Source-Free Domain Adaptation for RGB-D Semantic Segmentation with Vision Transformers: *Supplementary Material*

Giulia Rizzoli Donald Shenaj Pietro Zanuttigh
University of Padova, Italy

This document contains supporting material for the paper *Source-Free Domain Adaptation for RGB-D Semantic Segmentation with Vision Transformers*. Here, we include additional ablation experiments supporting some design choices of the proposed method along with additional qualitative visual results on the SELMA dataset [1].

Dataset	β				
	0.00	0.01	0.05	0.09	0.12
SYNTHIA	39.79	39.60	40.66	**40.82**	38.58
SELMA	38.25	39.52	39.17	**40.80**	38.60

Table 1. Performance with different values of β.

1. Tuning on input level adaptation

In this section, we present some additional results concerning the input level adaptation in the Fourier domain.

For this strategy, a key parameter is β as it selects which region of the amplitude spectrum is going to be replaced with the target style. As pointed out in the paper, larger values of β lead to a stronger adaptation effect but also introduce visual distortion. While this aspect has been discussed for standard images in previous works [3], its impact on depth data has never been analyzed.

The main paper already shows an example relative to the SYNTHIA to Cityscapes adaptation, while Figure 1 in this document shows an additional example relative to the SELMA to Cityscapes setting. Notice that, as pointed out in the main paper, the approach allows to better align the depth ranges and matches the fact that real-world target data computed with stereo vision has more artifacts and a less sharp distribution. The example in the image confirms these observations and shows that they are not dataset-dependent.

Using larger β introduces artifacts that are visually disturbing, yet the key question is if they affect also the network performances. Table 1 shows the mIoU after the pretraining step for different values of β. Notice how the relatively large value of $\beta = 0.09$ leads to depth maps not very visually appealing but instead effective when used to aid the segmentation model. The results in Table 1 refer to employing the average style computed over $2.5k$ patches of the target dataset. Based on the empirical findings, it was determined that employing the mean value of the objective variable across each training step batch leads to a further performance improvement (e.g., to 41.01 for the SYNTHIA-to-Cityscapes setting). The utilization of the per-batch style may explain why it effectively captures the significant variations in depth values based on the patch's position. This strategy was then adopted for all the subsequent tests.

2. Depth-dependent Entropy Loss Weighting

Another design choice that needs to be properly evaluated is the depth-dependent weight for the entropy loss. Various models can be selected for this task, following the rationale that closer regions should get a higher weight since they have a better image resolution and depth accuracy. We considered 4 possible options, where d is the normalized disparity in the $[0, 1]$ range:

1. Directly using disparity $w = d$ as the weight;
2. Using a piecewise linear function to crop the weights range, set to $w = \frac{1}{\sqrt{10}}$ for $d < 0.1$, to $w = \sqrt{10}$ for $d \geq 1$ and with a linear disparity-dependent increase in the middle;
3. Using the square root of disparity, i.e., $w = \sqrt{d}$;
4. Using an exponential function of the disparity value, i.e., $w = e^d$.

Table 2 shows how directly using the disparity as the weight led to the best performances.

