CrossMatch: Source-Free Domain Adaptive Semantic Segmentation via Cross-Modal Consistency Training

Yifang Yin[1], Wenmiao Hu[2,4], Zhenguang Liu[3]*, Guanfeng Wang[4]*, Shili Xiang[1], Roger Zimmermann[2]
[1]Institute for Infocomm Research, A*STAR [2]National University of Singapore
[3]Zhejiang Gongshang University [4]Grabtaxi Holdings Pte. Ltd.
{yin_yifang, sxiang}@i2r.a-star.edu.sg, hu.wenmiao@u.nus.edu
liuzhenguang2008@gmail.com, guanfeng.wang@grab.com, rogerz@comp.nus.edu.sg

Abstract

Source-free domain adaptive semantic segmentation has gained increasing attention recently. It eases the requirement of full access to the source domain by transferring knowledge only from a well-trained source model. However, reducing the uncertainty of the target pseudo labels becomes inevitably more challenging without the supervision of the labeled source data. In this work, we propose a novel asymmetric two-stream architecture that learns more robustly from noisy pseudo labels. Our approach simultaneously conducts dual-head pseudo label denoising and cross-modal consistency regularization. Towards the former, we introduce a multimodal auxiliary network during training (and discard it during inference), which effectively enhances the pseudo labels' correctness by leveraging the guidance from the depth information. Towards the latter, we enforce a new cross-modal pixel-wise consistency between the predictions of the two streams, encouraging our model to behave smoothly for both modality variance and image perturbations. It serves as an effective regularization to further reduce the impact of the inaccurate pseudo labels in source-free unsupervised domain adaptation. Experiments on GTA5 → Cityscapes and SYNTHIA → Cityscapes benchmarks demonstrate the superiority of our proposed method, obtaining the new state-of-the-art mIoU of 57.7% and 57.5%, respectively.

1. Introduction

Semantic segmentation predicts pixel-level category labels to given scenes. Although deep neural networks have been widely adopted, attaining state-of-the-art performance relies mainly on the assumption that the training and testing data follow the same distribution [62, 32, 33]. This assumption is impractical as target scenarios often exhibit a distribution shift, *e.g.*, street scenes collected under a cross-city [11] or cross-weather [44] environment. Unsupervised domain adaptation (UDA) techniques have been proposed to address the domain shift problem, which aim at transferring the knowledge learned from a labeled source domain to an unlabeled target domain [48, 50, 69, 67]. However, one major limitation of such UDA approaches lies in the requirement for full access to the source dataset. In practice, the source data may be restricted from being shared due to proprietary, privacy, or profit related concerns [26].

*The corresponding authors.

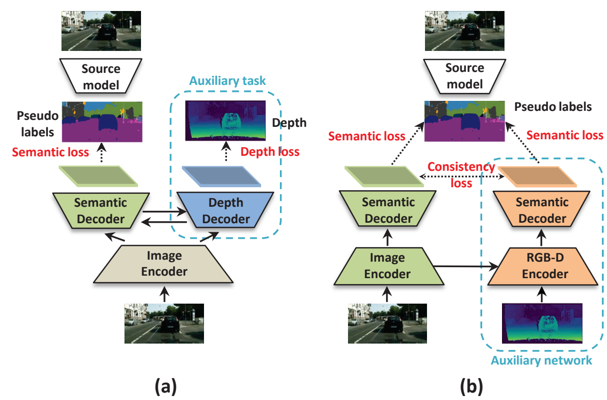

Figure 1. Comparison of our proposed framework with existing depth-aware semantic segmentation models. (a) Prior art mostly adopts a multitask learning framework by adding depth estimation as an auxiliary task. (b) We introduce a multimodal auxiliary network that takes depth modality as an additional input for effective pseudo label denoising and consistency regularization.

To cope with data sharing restrictions, recent efforts have investigated source-free domain adaptation, which transfers knowledge from a well-trained source model (rather than from the source data itself) to an unlabeled target domain [39, 31]. Early solutions introduce a generator to estimate the source domain based on the pre-trained source model [31], which can be used to generate fake source samples for supervision as in typical UDA. However, due to the lack of supervision from the real source domain, advanced techniques designed for typical UDA, such as depth-aware semantic segmentation and pseudo label denoising methods, may work less satisfactorily in a source-free setting.

With the above insights, we propose a novel two-stream segmentation network for source-free UDA. As shown in Figure 1 (a), existing depth-aware semantic segmentation for typical UDA mainly adopts a multitask learning framework where depth estimation is modeled as an auxiliary task [51, 53]. However, we observe through experiments that the regularization induced by the auxiliary task is quite limited for source-free UDA due to the lack of ground-truth semantic labels. It cannot effectively prevent the main segmentation network from overfitting to the incorrect over-confident pseudo labels of the target images. To solve this problem, we alternatively propose a multimodal auxiliary network, as shown in Figure 1 (b), which takes the depth information and the intermediate representations generated by the main stream image encoder as the input. We train both the main and the auxiliary streams on the segmentation task via self-training, and formulate an explicit cross-modal consistency loss between the output of the two streams for effective regularization. The benefits of our proposed segmentation network are threefold:

