Source-Free Domain Adaptation for RGB-D Semantic Segmentation with Vision Transformers

Giulia Rizzoli Donald Shenaj Pietro Zanuttigh
University of Padova, Italy

Abstract

With the increasing availability of depth sensors, multimodal frameworks that combine color information with depth data are gaining interest. However, ground truth data for semantic segmentation is burdensome to provide, thus making domain adaptation a significant research area. Yet most domain adaptation methods are not able to effectively handle multimodal data. Specifically, we address the challenging source-free domain adaptation setting where the adaptation is performed without reusing source data. We propose ***MISFIT****: **MultImodal Source-Free Information fusion Transformer**, a depth-aware framework which injects depth data into a segmentation module based on vision transformers at multiple stages, namely at the input, feature and output levels. Color and depth style transfer helps early-stage domain alignment while re-wiring self-attention between modalities creates mixed features, allowing the extraction of better semantic content. Furthermore, a depth-based entropy minimization strategy is also proposed to adaptively weight regions at different distances. Our framework, which is also the first approach using RGB-D vision transformers for source-free semantic segmentation, shows noticeable performance improvements with respect to standard strategies.*

1. Introduction

Semantic segmentation has traditionally been performed employing RGB images, which solely capture color information. Yet, as depth sensors become more widely available, multimodal frameworks that integrate RGB visuals with depth information have emerged [21]. This integration offers the potential for improved semantic segmentation performance due to the additional clues provided by depth. Depth information proves particularly beneficial in several scenarios, including distinguishing between objects with similar colors but different distances, as well as aiding the segmentation of objects with complex geometries. Although state-of-the-art approaches achieve good results on several benchmarks, most multimodal methods do not test the generalization capabilities of the model [2, 14, 31]. Domain adaptation is especially crucial in real-world scenarios where environments and conditions may vary, making it challenging to train a single model that performs well across all domains [27]. The majority of domain adaptation methods work with color data alone [10, 13, 38] or focus on the Unsupervised Domain Adaptation (UDA) setting. However, a more challenging yet realistic setting is Source Free Domain Adaptation, where the pre-trained model undergoes adaptation without accessing the source data, as opposed to the typical joint source supervised and target unsupervised training approach in standard UDA. In this scenario, jointly using color and depth data for pre-training is unexplored. Given the recent studies demonstrating the generalization [7, 18] and multimodal processing [2, 14] capabilities of transformer architectures, in this paper we explore the feasibility of employing a transformer architecture for source-free domain adaptation. Furthermore, we would like to exploit the potential of depth data in guiding the adaptation process. To this extent, we propose our method, MultI-

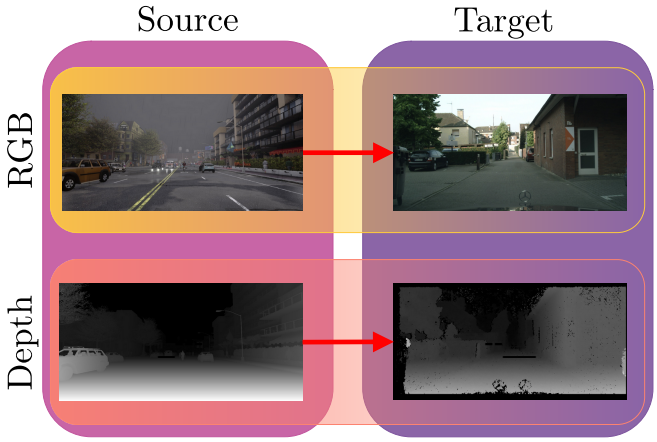

Figure 1. RGB-D Domain Adaptation: (a) Color domain shift, where variations in lighting conditions, color distribution, and texture affect the RGB images; (b) Depth domain shift, where differences in depth estimation strategies, surface geometry, and objects' scale impact the depth maps.

modal Source-Free Information fusion Transformer (MISFIT), which includes the following contributions:

1. The introduction of the first RGB-D framework for source-free domain adaptation semantic segmentation, exploiting vision transformers;

2. The exploration of input-level depth stylization in source pre-training using a fast and simple approach operating in the frequency domain;

3. The evaluation of generalization capabilities of RGB-D attention fusion within transformer architectures;

4. The development of a novel approach that leverages depth data in a self-teaching optimization scheme for source-free domain adaptation.

The proposed approach tackles the multimodal source-free domain adaptation task by introducing several provisions into a vision transformer architecture for semantic segmentation. The method involves distinct stages for both pre-training and adaptation. In the preliminary pre-training phase, we employ a domain stylization to tailor the input data, thereby enhancing the adaptability of the model. Within the internal network representation, we introduce modifications to the attention module of the transformer to handle the multimodal nature of the data effectively in both stages. Finally, during adaptation, our approach integrates a depth-guided self-teaching strategy to refine the segmentation results. We validated it on standard RGB-D benchmarks and the employed provisions allow to tackle the source-free domain adaptation task effectively.

After discussing the related works in Section 2, we will introduce the main components of our method in Section 3, detailing the input (Section 3.1), feature (Section 3.3) and output (Section 3.3) level provisions. Finally, we present the experimental results and ablation studies in Section 4 and draw the conclusions (Section 5).