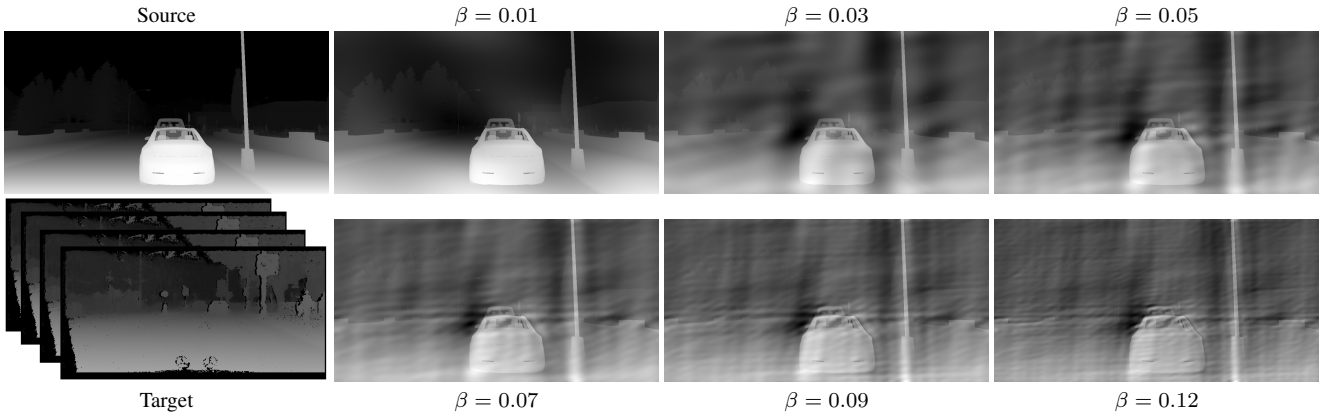

Figure 1. Effect of the Fourier Domain style transfer applied on depth images from the SELMA dataset, whereas $\beta = 0$ is equivalent to no transfer and $\beta = 1$ to the transfer of the full target amplitude.

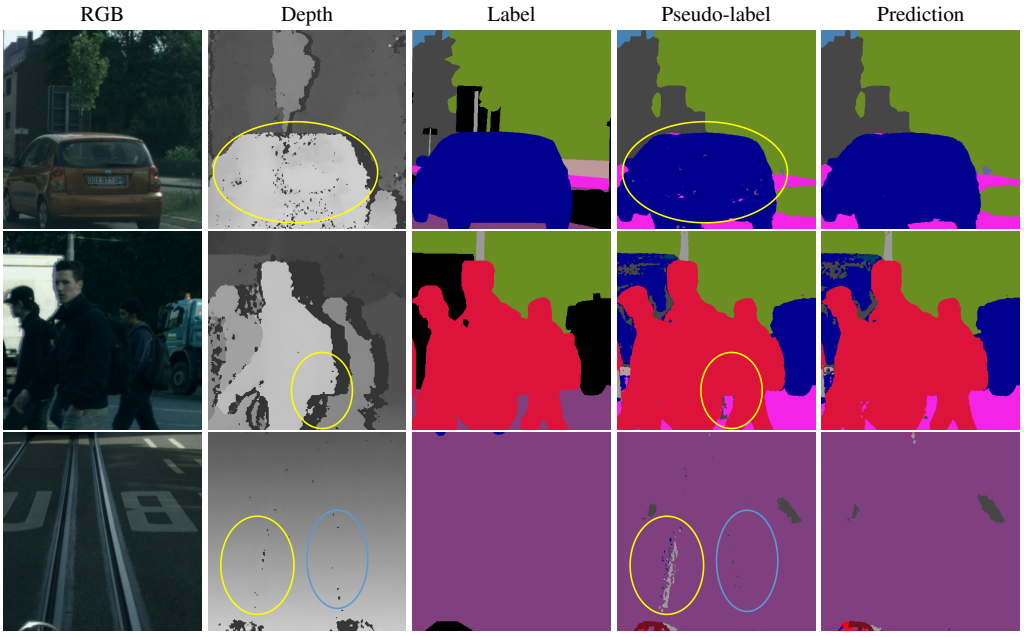

Figure 2. Self-training with depth masking allows masking areas where the depth fails to capture the scene content.