First, our inference-stage model consists of the main stream only, which is a unimodal model that infers from RGB images the same way as existing models. *Second*, the asymmetric design of our neural network introduces modality variance in addition to the typical input perturbations produced by data augmentation, dropouts, *etc*. On one hand, the auxiliary network better rectifies the pseudo labels with multimodal knowledge expansion [61]. On the other hand, the cross-modal consistency effectively transfers the knowledge learned from the multimodal auxiliary network to the unimodal main network. *Third*, our proposed framework has better feasibility compared to existing depth-aware UDA as ours only requires the depth information in the target domain. Without annotation cost, the depth information can be easily learned from video sequences or stereo images based on self-supervised depth estimation models [17, 71, 53]. Here we summarize our contributions as follows:

- We propose a novel *source-free* UDA framework by introducing a multimodal auxiliary network. It models the correlations between depth and semantics, and can be discarded completely at inference time.
- We enforce a cross-modal consistency between the predictions of the main and auxiliary streams with dual-head pseudo label denoising, to reduce the impact of inaccurate pseudo labels in *source-free* UDA.
- Our proposed method outperforms the prior art by a significant margin, obtaining an mIoU of 57.7% and 57.5% on the Cityscapes dataset when adapting from the GTA5 and SYNTHIA benchmarks, respectively.

2. Related Work

Unsupervised domain adaptation Unsupervised domain adaptation (UDA) aims to improve a model's performance on an unlabeled target domain by leveraging the features extracted from a labeled source domain [62]. Early works adopted adversarial training [18] to reduce the distribution mismatch between different domains [36, 15, 48, 50]. Efforts have been made on aligning the distributions at either the image level [21, 57], the intermediate feature level [11, 10] or the output level [48, 50]. Some recent attempts align the distributions in a class-wise manner in order to obtain a fine-grained feature alignment [36, 15]. However, these methods rely on cumbersome adversarial training that requires access to the source data.

UDA via self-training Pseudo label refinement under a self-training framework has achieved competitive results in the field of UDA for semantic segmentation [30, 68, 70, 23]. Early methods selected highly confident predictions as pseudo labels based on a confidence threshold [73, 72]. To improve the robustness of the pseudo labels, efforts have been made on prediction ensembling [6, 63], pseudo label denoising [37, 28, 45, 67], training sample reweighting [69], augmentation consistency [1, 38], leveraging high-resolution images [24], and pixel-level contrastive learning [58]. However, these approaches also rely on the source-target co-existence to retain task-specific source knowledge with self-training.

Source-free UDA Kundu *et al*. [26] focused on source model generalization and developed a multi-head framework trained by extending the source data with diverse data augmentations. Teja and Fleuret [39] focused on target domain adaptation and proposed to reduce the prediction uncertainty by feature corruption with entropy regularization. Liu *et al*. [31] leveraged a generator to estimate the source data distribution, based on which fake samples were synthesized for training. Qiu *et al*. [40] proposed to generate per-class prototypes based on a source prototype generator, which is used to align the pseudo-labeled target data based on contrastive learning. To the best of our knowledge, the prior approaches [64, 66] all focused on unimodal models. Inspired by existing work on cross-modal modeling between image features and acoustic clues [65], edge maps [34], or LiDAR points [25] in different applications, we develop a new cross-modal pseudo label denoising network for depth-aware source-free UDA.

Depth-aware UDA Motivated by multitask learning, depth estimation has been adopted as an auxiliary task to improve UDA for semantic segmentation [49, 9, 43, 3, 22]. The labels for depth estimation are mostly derived by self-supervised models using stereo pairs [16, 17] or video se-

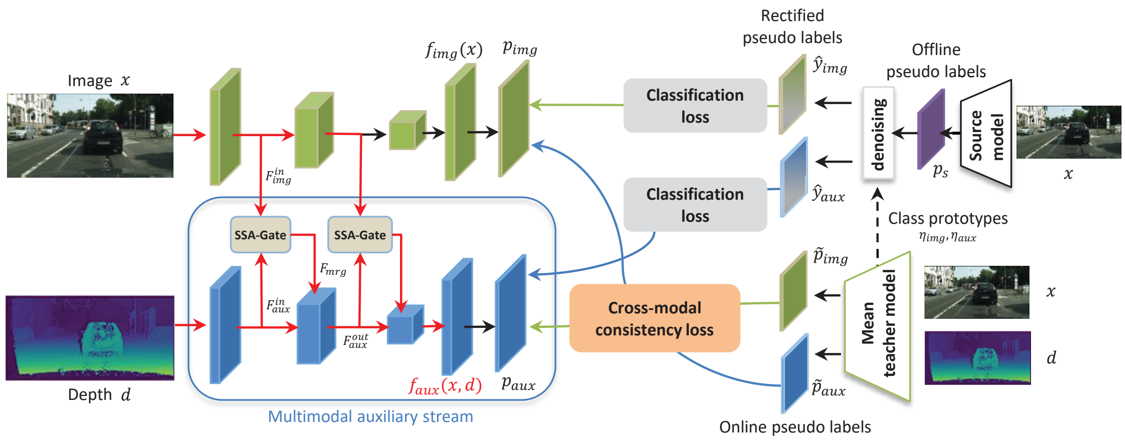

Figure 2. Illustration of our proposed two-stream segmentation network for source-free UDA.

quences [71]. The correlations between depth and semantics are next modeled by attention-based feature fusion [51, 53]. The depth distribution in different categories can be utilized to further reduce the domain gap [56]. However, these methods rely on the access to the source domain and assume the source and target images are available in stereo pairs or video sequences.