2. Related Works

Multimodal Semantic Segmentation Recent studies have highlighted the potential of additional representations, such as depth and thermal data, in extracting semantic cues [23, 31]. Early multimodal segmentation techniques involved combining RGB data with other modalities into multi-channel representations, which were then fed into standard semantic segmentation networks [21]. This simple fusion strategy fails to comprehensively capture the varied information conveyed by each modality. To address this limitation, current methods employ various fusion strategies at different levels in the deep network. These approaches typically rely on a multi-stream encoder with a network branch for each modality, along with additional network modules that combine modality-specific features into fused ones and carry information across branches [2, 14, 41].

Transformer-based Adaptation for Semantic Segmentation Several studies investigated the potential of transformers for semantic segmentation in unsupervised domain adaptation (UDA) settings [7, 18]. They showed the generalization potential of Transformers architectures [12, 34] compared to the widely used convolutional neural networks. Hoyer et al. [8] propose a multi-resolution training approach for UDA to preserve fine segmentation details and capture long-range context dependencies. Park et al. [18] apply an entropy-based re-weighting in the attention module to address domain discrepancy. Although these methods show the applicability of vision transformers in the UDA scenario, none of them investigates adaptation in the source-free setting data nor domain adaptation with multimodal data.

Multimodal Domain Adaptation for Semantic Segmentation Hu et al. [9] addresses a single-stage input-level fusion, summing the depth after being injected into one attention block. xMUDA [11] proposes an unsupervised domain adaptation scheme for 3D semantic segmentation where the output feature of two distinct networks (2D for RGB and 3D for LiDAR) are fused through mutual mimicking. In MM-TTA [25], they investigate the challenge of test-time adaptation for multi-modal 3D semantic segmentation. Due to the additional domain shift introduced by pre-training a model on depth, there is no currently existing framework for source-free domain adaptation exploiting depth on the task of 2D semantic segmentation.

Unsupervised vs Source-Free Domain Adaptation In the challenging domain of Source-free domain adaptation (SFDA), the availability of source domain data is limited to an initial pre-training stage, while the subsequent adaptation process relies solely on unlabeled target data. Domain adaptation methods can be classified into two categories: data-level approaches and model-level approaches [40]. Data-level approaches aim to mitigate the domain shift by manipulating the target data to resemble the source domain. This involves aligning various aspects such as the imaging style in the input space [6, 28, 36], feature space [16, 17], output space [29]. However, SFDA presents a greater challenge as domain alignment must be achieved without access to the source dataset, making traditional adversarial learning methods used in UDA unsuitable. On the other hand, *model-level* approaches for domain adaptation include self-training, where the model is used to generate pseudo-labels for the data from the unlabeled target domain [40, 43, 44]. In standard UDA, entropy minimization approaches enhance the quality of pseudo-labels by minimizing the entropy of the target data or using the entropy map as input for a domain discriminator with an adversarial learning strategy [3, 33]. However, in the absence of labeled source domain

data in the source data-free setting, the effectiveness of entropy minimization-based methods can be compromised.

Source-Free Domain Adaptation Apart from standard domain adaptation, more recent works, investigate unimodal source-free adaptation. Liu et al. [15] leverage self-supervised learning to learn representations that are robust to domain shift and a knowledge distillation loss function is used to align the representations of the source and target domains. Fleuret et al. [5] exploit posterior probabilities to estimate uncertainty in the adaptation process. Huang et al. [10] uses contrastive category discrimination on pseudo-labels target samples to learn category-discriminative representations. You et al. [38] adopted a positive-negative learning strategy in combination with intra-class pseudo-labels thresholding. Kundu et al. [13] employ several encoder-level heads which are further pruned to select the optimal one. Ye et al. [37] employ uncertainty and prior distribution-aware domain adaptation techniques, incorporating both adversarial learning and self-training strategies, to create a set of virtual source domain data. Yang et al. [35] propose a weight-regularized distribution transfer method, followed by class-balanced thresholding and multi-class negative techniques during the adaptation phase. Zhao et al. [42] introduce a dynamic teacher update mechanism and a resampling strategy based on training consistency. Furthermore, to tackle diverse practical contexts, some approaches integrate source-free domain adaption with federated learning [24], black box test [19], or robust transfer [1].

3. Method

In this section, we introduce the three main strategies that we adopt for source-free domain adaptation, organizing them according to the stage at which they are employed: at the input level we exploited style transfer during pre-training (Section 3.1); at the feature level we tackle the multimodal setting by exchanging information in the attention module of the transformer (Section 3.2); at the output level a depth-based self-teaching strategy is used for domain adaptation (Section 3.3). An overview of the framework is shown in Figure 2.

We denote the labeled source domain data samples as $\mathcal{D}_s = ((x^s_{rgb}, x^s_d), y^s)$, where y^s is the label corresponding to the multimodal input (x^s_{rgb}, x^s_d) (as expected x^s_{rgb} is the color image and x^s_d the corresponding depth map). The target domain is unlabeled and drawn from the distribution $\mathcal{D}_t = (x^t_{rgb}, x^t_d)$. In the source-free setting, we assume that $\mathcal{D}_s$ is only available during model pre-training. The target is to assign to each pixel one of the C possible classes. The transformer architecture is constituted by a multi-head attention, which constitutes the encoder part, and a segmentation decoder, in particular, we employed SegFormer [34] as the starting architecture. The probability output of the network is denoted as $p(x) \coloneqq p \in \mathcal{R}^C$.