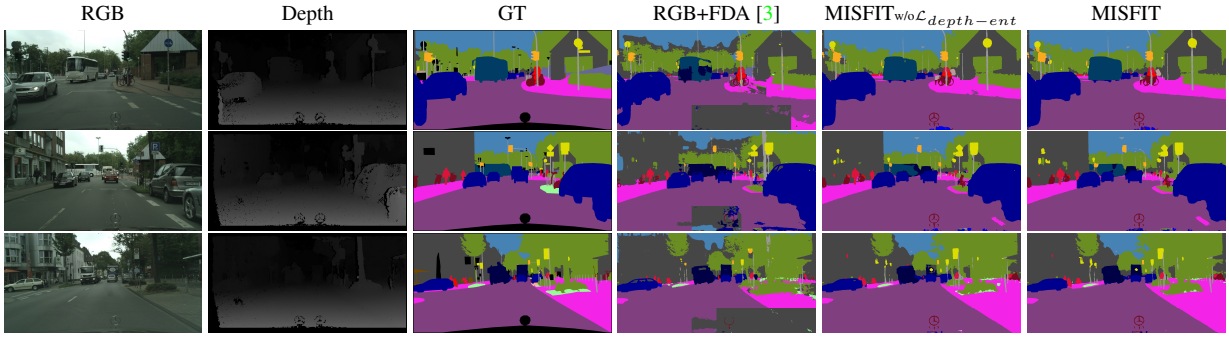

Figure 3. Qualitative semantic segmentation results for the SELMA-to-Cityscapes adaptation task.

Weighting Scheme	mIoU
(1)	**54.52**
(2)	53.05
(3)	52.93
(4)	52.80

Table 2. Ablation on different weighting schemes.

3. Ablation on Self-training with depth

As mentioned in the main paper, pseudo-label filtering has the capability to help the adaptation process considerably. In Figure 2, we show that masking pixels with missing or distorted depth data improves the prediction in critical regions. In the first row, the depth of the car is noisy and has discontinuities determining misleading predictions in the pseudo-labels, as evidenced by the circles. Therefore, it is important to discard the wrong supervision of the labels corresponding to those regions, by masking the pseudo-labels. In this way, the model learns to handle the gaps coherently. Analogously, in the second row, holes in the depth are introduced by the movement of the people, and effectively masking them, makes the employed pseudo-labels more consistent. Finally, in the last row, noisy artifacts in the depth are introduced by railway tracks on the road, compromising significantly the pseudo-labels.

Furthermore, in Table 3, a comparative evaluation is conducted between standard self-training and depth-aware self-training. The depth usage allows for a noticeable improvement over the RGB and RGB-D self-training. In particular with the standard self-training, the approach starts the learning properly, still, it becomes too confident about the wrong self-predictions in regions like the ones of Figure 2. Using instead the depth guidance the corrupted pseudo-labels are discarded and the learning continues to improve up to convergence at around 52.5%.

Input	ST	ST_d	mIoU
RGB			36.93
RGB	✓		40.59
RGB+D			39.79
RGB+D	✓		41.29
RGB+D		✓	**52.50**

Table 3. Improvement of the self-training using depth data. ST corresponds to the standard self-training, while ST_d is the depth-guided self-training.

4. Ablation on different backbones

With the aim of testing the generalization ability of the transformer architecture, we tested MISFIT over different backbones (see Table 4). Even with the utilization of less complex backbones (MiT-B2 and MiT-B4), the results obtained using the model of [2] are comparable to the state-of-the-art performance. This observation suggests that the architecture exhibits a level of robustness and efficiency, enabling competitive outcomes with lower parameter counts.

Encoder	(M)Params.	mIoU
MiT-B2	27.7	43.2
MiT-B4	64.1	52.55
MiT-B5	84.7	**54.5**

Table 4. Ablation on different backbones.

5. Additional Visual Results

Since in the main paper visual results are shown only for the SYNTHIA-to-Cityscapes setting, in Figure 3, we provide some qualitative results on the SELMA-to-Cityscapes setting. In the first row, the feature-level adaptation technique enhances the accuracy of car shapes and eliminates segmentation artifacts on the road. Furthermore, incorporating the entropy output level loss refines the shape of the sidewalk on the right side. In the second row, in addition to similar improvements observed on the road and the car on the right, it is evident that using only input-level adaptation results in a complete failure to detect the bus in the background. In contrast, our approach properly detects it. In the third row, the approach is able to refine the shape of the vegetation region on the right side and removes artifacts on the road.

References

[1] Paolo Testolina, Francesco Barbato, Umberto Michieli, Marco Giordani, Pietro Zanuttigh, and Michele Zorzi. Selma: Semantic large-scale multimodal acquisitions in variable weather, daytime and viewpoints. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 1

[2] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[3] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 1, 2