3. Problem Formulation

Efforts on source-free domain adaptive semantic segmentation can be divided into 1) vendor-side domain generalization, and 2) client-side domain adaptation [26]. The vendor and the client have access to the labeled source and the unlabeled target datasets, respectively. The goal of the vendor is to train a source model with good generalization ability to unseen domains [27]. This trained source model is next passed to the client to be adapted to the unlabeled target domain via self-training [39, 31].

In this work, we propose to improve client-side domain adaptation by leveraging depth information as the auxiliary modality. Let $\mathcal{X} = \{(x_i, d_i)\}_{i=1}^{n}$ denote the target dataset where (x_i, d_i) represent the RGB and the depth modality of the i-th sample, respectively. Our goal is to adapt a unimodal source model $h^s(x)$ to a unimodal target model $h^t(x)$ more robustly via a multimodal auxiliary network. To achieve this goal, we present a novel two-stream neural network with a main stream and an auxiliary stream that perform semantic segmentation based on RGB and RGB-D modalities, respectively. Facilitated by the depth modality, pseudo labels obtained from the source model can be better rectified, leading to improved source-free UDA performance. Moreover, the auxiliary stream is only required during training, and will be discarded at inference time. Thus, our inference-stage model shares the same network architecture (*e.g.*, DeepLabv2 [4]) but obtains improved segmentation results compared to the prior art.

4. Approach

We follow the pseudo-label based self-training strategies to train our source-free UDA model [26]. Target samples are passed through the source model to generate a set of pseudo labels that are used to supervise the network. One main challenge in a self-training framework is reducing the uncertainty of the pseudo labels for the target images. To tackle this challenge, we propose to denoise the offline target pseudo labels with online cross-modal consistency training. Next, we introduce the technical details of our proposed framework.

4.1. Two-stream Segmentation Network

The overall architecture of our proposed asymmetric two-stream segmentation network is shown in Figure 2. The main stream is unimodal, which takes RGB images as the only input, and can be implemented by any of the existing segmentation models such as DeepLabv2. The auxiliary stream is multimodal, which ingests depth and the intermediate features generated by the main stream image encoder to exploit the correlations between the depth and semantic information. To achieve this, we build upon the Separation-and-Aggregation Gate (SA-Gate) [8] and present a single-sided SA-Gate, termed SSA-Gate, which is placed after each of the encoder blocks. Formally, let F_{img}^{in} and F_{aux}^{in} denote the input features of the SSA-Gate from the main and auxiliary streams, respectively. SSA-Gate first recalibrates the input features with the help from the other modality by

$$
\begin{aligned}
F_{img}^{rec} &= F_{img}^{in} + Attn^{a}(F_{img}^{in} || F_{aux}^{in}) \circledast F_{aux}^{in} \\
F_{aux}^{rec} &= F_{aux}^{in} + Attn^{i}(F_{img}^{in} || F_{aux}^{in}) \circledast F_{img}^{in}
\end{aligned}
\tag{1}
$$

where $F_{img}^{in} || F_{aux}^{in}$ is the concatenation of the input features along the channel dimension. $Attn^a$ and $Attn^i$ compute the channel-wise attention for F_{aux}^{in} and F_{img}^{in}, respectively, and $\circledast$ denotes the channel-wise multiplication. Next, SSA-Gate merges the features from the two streams based on the spatial-wise gates proposed in [8]. Let F_{mrg} denote the merged feature, SSA-Gate updates the feature of the auxiliary stream as $F_{aux}^{out} = 0.5 \cdot (F_{aux}^{in} + F_{mrg})$ and keeps the feature in the main stream unchanged. With known camera parameters, we follow prior work [5, 8] and extract the HHA representation, which encodes the depth image with three channels of horizontal disparity, height above ground, and the angle of the pixel's local surface normal, as the input of our target network [19]. According to previous studies [5, 8], the HHA representation is more effective for semantic segmentation tasks. Alternatively, the 1-channel disparity maps can be directly used as the input to our framework if the camera parameters are not available.

4.2. Dual-head Pseudo Label Denoising with Cross-modal Consistency Regularization

Given a target sample (x, d), we use $f_{img}(x)$ and $f_{aux}(x, d)$ to denote the features extracted by the main and auxiliary streams as shown in Figure 2. The extracted features are next passed to the respective classifiers g_{img} and g_{aux} to obtain the predictions p_{img} and p_{aux}. A mean-teacher model [47] is maintained whose parameters are updated as the exponential moving average of the parameters of the target network. This is used to generate more reliable online pseudo labels, denoted as $\tilde{p}_{img}$ and $\tilde{p}_{aux}$. Offline pseudo labels are generated using the source model based on RGB images only, *i.e.*, $p_s = h^s(x)$. Next, we will introduce how to formulate the objectives to optimize our proposed framework.

4.2.1 Cross-modal Consistency Training

Consistency regularization is a popular and essential technique in semi-supervised learning [60, 46]. Based on the model smoothness assumption, model predictions should be constrained to be invariant to small perturbations of either inputs or model hidden states [38], which can be introduced by data augmentation, dropouts, *etc*. To prevent the target model from overfitting to the noisy pseudo labels, we present a new cross-modal consistency regularization loss that works effectively with pseudo labeling in source-free UDA. The predictions for pixels with low-confidence pseudo labels tend to be more sensitive to input perturbations [69]. Thus, the impact of the noise in pseudo labels can be significantly reduced by enforcing a consistency regularization between the predictions of the two streams.