3.1. Input Level Pre-training Adaptation

Style transfer techniques allow the alignment of the visual appearance of samples from the source to the target domain, thus increasing the model's generalization capabilities. These techniques have been widely used on color data, but their applicability to depth representations has not been explored.

In our setting, the segmentation network is pre-trained by applying image-to-image translation on both the color and depth data of the input samples from $\mathcal{D}_s$. In particular, we opted for a frequency domain image translation algorithm for style transfer, i.e., FDA [36], preserving the advantage of using a simple module that does not require training complex adversarial deep networks modules for image translation. This allows for avoiding additional computational complexity at inference time and for keeping simpler the training procedure. We retrieve the frequency space representation of a sample $x_i \in \mathbb{R}^{HxWxC}$ through Fast Fourier Transform (FFT) [36] as:

$$\mathcal{F}(x_i)(u,v,c) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x_i(h,w,c)e^{-j2\pi(\frac{h}{H}u+\frac{w}{W}v)} \tag{1}$$

The frequency space signal $\mathcal{F}(x_i)$ can be decomposed into an amplitude spectrum $\mathcal{A}(x_i) \in \mathbb{R}^{HxWxC}$ and a phase angle $\mathcal{P}(x_i) \in \mathbb{R}^{HxWxC}$. Low-level distributions reflect the image style, thus replacing lower frequencies in the source spectrum with the target ones - in our case we used the average of a small set of target samples - can improve the domain adaptation performances:

$$x^{s\to t} = \mathcal{F}^{-1}(\mathcal{A}_{low}(x^t) + \mathcal{A}_{high}(x^s), \mathcal{P}(x^s)) \tag{2}$$

where x can be both the color image x_{rgb} or the depth map x_d. The fraction of the replaced low-level details is set by the parameter β, which controls the amplitude window.

In particular, different choices of β affect the source representation: a larger β increases the domain translation effect but also introduces visual artifacts.

The predicted semantics should not be influenced by the sensor's properties or other causes of variation linked to the acquisition procedure. Yet the generalization ability of the network is affected by these aspects [36]. Hence, in domain adaptation settings, perceptually minor changes in the low-level data might result in a considerable decline in the trained model's performance. Depth maps are influenced not only by the acquisition sensor, which can be based on completely different technologies, e.g., time-of-flight, active or passive-stereo, etc., leading to very different frequency responses, but also by the characteristics of the scene and by the camera viewpoint. Performing alignment of low-frequency coefficients allows getting a better invariance to the characteristics of the sensors and to the

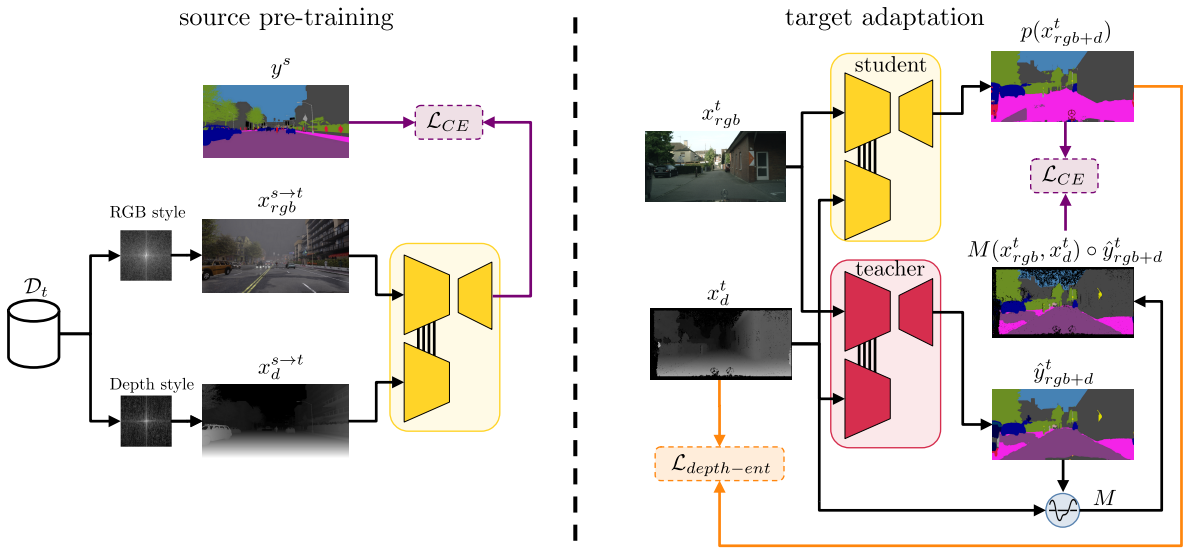

Figure 2. Overview of the training procedure for the proposed method. First, the network is trained on the (synthetic) supervised source dataset, while style transfer is applied both on RGB and depth images (Sec. 3.1). Then, the model is trained on the (real) unsupervised target dataset via the masked self-training strategy and depth entropy minimization (Sec. 3.3). In both steps, the fusion between RGB and depth features is performed with cross-modality attention (Sec. 3.2).

