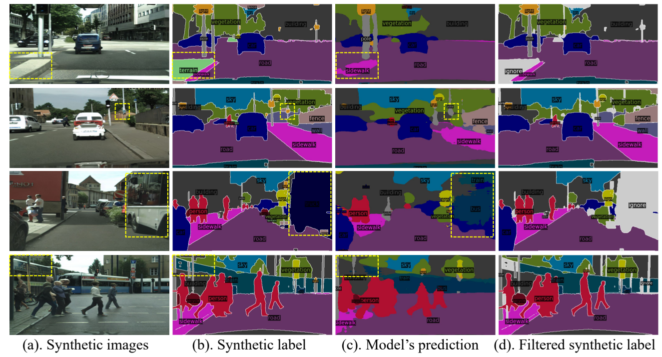


Figure 7. Display of more synthetic images and its corresponding semantic mask with 1024×512 resolution. The data is sampled from semi-supervised cityscapes (64 labeled data) task.

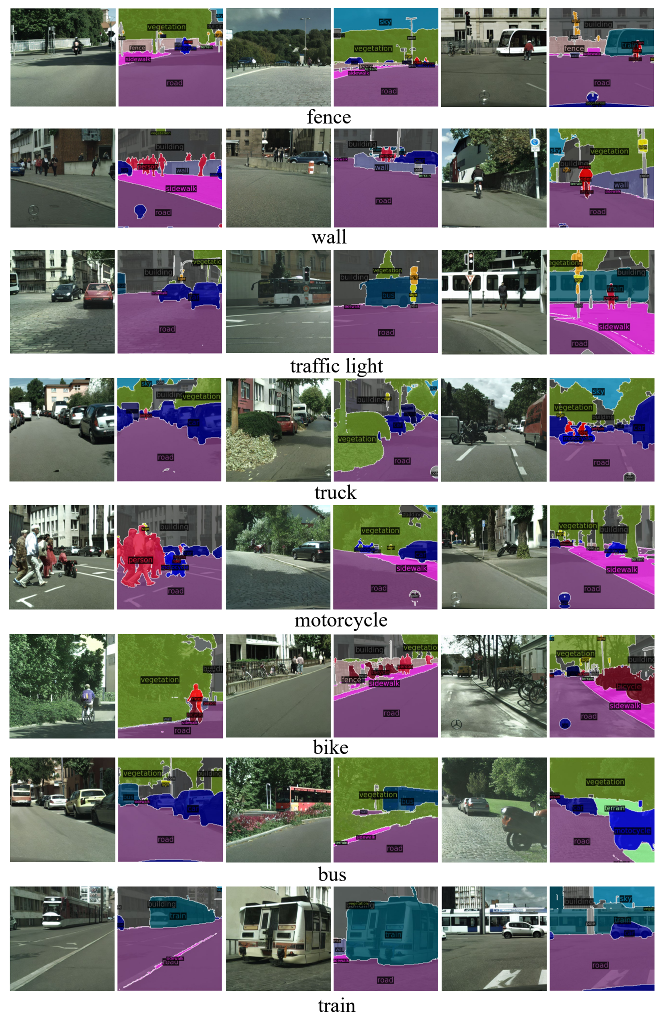
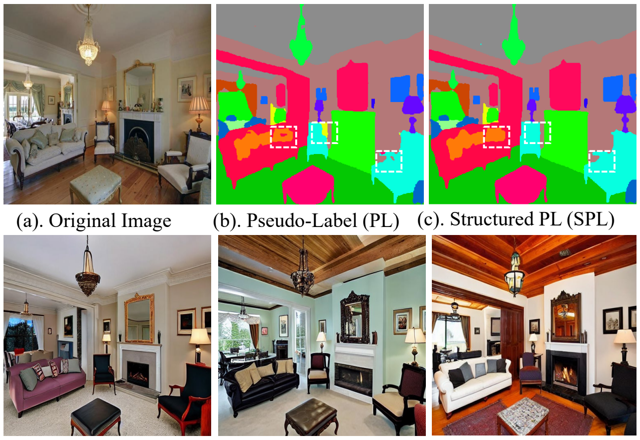


Figure 8. Display of 512×512 resolution synthetic images used for training, containing more minority categories. The data is sampled from semi-supervised cityscapes (64 labeled data) task.

training time compared to the SSSS baselines but does not affect the inference speed. Besides, we would like to clarify



(a). Original Image (b). Pseudo-Label (PL) (c). Structured PL (SPL)

(d). Synthetic Image 1 (e). Synthetic Image 2 (f). Synthetic Image 3

Figure 9. Indoor scenes on ADE20K.

(a). Structured pseudo label (b). Distorted synthetic images

Figure 10. Failure cases. Although the given structured pseudo-labels are semantically correct, their corresponding synthetic images exhibit semantic distortions.

that the main goal of SSSS is to enhance the performance on target data and few methods focus on the training cost in the community. Considering the effectiveness of our method, we thus think our extra training cost is acceptable for the task of SSSS.

Semantically distorted synthetic images. Although our results show that Pseudo-SD generates high-quality samples for most images, some low-quality generated images still exist. As shown in Fig. 10, despite well-structured pseudo-label masks, certain generated images remain partially incomprehensible or perceptually unclear, posing potential risks to SSSS and CDSS tasks. Further investigation and mitigation of these issues will be the focus of our future research.