Given an unlabeled target image x, we pass it through the source model to generate the soft pseudo labels $p_s^{(i,k)}$. The hard pseudo labels $\hat{y}^{(i,k)}$ are computed as

$$\hat{y}^{(i,k)} = \begin{cases} 1 & \text{if } k = \arg\max_{k'} p_s^{(i,k')} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $p_s^{(i,k)}$ represents the softmax probability of pixel $x^{(i)}$ belonging to the k-th class. Thereafter, the classification loss can be computed based on $\hat{y}^{(i,k)}$ as

$$\ell_{cla} = \ell_{ce}(\hat{y}, p_{img}) + \ell_{ce}(\hat{y}, p_{aux}) \tag{3}$$

where $\ell_{ce}(\hat{y}, p) = -\sum_{i=1}^{H\times W} \sum_{k=1}^{K} \hat{y}^{(i,k)} \log p^{(i,k)}$ is the cross-entropy loss. p_{img} and p_{aux} are the predicted outputs of the main and auxiliary streams, respectively. In addition to the pseudo labeling, we introduce a cross-modal consistency loss to regularize the output between the two streams. The goal is to reduce the impact of inaccurate pseudo labels, and this consistency loss is formulated as

$$\ell_{reg} = \mathcal{D}_{kl}(\tilde{p}_{aux}||p_{img}) + \mathcal{D}_{kl}(\tilde{p}_{img}||p_{aux}) \tag{4}$$

where $\tilde{p}_{img}$ and $\tilde{p}_{aux}$ are the predicted outputs of the mean-teacher model, and $\mathcal{D}_{kl}(\tilde{p}_{aux}||p_{img}) = -\sum_{i=1}^{H\times W} \tilde{p}_{aux}^{(i)} \log(p_{img}^{(i)}/\tilde{p}_{aux}^{(i)})$ is the Kullback Leibler (KL) divergence. We perturb the input based on strong and weak augmentations, and feed them to the target network and its mean-teacher model, respectively. Since $\tilde{p}_{img}$ and $\tilde{p}_{aux}$ are generated based on weak augmented views, they are more reliable. They thus can be used as online soft pseudo labels to regularize the predictions p_{img} and p_{aux} inferred over the strong augmented views.

In addition to data augmentations, recall that $p_{img} = g_{img}(f_{img}(x))$ and $p_{aux} = g_{aux}(f_{aux}(x, d))$ also predict based on different input modalities. Therefore, our proposed regularization loss enforces that the target network gives consistent predictions not only for small perturbations but also over cross-modal views.

4.2.2 Dual-head Pseudo Label Denoising

Though the pseudo labels p_s generated by the source model can be directly used to train the target network, rectifying p_s from a parallel aspect to consistency training will gain additional benefits. To this end, we adapt a recent state-of-the-art prototypical pseudo label denoising method [67] to our framework. This approach fixes p_s and rectifies p_s based on class-wise dynamic weights ω as

$$\hat{p}_s^{(i,k)} = \frac{\exp\left(\omega^{(i,k)} \cdot p_s^{(i,k)}\right)}{\sum_{k'=1}^{K} \exp\left(\omega^{(i,k')} \cdot p_s^{(i,k')}\right)} \tag{5}$$

where $p_s^{(i,k)}$ and $\hat{p}_s^{(i,k)}$ represent the softmax probability of pixel $x^{(i)}$ belonging to the k-th class before and after denoising. We perform prototypical pseudo label denoising for the main and the auxiliary streams separately. Take the main stream as an example, let $f_{img}(x)^{(i)}$ represent the feature at pixel i. The weights ω_{img} are updated in each training epoch based on the feature distance to the class prototypes by

$$\omega_{img}^{(i,k)} = \frac{\exp\left(-||\tilde{f}_{img}(x)^{(i)} - \eta_{img}^{(k)}||/\tau\right)}{\sum_{k'=1}^{K} \exp\left(-||\tilde{f}_{img}(x)^{(i)} - \eta_{img}^{(k')}||/\tau\right)} \tag{6}$$

where $\eta_{img}^{(k)}$ is the prototype (*i.e.*, the feature centroid) of class k in the main stream. We use $\tilde{f}_{img}$ (*i.e.*, the image encoder in the mean-teacher model) instead of f_{img}, as we desire a more reliable feature estimation for the input sample. τ is the softmax temperature empirically set to 1. Similarly, we maintain class prototypes $\eta_{aux}^{(k)}$ for the auxiliary stream, compute ω_{aux} based on $\tilde{f}_{aux}(x, d)$ and $\eta_{aux}^{(k)}$, and correct p_s based on ω_{aux} using Eq. 5. The classification loss can then be computed based on the rectified pseudo labels $\hat{y}_{img}$ and $\hat{y}_{aux}$, which are more accurate than $\hat{y}$.

4.2.3 Optimization

We perform two rounds of self-training to optimize our proposed two-stream segmentation network. In both stages, we formulate the overall loss as a linear combination of the classification loss and the regularization loss

$$\ell^{stg} = \ell_{cla}^{stg} + \gamma \ell_{reg} \tag{7}$$

where the superscript $stg \in \{1, 2\}$ distinguishes the loss computed in stage 1 or stage 2. γ is a balancing coefficient that controls the weight of the regularization loss. We empirically set $\gamma = 1$ in our experiments. We train the same two-stream segmentation model with the same cross-modal consistency loss as the regularization for self-training. The only difference between the two stages is how we compute the hard pseudo labels and the classification loss.