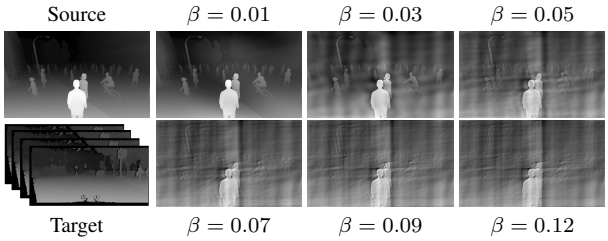

Figure 3. Effect of the Fourier domain style transfer applied on depth images, whereas $\beta = 0$ is equivalent to no transfer and $\beta = 1$ to the transfer of the full target amplitude.

depth values distribution due to the employed camera. Furthermore, when performing synthetic-to-real adaptation the difference between noise-free ideal synthetic depth and the inaccurate data of most real-world depth acquisition strategies is another critical challenge. It can be mitigated by the frequency domain adaptation forcing the network to focus more on the semantic structure of the scene than on acquisition device peculiarities.

As we can see from Figure 3, the approach allows to better align the depth ranges and makes the synthetic source data less "clean", matching the fact that real-world target data computed with stereo vision has more artifacts and a less sharp distribution (see Section 4.2 for more details on the employed datasets). On the other side, using an excessively large β introduces visual artifacts that can affect the network's performance.

3.2. Feature Level Adaptation

Cross-Modalities Attention In order to perform feature exchange, the multi-head attention module is shared between the two input modalities x_{rgb} and x_d. Cross-modal attention was proposed to provide latent adaptation across modalities in visual-text multimodal tasks [30]:

$$\text{Cross-Att}_{\beta\to\alpha}(Q_\alpha, K_\beta, V_\beta) = softmax\left(\frac{Q_\alpha K_\beta^T}{\sqrt{d_{head}}}\right) V_\beta \tag{3}$$

where Q is the query, K is the key, V is the value and d_{head} is the dimension of the head [32]. Transformer attention can be seen as an information retrieval mechanism: the generated query is specified from a key that returns a value. Differently from previous Transformer-based fusion approaches which considered feature-fusion at the end of each attention head [14], following the idea introduced in [2], the proposed framework acts directly at the core of the architecture by swapping the keys as in the following equation:

$$\text{Att}(Q_{rgb}, K_d, V_{rgb}) = softmax\left(\frac{Q_{rgb} K_d^T}{\sqrt{d_{head}}}\right) V_{rgb} \tag{4}$$

Unlike [2] where they investigated the effect of the interaction across each modality, we focused on the impact of this operation on the generalization ability of the network, i.e., multi-modal pre-training. We assume the interaction between the two modalities should be consistent across different data distributions, as proved by the ablation studies in

Section 4.5. The multi-head mixed attention feature x_{rgb+d} (Eq. 4) is served at the decoder level as in [34].

3.3. Output Level Adaptation

Self-Training During target adaptation, pseudo-labels $\hat{y}^t_{rgb+d}$ are assigned to unlabeled target data by the model through a self-training procedure. These labels are not always accurate, and filtering them can improve the performance of the model. The adopted filtering function uses a combination of probability thresholds and top-k filtering to select high-quality pseudo-labeled data points for training and discard the unlabeled ones. Following [36], we considered valid only the predictions with a confidence score above 0.9 or the ones that are within the top-66% confidence values.

Furthermore, we took under consideration that in the model pre-trained on the source dataset -thus synthetic data- depth maps are typically ground truth rendered maps free from noise, artifacts and missing points. Real-world depth maps, especially if obtained through stereo-matching, as in the case of the cityscapes dataset, are corrupted by noise and have missing disparity values due to occlusions or to limitations of the stereo-matching strategy. In a vanilla multimodal approach, color and depth features equally contribute to the loss term. In our setting, the depth information does not directly produce the semantic data estimation but contributes to the attention mechanism employed to construct the actual feature. Nevertheless, masking pixels with missing or corrupted depth data during the computation of pseudo-labels has the potential to significantly aid in the adaptation process.

The loss driven by the self-teaching module is thus computed as:

$$\mathcal{L}_{pseudo} = L_{CE}(p(x^t_{rgb+d}), M(x^t_{rgb}, x^t_d) \circ \hat{y}^t_{rgb+d}) \quad (5)$$

where the pseudo-label selection mask is:

$$M(x_{rgb}, x_d) = \begin{cases} 1 & \text{if } x_d \text{ is valid and} \\ & [p > 0.9 \text{ or } p \in \text{Top-66\%}] \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Notice how the resulting binary mask is a combination of the probability-based and top-k masks with the depth validity constraint. By employing depth as an indicator of the model's uncertainty, it becomes possible to enhance the model without the need for additional hyperparameters.

Depth Entropy minimization Pseudo-labels allow network training on unlabeled target data imitating the label's existence. However, it can be experimentally noticed that, as training progresses, after a certain point the learning curve begins to decline [5]. Initially, self-training serves as a means to narrow the discrepancy between the knowledge obtained from the source dataset $\mathcal{D}_s$ and that required for satisfactory performance on the target dataset $\mathcal{D}_t$. However, as the training proceeds, the network becomes overly self-assured in its predictions, thereby diminishing its efficacy and resulting in increased misclassification errors.

Under the assumption that real-world depth data often contains inconsistencies, we have chosen to assign greater weight to images captured at shorter distances, on the basis that disparity values for distant objects are more prone to error. Furthermore, close objects have a higher resolution in terms of pixels in the image and thus are better represented and easier to be properly classified also in color data.