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proc. CVPR*, pages 15384–15394, 2021. 1

[2] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proc. CVPR*, pages 2613–2622, 2021. 1

[3] Guillaume Couairon, Marlene Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proc. ICCV*, pages 2174–2183, 2023. 1

[4] Rui Gong, Martin Danelljan, Han Sun, Julio Delgado Mangas, and Luc Van Gool. Prompting diffusion representations for cross-domain semantic segmentation. *arXiv preprint arXiv:2307.02138*, 2023. 1

[5] Dayan Guan, Jiaxing Huang, Aoran Xiao, and Shijian Lu. Unbiased subclass regularization for semi-supervised semantic segmentation. In *Proc. CVPR*, pages 9968–9978, 2022. 1

[6] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *Proc. ICCV*, pages 6930–6940, 2021. 1

[7] 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 Jennifer Dy and Andreas Krause, editors, *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 1

[8] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proc. CVPR*, June 2018. 1

[9] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proc. CVPR*, pages 9924–9935, 2022. 3

[10] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Proc. ECCV*, pages 372–391. Springer, 2022. 1, 2, 3

[11] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proc. CVPR*, pages 11721–11732, 2023. 1, 3

[12] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Proc. NeurIPS*, 34:22106–22118, 2021. 1

[13] Huimin Huang, Shiao Xie, Lanfen Lin, Ruofeng Tong, Yen-Wei Chen, Yuexiang Li, Hong Wang, Yawen Huang, and Yefeng Zheng. Semicvt: Semi-supervised convolutional vision transformer for semantic segmentation. In *Proc. CVPR*, pages 11340–11349, 2023. 1

[14] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proc. CVPR*, pages 1205–1214, 2021. 1

[15] Peixia Li, Pulak Purkait, Thalaiyasingam Ajanthan, Majid Abdolshah, Ravi Garg, Hisham Husain, Chenchen Xu, Stephen Gould, Wanli Ouyang, and Anton van den Hengel. Semi-supervised semantic segmentation under label noise via diverse learning groups. In *Proc. ICCV*, pages 1229–1238, 2023. 1

[16] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *Proc. CVPR*, pages 11593–11603, 2022. 1

[17] Yijiang Li, Xinjiang Wang, Lihe Yang, Litong Feng, Wayne Zhang, and Ying Gao. Diverse cotraining makes strong semi-supervised segmentor. In *Proc. ICCV*, pages 7273–7282, 2023. 1

[18] Xin Luo, Wei Chen, Zhengfa Liang, Longqi Yang, Siwei Wang, and Chen Li. Crots: Cross-domain teacher–student learning for source-free domain adaptive semantic segmentation. *IJCV*, 132(1):20–39, 2024. 3

[19] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *Proc. ECCV*, 2020. 1

[20] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Proc. NeurIPS*, 36, 2024. 1

[21] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proc. CVPR*, pages 12674–12684, 2020. 1

[22] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proc. CVPR*, June 2020. 1

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pages 8748–8763. PMLR, 2021. 2

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pages 10684–10695, 2022. 1

[25] 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 *Proc. CVPR*, June 2018. 1

[26] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proc. CVPR*, pages 2517–2526, 2019. 1

[27] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 642–659, 2020. 1

[28] Shuang Wang, Dong Zhao, Chi Zhang, Yuwei Guo, Qi Zang, Yu Gu, Yi Li, and Licheng Jiao. Cluster alignment with target knowledge mining for unsupervised domain adaptation semantic segmentation. *TIP*, 31:7403–7418, 2022. 1

[29] 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 *Proc. CVPR*, pages 4248–4257, 2022. 1

[30] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proc. CVPR*, June 2020. 1

[31] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Proc. NeurIPS*, 36, 2024. 1

[32] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *Proc. ICCV*, 2023. 1

[33] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proc. CVPR*, pages 2955–2966, 2023. 1

[34] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proc. CVPR*, 2023. 1, 2

[35] Lihe Yang, Xiaogang Xu, Bingyi Kang, Yinghuan Shi, and Hengshuang Zhao. Freemask: Synthetic images with dense annotations make stronger segmentation models. *arXiv preprint arXiv:2310.15160*, 2023. 1

[36] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proc. CVPR*, 2022. 1

[37] Yifang Yin, Wenmiao Hu, Zhenguang Liu, Guanfeng Wang, Shili Xiang, and Roger Zimmermann. Crossmatch: Source-free domain adaptive semantic segmentation via cross-modal consistency training. In *Proc. ICCV*, pages 21786–21796, 2023. 3

[38] Dong Zhao, Shuang Wang, Qi Zang, Dou Quan, Xiutiao Ye, and Licheng Jiao. Towards better stability and adaptability: Improve online self-training for model adaptation in semantic segmentation. In *Proc. CVPR*, pages 11733–11743, 2023. 2, 3

[39] Dong Zhao, Shuang Wang, Qi Zang, Dou Quan, Xiutiao Ye, Rui Yang, and Licheng Jiao. Learning pseudo-relations for cross-domain semantic segmentation. In *Proc. ICCV*, pages 19191–19203, October 2023. 1, 3

[40] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023. 1

[41] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proc. ECCV*, pages 289–305, 2018. 1

[42] Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proc. ICCV*, October 2019. 1