Stage one The source model extracts the pseudo labels for the target images in the first stage. As the source model was trained on the labeled source data, the uncertainty in the pseudo labels for target images is high. Thus, applying pseudo label denoising techniques is beneficial, based on which a more robust classification loss can be computed. In our implementation, we compute the symmetric cross-entropy (SCE) [54] based on $\hat{y}_{img}$ and $\hat{y}_{aux}$ as

$$\ell_{cla}^{1} = \ell_{sce}(\hat{y}_{img}, p_{img}) + \ell_{sce}(\hat{y}_{aux}, p_{aux}) \tag{8}$$

where p_{img} and p_{aux} are the predicted outputs of the main and auxiliary streams, $\hat{y}_{img}$ and $\hat{y}_{aux}$ are the hard pseudo labels denoised by ω_{img} and ω_{aux}, and $\ell_{sce}(\hat{y}, p) = \alpha \ell_{ce}(p, \hat{y}) + \beta \ell_{ce}(\hat{y}, p)$. Following previous work [67], we set the balancing coefficients α and β to 0.1 and 1.

Stage two The pseudo labels for the target images are extracted by our learned target model in the first stage, which are derived from the fusion of the two streams: $\hat{y} = \frac{1}{2}(p_{img} + p_{aux})$. No advanced denoising methods are required in this stage as the quality of the pseudo labels is already relatively high. We compute the classification loss using Eq. 3 as $\ell_{cla}^{2} = \ell_{ce}(\hat{y}, p_{img}) + \ell_{ce}(\hat{y}, p_{aux})$.

This stage is usually referred to as self-distillation, which has been successfully applied to typical UDA to boost a model's performance [67, 26]. Here we show that with our proposed cross-modal consistency training, one or more rounds of self-distillation can also bring substantial performance gain to source-free UDA.

4.3. Test-time Inference

Considering that the depth information may not always be available during test-time inference, we discard the multimodal auxiliary network and keep only the main stream as our inference-stage model. The reasons behind this are twofold. First, it improves the feasibility of our model as the main stream takes the RGB image as the only input. Second, we observe that the multimodal auxiliary stream only marginally outperforms the main stream after the model converges. Therefore, the accuracy loss as a trade-off for model feasibility is relatively slim. Formally, given a test image x, we compute its pixel-level semantic labels as $p_{img} = g_{img}(f_{img}(x))$.

5. Experiments

5.1. Experimental Settings

Dataset We evaluate our proposed method by adapting from the game scenes GTA [41] and SYNTHIA [42] to the real scenes Cityscapes [12]. The Cityscapes dataset contains 2,975 training and 500 validation images with a resolution of 2048 × 1024. For depth, we use the disparity maps provided by the official Cityscapes dataset by default. In the ablation study, we also evaluate our method with self-supervised stereoscopic depth [44, 53] and monocular depth [55], which were trained on the stereo images and video sequences in the Cityscapes training set, respectively.

Evaluation metric We report the Intersection over Union (IoU) on the 19 common categories shared by GTA5 and Cityscapes and the 16 common categories shared by SYNTHIA and Cityscapes. Following previous studies, we also report the results on 13 of the 16 common categories shared by the SYNTHIA and Cityscapes datasets.

Implementation details For the source-only model, we adopt the pre-trained models on GTA5 and SYNTHIA provided by Kundu *et al*. [26]. Both the source model and our target model use DeepLabv2 [4] for segmentation with ResNet-101 [20] as the backbone. We insert four SSA-Gates, one after each of the four encoder blocks in ResNet-101. We train our model using the SGD solver with a momentum of 0.9 and weight decay of 2×10^{-4}. We use a mini-batch size of 4 and an initial learning rate of 6×10^{-4}. Following [67], we set the parameters for the prototypical pseudo label denoising α, β, and τ to 0.1, 1, and 1, re-

Table 1. Per-class IoU (%) and mIoU (%) comparison of GTA5 → Cityscapes adaptation. The best score for each column is highlighted.