To exploit this, we developed an entropy minimization strategy exploiting distance information through the disparity map. Recall that the disparity map, which is the typical output of stereo vision methods, is inversely proportional to depth data. Therefore a smaller disparity corresponds to objects that are captured with a lower spatial resolution by pinhole cameras and have less reliable depth values.

We started from the standard entropy minimization target proposed in [33] to aid the domain adaptation task:

$$\mathcal{L}_{ent}(x) = -\sum_{c}^{C} p(x)^{(c)} log p(x)^{(c)} \quad (7)$$

We modified the loss by adding a weighting term that depends on the distance from the camera giving more relevance to close points. We found that simply weighting the entropy loss with the disparity values $x^{(w,h)}_{disp}$ led to the best performances:

$$\mathcal{L}_{depth-ent} = \sum_{h}^{H} \sum_{w}^{W} \mathcal{L}_{ent}(x_{rgb+d})^{(w,h)} * (x^{(w,h)}_{disp}) \quad (8)$$

4. Results

In this section, we introduce the experimental framework and the employed datasets, then we present the numerical results obtained by our method. Finally we present some ablation studies to evaluate the impact of the different components of the approach.

4.1. Datasets

We evaluate our method on two synthetic-to-real road scene segmentation scenarios: (a) SYNTHIA-to-Cityscapes and (b) SELMA-to-Cityscapes. For the supervised pre-training on source data, we employed the widely used SYNTHIA dataset [22] that contains 9400 total samples with a resolution of 1280×760. Furthermore, we made some tests also with the more recent SELMA dataset [26], which comprises 31k scenes with a resolution of 1280x640 in a wide range of different acquisition conditions. While well-known domain adaptation datasets like GTAV [20] (Synthetic) or BDD100K [39] (Real) cannot be utilized in our

Method	Backbone	road	side.	build.	wall*	fence*	pole*	light	sign	vege.	sky	pers.	rider	car	bus	motor	bike	mIoU$_{16}$	mIoU$_{13}$
SFDA [15]	ResNet-50	81.9	44.9	81.7	4.0	0.5	26.2	3.3	10.7	86.3	89.4	37.9	13.4	80.6	25.6	9.6	31.3	39.2	45.9
URMA [5]	ResNet-101	59.3	24.6	77.0	14.0	1.8	31.5	18.3	32.0	83.1	80.4	46.3	17.8	76.7	17.0	18.5	34.6	39.6	45.0
DT+AC [35]	ResNet-101	77.5	37.4	80.5	13.5	1.7	30.5	24.8	19.7	79.1	83.0	49.1	20.8	76.2	12.1	16.5	46.1	41.8	47.9
LD [38]	ResNet-101	77.1	33.4	79.4	5.8	0.5	23.7	5.2	13.0	81.8	78.3	56.1	21.6	80.3	49.6	28.0	48.1	42.6	50.1
HCL [10]	ResNet-101	80.9	34.9	76.7	6.6	0.2	36.1	20.1	28.2	79.1	83.1	55.6	25.6	78.8	32.7	24.1	32.7	43.5	50.2
SFUDA [37]	ResNet-101	90.9	45.5	80.8	3.6	0.5	28.6	8.5	26.1	83.4	83.6	55.2	25.0	79.5	32.8	20.2	43.9	44.2	51.9
DT-ST [42]	ResNet-101	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	58.8
SOMAN+cPAE [13]	ResNet-101	90.5	50.0	81.6	13.3	2.8	34.7	25.7	33.1	83.8	89.2	66.0	34.9	85.3	53.4	46.1	46.6	52.0	60.1
Source Only RGB	MiT-B5	28.5	19.7	56.7	3.4	0.2	39.1	34.9	18.0	81.0	86.1	64.0	11.6	82.6	28.2	7.5	29.4	36.9	42.2
RGB + FDA [36]	MiT-B5	41.1	27.2	60.6	6.3	0.3	42.7	31.0	27.2	82.2	87.8	65.8	15.0	61.4	38.9	9.0	30.8	39.2	44.5
MISFIT (Ours)	MiT-B5	80.2	38.5	85.9	30.3	1.2	52.3	56.8	29.0	89.9	88.3	68.1	10.8	92.1	69.0	26.3	52.6	**54.5**	**60.6**

Table 1. Semantic segmentation results for the SYNTHIA-to-Cityscapes source-free adaptation task. $mIoU_{13}$ denotes performance over 13 classes excluding those marked with *.

Method	road	side.	build.	wall	fence	pole	light	sign	vege.	terr.*	sky	pers.	rider	car	truck*	bus	train*	motor	bike	mIoU$_{19}$	mIoU$_{16}$
Source Only RGB	70.9	45.7	71.2	12.4	7.8	37.4	37.5	35.4	84.8	24.8	81.7	65.9	23.4	65.7	11.5	21.5	2.8	41.0	45.7	41.4	46.8
RGB + FDA [36]	64.0	47.2	60.0	7.1	6.8	41.2	38.6	43.9	84.1	20.3	79.0	66.8	23.7	73.4	18.6	34.7	4.1	37.5	39.8	41.6	46.7
MISFIT (w/o $\mathcal{L}_{depth-ent}$)	74.2	61.8	66.0	8.6	15.9	51.3	55.2	60.7	87.0	24.4	73.0	72.2	32.7	86.8	31.6	62.9	0.0	39.9	57.1	50.6	56.6
MISFIT (Ours)	76.2	63.2	68.7	5.6	13.7	50.9	57.2	60.8	87.2	21.6	89.8	72.2	33.3	86.6	30.0	54.8	7.8	43.9	58.2	**51.7**	**57.6**

Table 2. Semantic segmentation results for the SELMA-to-Cityscapes source-free adaptation task. $mIoU_{16}$ denotes performance over 16 classes - corresponding to SYNTHIA classes - excluding those marked with *.