Method	SF	road	sidewalk	building	wall	fence	pole	light	sign	vege.	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
FADA [52]	✗	91.0	50.6	86.0	43.4	29.8	36.8	43.4	25.0	86.8	38.3	87.4	64.0	38.0	85.2	31.6	46.1	6.5	25.4	37.1	50.1
CAG-UDA [68]	✗	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
Seg-Uncertainty [69]	✗	90.4	31.2	85.1	36.9	25.6	37.5	48.8	48.5	85.3	34.8	81.1	64.4	36.8	86.3	34.9	52.2	1.7	29.0	44.6	50.3
IAST [37]	✗	94.1	58.8	85.4	39.7	29.2	25.1	43.1	34.2	84.8	34.6	88.7	62.7	30.3	87.6	42.3	50.3	24.7	35.2	40.2	52.2
CorDA [53]	✗	94.7	63.1	87.6	30.7	40.6	40.2	47.8	51.6	87.6	**47.0**	**89.7**	66.7	35.9	90.2	48.9	57.5	0.0	39.8	56.0	56.6
ProDA [67]	✗	87.8	56.0	79.7	**46.3**	**44.8**	**45.6**	53.5	53.5	**88.6**	45.2	82.1	**70.7**	39.2	88.8	45.5	59.4	1.0	**48.9**	56.4	57.5
EHTDI [29]	✗	**95.4**	**68.8**	**88.1**	37.1	41.4	42.5	45.7	**60.4**	87.3	42.6	86.8	67.4	38.6	**90.5**	**66.7**	61.4	0.3	39.4	56.1	58.8
BiSMAP [35]	✗	89.2	54.9	84.4	44.1	39.3	41.6	**53.9**	53.5	88.4	45.1	82.3	69.4	**41.8**	90.4	56.4	**68.8**	**51.2**	47.8	**60.4**	**61.2**
SFDA [31]	✓	84.2	39.2	82.7	27.5	22.1	25.9	31.1	21.9	82.4	30.5	85.3	58.7	22.1	80.0	33.1	31.5	3.6	27.8	30.6	43.2
URMA [39]	✓	92.3	55.2	81.6	30.8	18.8	37.1	17.7	12.1	84.2	35.9	83.8	57.7	24.1	81.7	27.5	44.3	6.9	24.1	40.4	45.1
LD [66]	✓	91.6	53.2	80.6	36.6	14.2	26.4	31.6	22.7	83.1	42.1	79.3	57.3	26.6	82.1	41.0	50.1	0.3	25.9	19.5	45.5
SRDA [2]	✓	90.5	47.1	82.8	32.8	28.0	29.9	35.9	34.8	83.3	39.7	76.1	57.3	23.6	79.5	30.7	40.2	0.0	26.6	30.9	45.8
SFUDA [64]	✓	95.2	40.6	85.2	30.6	26.1	35.8	34.7	32.8	85.3	41.7	79.5	61.0	28.2	86.5	41.2	45.3	**15.6**	33.1	40.0	49.4
GtA w/o cPAE [26]	✓	90.9	48.6	85.5	35.3	31.7	36.9	34.7	34.8	86.2	47.8	88.5	61.7	32.6	85.9	46.9	50.4	0.0	38.9	52.4	51.6
GtA w/ cPAE [26]	✓	91.7	53.4	86.1	37.6	32.1	37.4	38.2	35.6	86.7	48.5	**89.9**	62.6	**34.3**	87.2	51.0	50.8	4.2	42.7	**53.9**	53.4
Ours	✓	93.0	60.4	87.2	**46.4**	41.4	38.0	45.1	51.5	87.5	**48.6**	83.7	63.2	31.8	88.6	49.5	**60.3**	0.0	**47.1**	47.8	56.4
Ours w/ distillation	✓	**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**
Ours (mono)	✓	95.0	67.0	87.4	44.0	42.2	40.7	47.5	50.8	87.1	51.0	77.5	67.7	29.9	88.5	42.0	57.4	0.0	45.3	42.5	56.0
Ours (stereo)	✓	95.1	67.8	87.7	51.3	41.5	36.3	47.4	51.3	87.8	47.8	87.3	67.0	34.2	87.5	41.0	51.8	0.0	42.6	46.4	56.4

spectively. We conduct an ablation study on the balancing coefficient γ in Eq. 7 and set $\gamma = 1$ in the rest of the experiments. For consistency regularization, we employ random crop as the weak augmentation and apply RandAugment [13] and Cutout [14] in addition to random crop as the strong augmentation. As the class prototypes are required for pseudo label denoising, we first train our target model on the pseudo labels generated by the source model before denoising as a warm-up. Next, we initialize the class prototypes with the learned warm-up model and continue optimizing it based on Eq. 7 for 60 epochs. In the warm-up stage, we choose the top 33% of the most confident predictions per class over the entire training set to select balanced and reliable hard pseudo labels [30, 26].

5.2. Comparisons with State-of-the-Art Methods

We compare our proposed method with the prior art in Tables 1 and 2. The column SF indicates if the comparison method is source-free or not. As shown, our method outperforms the existing source-free methods by a large margin, achieving a state-of-the-art mIoU of 57.7% and 56.4% (57.5% and 55.6%) with or without self-distillation on GTA5 → Cityscapes (SYNTHIA → Cityscapes). We achieve the best score on 15 out of 19 common categories shared by GTA5 and Cityscapes, and on 12 out of 16 common categories shared by SYNTHIA and Cityscapes. The experimental results indicate the effectiveness of our proposed pseudo label denoising with cross-modal consistency training. As we are exploring a new direction that has not been studied in previous source-free methods, our solution is orthogonal to existing techniques such as source domain estimation [31] and conditional Prior-enforcing AutoEncoder (cPAE) [26]. Such techniques can be combined with our proposed method for further performance gains.

Next, we compare our method to the non-source-free prior art. Starting with a well-trained source model (44.0% or 41.0% mIoU on GTA5 or SYNTHIA → Cityscapes), our method obtains competitive or even better results compared to most of the existing non-source-free UDA methods. It is worth noting that our method can be easily integrated with non-source-free UDA methods. A naive implementation is to start with an adapted model instead of the source model to generate pseudo labels for target images in stage one self-training.