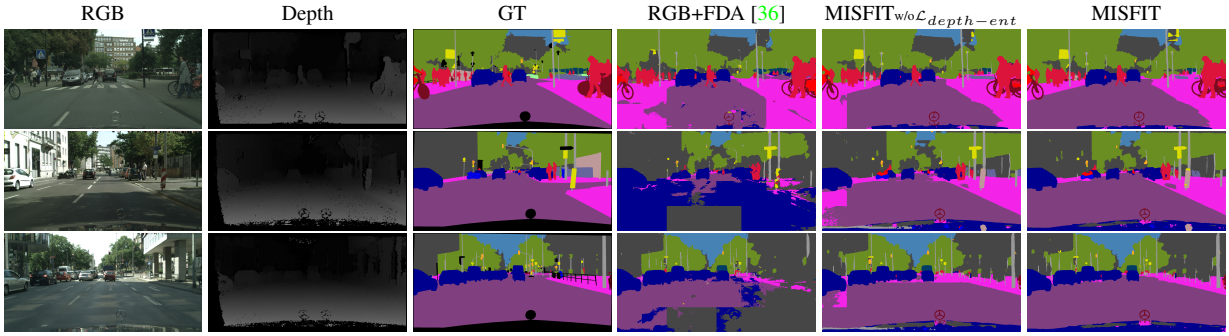

Figure 4. Qualitative semantic segmentation results for the SYNTHIA-to-Cityscapes adaptation task.

approach due to the absence of depth maps, employing recently available datasets like SELMA offers a distinct advantage. The primary benefit lies in SELMA's provision of all the 19 classes present in Cityscapes, enabling a better matching between the datasets. As a target real-world dataset we used Cityscapes [4], which is the most common benchmark for semantic segmentation in the driving environment. It includes 2975 training samples and 500 validation ones. Each image is provided with the associated depth map computed with stereo vision, while the resolution is 2048×1024. Notice that depth maps are the result of a stereo matching algorithm and consequently present many issues and artifacts, differently from the ones of SYNTHIA and SELMA, that contain ground truth data extracted from the rendering engine. This makes the domain adaptation task more challenging since it must adapt both from synthetic to real data and from ground truth depth to stereo vision data.

4.2. Implementation Details

We adopted SegFormer [34] as the basic segmentation framework since it is a widespread well-performing approach based on vision transformers. Furthermore, previous studies have demonstrated its small generalization gap [18]. In our framework, the encoder is shared between the two modalities thus reducing the number of parameters to be estimated and at the same time supporting depth attention processing and multimodal fusion as described in Section 3.2. The architecture is pre-trained on source data using the Adam optimizer for 40 epochs (160k iterations) with batch size 4 and learning rate starting from $6e-5$ with a weight decay rate of 0.01. For the unsupervised target adaptation, where only target data is used, the batch size was set equal to 2. We employed the same data augmentation and input resolution (512×512) as those in [34]. For generating

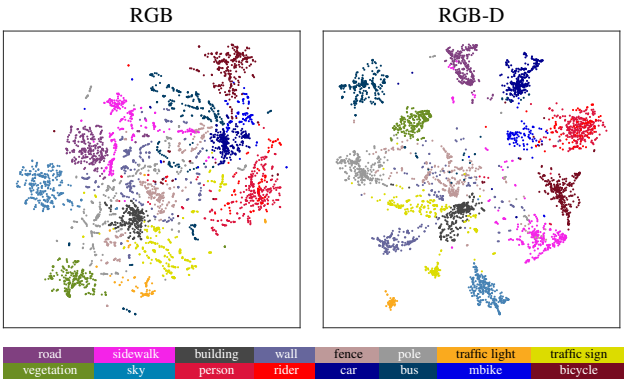

Figure 5. T-SNE on the SYNTHIA-to-Cityscapes setting: (Left) RGB-only adaptation (Right) RGB-D adaptation from our method MISFIT.

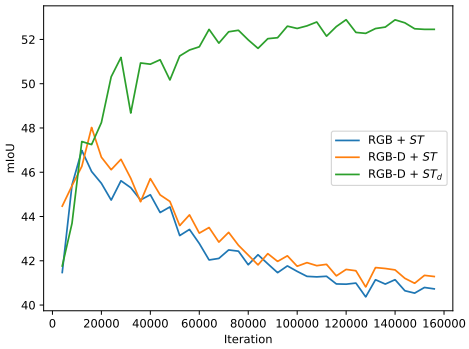

Figure 6. Comparison of learning curves for various self-training setups: standard pseudo-label masking based on confidence scores and ST_d employing depth-based masking on top.

the target pseudo-labels, we utilize the depth-masked self-training strategy of Section 3.3, where the teacher model is updated every 100 steps with update momentum 0.99. The FFT style transfer parameters are empirically chosen as $\beta = 0.01$ for color and $\beta = 0.09$ for depth.

4.3. SYNTHIA-to-Cityscapes adaptation results

Table 1 shows the performances of our approach in the SYNTHIA-to-Cityscapes benchmark and compares it with the other source-free domain adaptation approaches from the literature. To showcase the effectiveness of our method, we compared it with several recent convolutional architecture-based methods. The table presents the state-of-the-art results for source-free domain adaptation, highlighting the superiority of our proposed approach.

In this setting, training on source color data leads to a relatively low accuracy of approximately 37%. Even when incorporating FDA-style transfer, there is only a marginal improvement observed, reaching up to 39.2%. On the other hand, our approach, which leverages multimodal domain adaptation techniques, achieves remarkable results, as evidenced by an impressive mIoU score of 54.5%. This performance surpasses most of the competing methods by a significant margin of over 10%. Furthermore, our method demonstrates consistent and outstanding performance across different classes, particularly excelling in classes such as bus, pole, and wall. Notably, our scores in the wall class are more than double the best scores achieved by our competitors.

The visual results align with the numerical evaluation, as evident in Figure 4, where challenging objects such as the bike with the rider and the poles are accurately identified. At the same time, the domain shift causes issues on textured areas like the road or the sidewalk that need all the components of the proposed approach to be correctly handled. Notably, as observed in the Figure, the regions in question encounter difficulties when only certain components are employed. This outcome is reinforced by the improved disambiguation observed between the road and sidewalk classes in the T-SNE representation (see Figure 5). Moreover, the numerical results from Table 1 validate the improved class separation observed for the driving classes such as bus, motorbike, and bike.

4.4. SELMA-to-Cityscapes adaptation results

The evaluation of the SELMA-to-Cityscapes benchmark is presented in Table 2. On this recent dataset there are no results for the source-free setting from previous works, so we can compare the performances of our approach only with some baselines. The training on source color data in this setting leads to an accuracy of 41.4% and the FDA style transfer on color data has a very limited impact (the gain is 0.2%). The higher performance observed when training on the source data, combined with the limited impact of style transfer, indicates that the dataset, compared to other benchmarks, exhibits a higher visual quality that closely aligns with real-world data. Our multimodal domain adaptation approach allows to increase the mIoU to 51.7% with a remarkable gain of more than 10% over the RGB baseline. Furthermore, it is worth mentioning that the depth-driven entropy minimization loss exhibits a slightly larger impact. This is particularly significant considering that while the performance may be comparatively lower in the SYNTHIA-to-Cityscapes scenario, SELMA encompasses all 19 classes present in Cityscapes.

4.5. Ablation Studies

Modules of the proposed framework First of all, we performed some ablation studies on the SYNTHIA-to-Cityscapes benchmark to evaluate the impact of the various modules of the proposed approach on the final performances. Results are shown in Table 3: as already

pointed out simply training on source color data and testing on the target dataset leads to an accuracy of around 37%, which represents the starting point. Adding a self-teaching step with target color data allows improving performances of almost 4%. Moving to the multimodal setting, the source-only training accuracy exploiting depth data is 39.8%. The depth-aware self-teaching scheme proposed in Section 3.3 allows for an impressive improvement up to 52.5% of mIoU. In particular, as visible in Figure 6, incorporating depth-based masking serves as a regularization method, enhancing training performance compared to the ones on confidence only (refer to the *Supplementary Material* for details). During standard self-training, the approach initially learns effectively, reaching an accuracy of 48.0%. Nevertheless, it becomes excessively confident in inaccurate predictions, leading to a decline in the learning curve —a trend consistent with observations in [5]. Further adding the input level style transfer boosts performances to 54%. Finally adding also the entropy minimization target allows us to get the full model accuracy of 54.5%.

Mode	ST	ST_d	Style	Entropy	mIoU
RGB					36.93
RGB	✓				40.59
RGB+D					39.79
RGB+D		✓			52.50
RGB+D		✓	✓		54.00
RGB+D		✓	✓	✓	**54.52**

Table 3. Impact of the various proposed adaptation framework modules, performed on the SYNTHIA-to-Cityscapes setting. When all modules are disabled, it corresponds to source-only.

Cross-Modality Attention We conducted tests on the impact of crossing between the two modalities (i.e., color and depth) in the source-only setting. Supported by the fact that previous works exploit asymmetric feature fusion [21], we tested the transformer cross-attention swap in the direction of color [30] and the key-swap algorithm [2]. The key-swapping strategy achieves better results and for this reason, it has been selected for our approach. More in detail, the informative content of the depth keys is able to gain a 2.86% over the use of color alone in the source-only setup (see Table 4).

Input Depth Style In order to prove the effectiveness of the style transfer method we compare the performances of the algorithm described in Section 3.1 when applying it to color data or to both modalities in the SYNTHIA-to-Cityscapes benchmark. The target style is transferred to the source domain to perform network pre-training. Results are shown in Table 5: if working with color data alone the style transfer allows for a gain of around 2%. Multimodal data allows for a higher starting point and adding the style transfer on both modalities allows to further boost performances from 39.8 to 41%.