5.3. Ablation Study and Discussion

Impact of the source for depth information Our proposed method is agnostic to the acquisition of the depth information. To evaluate, we replace the depth information provided by the official Cityscapes dataset[1] by 1) the self-supervised stereoscopic depth [44] used in CorDA [53], and 2) the self-supervised monocular depth learned by the ManyDepth model [55], denoted as *Ours (stereo)* and *Ours (mono)*, respectively. For the monocular depth, we directly use the 1-channel disparity map as the input; while for the stereo depth, we use the 3-channel HHA representation derived from the depth information with camera parameters as the input (see Figure 3 for the visualized examples). Generally speaking, stereo depth is more accurate but its acquisition requires more expensive stereo cameras. Monocular depth can be estimated based on video sequences recorded by regular cameras. However, it is less accurate and it requires significantly more storage to manage the video sequences. We show that our proposed method is effective with different sources of depth information. In real-world scenarios, users should choose based on their own requirements and available devices.

Utilization strategies on the depth information Existing depth-aware domain adaptive semantic segmentation meth-

[1]The depth provided in the official Cityscapes dataset is not the ground truth but also estimated based on stereo images.

Table 2. Per-class IoU (%) and mIoU (%) comparison of SYNTHIA → Cityscapes adaptation. The best score for each column is highlighted. mIoU and mIoU* denote the averaged scores across 16 and 13 categories, respectively.

Method	SF	road	sidewalk	building	wall*	fence*	pole*	light	sign	vege.	sky	person	rider	car	bus	motor	bike	mIoU	mIoU*
CAG-UDA [68]	✗	84.7	40.8	81.7	7.8	0.0	35.1	13.3	22.7	84.5	77.6	64.2	27.8	80.9	19.7	22.7	48.3	44.5	51.5
FADA [52]	✗	84.5	40.1	83.1	4.8	0.0	34.3	20.1	27.2	84.8	84.0	53.5	22.6	85.4	43.7	26.8	27.8	45.2	52.5
Seg-Uncertainty [69]	✗	87.6	41.9	83.1	14.7	1.7	36.2	31.3	19.9	81.6	80.6	63.0	21.8	86.2	40.7	23.6	53.1	47.9	54.9
IAST [37]	✗	81.9	41.5	83.3	17.7	4.6	32.3	30.9	28.8	83.4	85.0	65.5	30.8	86.5	38.2	33.1	52.7	49.8	57.0
CorDA [53]	✗	**93.3**	61.6	**85.3**	19.6	5.1	37.8	36.6	42.8	84.9	**90.4**	69.7	**41.8**	85.6	38.4	32.6	53.9	55.0	62.8
ProDA [67]	✗	87.8	45.7	84.6	**37.1**	0.6	**44.0**	**54.6**	37.0	88.1	84.4	**74.2**	24.3	**88.2**	51.1	40.5	45.6	55.5	62.0
EHTDI [29]	✗	93.0	**69.8**	84.0	36.6	**9.1**	39.7	42.2	43.8	**88.2**	88.1	68.3	29.0	85.5	**54.1**	37.1	56.3	**57.8**	**64.6**
BiSMAP [35]	✗	81.9	39.8	84.2	-	-	-	41.7	**46.1**	83.4	88.7	69.2	39.3	80.7	51.0	**51.2**	**58.8**	-	62.8
SFDA [31]	✓	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 [39]	✓	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
LD [66]	✓	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
SFUDA [64]	✓	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
GtA w/o cPAE [26]	✓	89.0	44.6	80.1	7.8	0.7	34.4	22.0	22.9	82.0	86.5	65.4	33.2	84.8	45.8	38.4	31.7	48.1	55.5
GtA w/ cPAE [26]	✓	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
Ours	✓	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	62.1
Ours w/ distillation	✓	**91.5**	**56.3**	**85.9**	**37.9**	**9.2**	**42.1**	**42.6**	**47.6**	**87.2**	86.1	64.5	23.3	**89.3**	**64.5**	45.0	**47.7**	**57.5**	**64.0**
Ours (mono)	✓	91.2	56.6	85.0	36.5	6.8	41.6	45.5	18.8	86.5	86.2	66.4	26.7	88.7	58.2	44.3	48.0	55.4	61.7
Ours (stereo)	✓	91.6	56.4	85.7	29.3	7.8	41.2	42.0	37.6	86.8	85.9	65.2	27.3	88.4	59.5	44.4	47.8	56.0	63.0

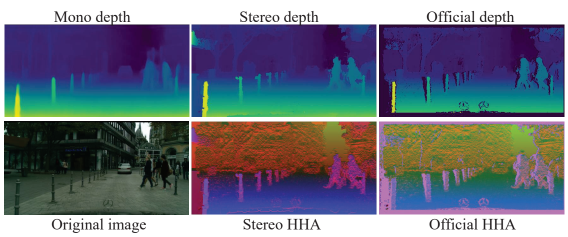

Figure 3. Visualization of the depth and the HHA representation obtained by different methods.

Table 3. Comparison of different utilization strategies of the depth information for source-free UDA on GTA5 → Cityscapes. * indicates we made minimum modifications to make the method compatible with source-free settings.