Mode	Cross-Attention	mIoU
RGB		36.93
RGB+D	RGB→D	29.83
RGB+D	D→RGB	38.02
RGB+D	Key Swap	**39.79**

Table 4. Ablation on source-only generalization ability of cross-modalities attention, performed on the SYNTHIA-to-Cityscapes setting.

Mode	Style-Transfer	mIoU
RGB		36.9
RGB	✓	39.2
RGB+D		39.8
RGB+D	✓	**41.0**

Table 5. Ablation on source-only generalization ability with different input-level style-transfers, performed on the SYNTHIA-to-Cityscapes setting.

5. Conclusions

Ultimately, the use of multimodal information for semantic segmentation is a relevant area of research that can help to address the challenges of adapting segmentation models to new domains. However, the domain adaptation capabilities of pre-trained multimodal schemes have seldom been explored, especially in conjunction with vision transformer architectures that represent the current state-of-the-art in many vision tasks. By leveraging multiple adaptation strategies driven by the complementary information provided by depth data, the proposed multimodal framework allows for improving the robustness and generalization ability of segmentation models, enabling them to be used in a wider range of applications. Experimental results show how it achieves state-of-the-art performance in the challenging source-free domain adaptation setting.

Further research will be devoted to improving the exploitation of depth data in transformer-based segmentation models and to the development of domain adaptation strategies explicitly targeted at the inconsistencies between ground truth and estimated depth data.

Acknowledgment

This work was supported in part by the European Union through the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (Program "RESTART") under Grant PE0000001.

References

[1] Peshal Agarwal, Danda Pani Paudel, Jan-Nico Zaech, and Luc Van Gool. Unsupervised robust domain adaptation without source data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2009–2018, 2022. 3

[2] Francesco Barbato, Giulia Rizzoli, and Pietro Zanuttigh. Depthformer: Multimodal positional encodings and cross-input attention for transformer-based segmentation networks. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1, 2, 4, 8

[3] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2090–2099, 2019. 2

[4] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[5] Francois Fleuret et al. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9613–9623, 2021. 3, 5, 6, 8

[6] 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. 2

[7] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 1, 2

[8] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 372–391. Springer, 2022. 2

[9] Sijie Hu, Fabien Bonardi, Samia Bouchafa, and Désiré Sidibé. Multi-modal unsupervised domain adaptation for semantic image segmentation. *Pattern Recognition*, page 109299, 2023. 2

[10] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34:3635–3649, 2021. 1, 3, 6

[11] Maximilian Jaritz, Tuan-Hung Vu, Raoul De Charette, Émilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1533–1544, 2022. 2

[12] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2

[13] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7046–7056, 2021. 1, 3, 6

[14] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838*, 2022. 1, 2, 4

[15] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021. 3, 6

[16] 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/CVF conference on computer vision and pattern recognition*, pages 2507–2516, 2019. 2

[17] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020. 2

[18] Jinyoung Park, Minseok Son, Sumin Lee, and Changick Kim. Dat: Domain adaptive transformer for domain adaptive semantic segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 4183–4187. IEEE, 2022. 1, 2, 6

[19] Qucheng Peng, Zhengming Ding, Lingjuan Lyu, Lichao Sun, and Chen Chen. Toward better target representation for source-free and black-box domain adaptation. *arXiv preprint arXiv:2208.10531*, 2022. 3

[20] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 5

[21] Giulia Rizzoli, Francesco Barbato, and Pietro Zanuttigh. Multimodal semantic segmentation in autonomous driving: A review of current approaches and future perspectives. *Technologies*, 10(4):90, 2022. 1, 2, 8

[22] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 5

[23] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531. IEEE, 2021. 2

[24] Donald Shenaj, Eros Fanì, Marco Toldo, Debora Caldarola, Antonio Tavera, Umberto Michieli, Marco Ciccone, Pietro

Zanuttigh, and Barbara Caputo. Learning across domains and devices: Style-driven source-free domain adaptation in clustered federated learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 444–454, 2023. 3

[25] Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16928–16937, 2022. 2

[26] 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. 5

[27] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation: a review. *Technologies*, 8(2):35, 2020. 1

[28] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 2

[29] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1456–1465, 2019. 2

[30] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 4, 8

[31] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5):1239–1285, 2020. 1, 2

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[33] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 2, 5

[34] 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. 2, 3, 5, 6

[35] Cheng-Yu Yang, Yuan-Jhe Kuo, and Chiou-Ting Hsu. Source free domain adaptation for semantic segmentation via distribution transfer and adaptive class-balanced self-training. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 3, 6

[36] 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. 2, 3, 5, 6

[37] Mucong Ye, Jing Zhang, Jinpeng Ouyang, and Ding Yuan. Source data-free unsupervised domain adaptation for semantic segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2233–2242, 2021. 3, 6

[38] Fuming You, Jingjing Li, Lei Zhu, Zhi Chen, and Zi Huang. Domain adaptive semantic segmentation without source data. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3293–3302, 2021. 1, 3, 6

[39] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 5

[40] Zhiqi Yu, Jingjing Li, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *arXiv preprint arXiv:2302.11803*, 2023. 2

[41] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. 2

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11733–11743, 2023. 3, 6

[43] Yang Zou, Zhiding Yu, BVK 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. 2

[44] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5982–5991, 2019. 2