Method	BG	MC	RIV	RIG	DS	mIoU	gain
Source only [26]	55.3	19.4	28.7	62.9	53.7	44.0	-
DADA* [51]	61.5	26.9	36.1	72.1	55.8	50.1	+6.1
CorDA* [53]	60.5	27.3	39.0	73.8	55.6	50.5	+6.5
MKE* [61]	62.2	27.8	40.4	70.9	57.8	51.5	+7.5
Ours	65.8	31.7	44.9	76.7	65.4	56.4	+12.4

ods mostly follow a multitask learning framework where depth estimation is modeled as the auxiliary task [51, 53]. We modified two depth-aware UDA methods to make them applicable in a source-free setting by calculating the classification loss based on the pseudo labeled target images only. The results are reported in Table 3[2]. As shown, without the supervision of the labeled source data, the regularization induced by the auxiliary task is quite limited. Moreover, we compare our approach to a Multimodal Knowledge Expansion (MKE) method [61] that transfers knowledge from a unimodal teacher network to a multimodal student network.

[2]Background (BG) - building, wall, fence, vegetation, terrain, sky; Minority Class (MC) - rider, train, motorcycle, bicycle; Road Infrastructure Vertical (RIV) - pole, traffic light, traffic sign; Road Infrastructure Ground (RIG) - road, sidewalk; and Dynamic Stuff (DS) - person, car, truck, bus.

Table 4. Model justification of our proposed framework on GTA5 → Cityscapes. The auxiliary modality column indicates if depth modality is used during training or not.

	components				mIoU	gain
	source model				44.0	-
stage 1	auxiliary modality	self training	consistency regularization	pseudo label denoising	mIoU	gain
		✓			50.5	+6.5
		✓	✓		51.2	+7.2
		✓		✓	52.7	+8.7
		✓	✓	✓	55.1	+11.1
	✓	✓			50.9	+6.9
	✓	✓	✓		51.6	+7.6
	✓	✓		✓	54.2	+10.2
	✓	✓	✓	✓	56.4	+12.4
stage 2	auxiliary modality	self distillation	stage 1 initialization	self-supervised initialization	mIoU	gain
	✓	✓	✓		57.6	+13.6
	✓	✓		✓	57.7	+13.7

However, as this method did not address the domain shift issue between the source model and the target images, it performs less effectively than our proposed approach. Furthermore, the inference-stage model in MKE is multimodal, while ours is unimodal with better feasibility.

Effectiveness of cross-modal pseudo label denoising Our proposed framework consists of two major components, namely the multimodal auxiliary network and cross-modal consistency training. As shown in Table 4, we start with a source model that obtains an mIoU of 44.0% on the GTA5 → Cityscapes. By training the network without our proposed consistency regularization, it achieves an mIoU of 50.9% and 54.2%, respectively, based on the supervision of the classification loss only before and after the pseudo label denoising. By combining our proposed consistency regularization with pseudo label denoising, we obtain a new state-of-the-art mIoU of 56.4%, outperforming the source model significantly by 12.4%. To evaluate the benefits introduced by the depth modality, we replaced our multimodal auxiliary network with a unimodal network with the same ar-

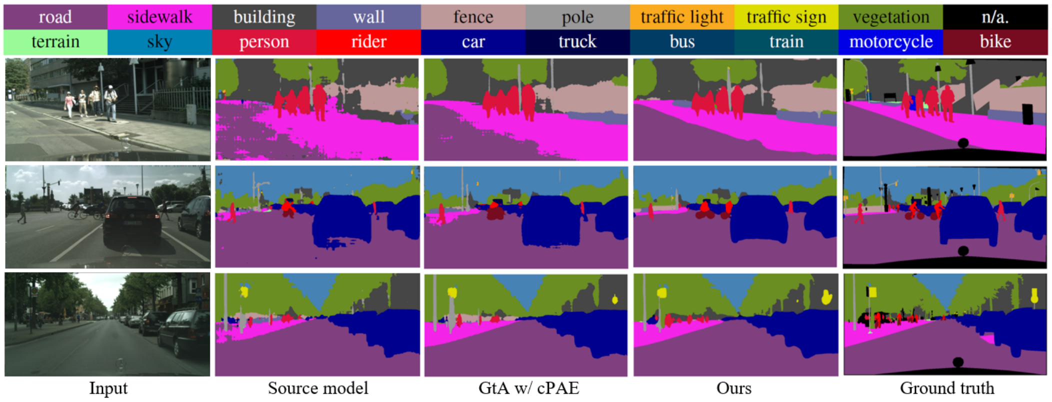

Figure 4. Qualitative results of source-free semantic segmentation on the Cityscapes dataset. From left to right: input, output of the source model, output of the GtA model with cPAE [26], output of our proposed model without self-distillation, ground-truth segmentation mask.

Table 5. Impact of the source model on GTA5 $\rightarrow$ Cityscapes.

source model	source training	target model	target adaptation	mIoU
DeepLabv2	data aug.	-	-	38.6
DeepLabv2 [26]	multi-head	-	-	44.0
DeepLabv2	multi-head	SegFormer	self-training	51.3
DeepLabv2	multi-head	DeepLabv2	self-training	50.5
DeepLabv2	multi-head	DeepLabv2*	our proposed	56.4
SegFormer [59]	data aug.	-	-	43.2
SegFormer	data aug.	SegFormer	self-training	50.5
SegFormer	data aug.	DeepLabv2	self-training	49.4
SegFormer	data aug.	DeepLabv2*	our proposed	55.5
GtA w/ cPAE	SF adapted	-	-	53.4
GtA w/ cPAE	SF adapted	DeepLabv2*	our proposed	57.3
ProDA [67]	non-SF adapted	-	-	57.5
ProDA	non-SF adapted	DeepLabv2*	our proposed	59.5

chitecture as the main stream. The mIoU decreases in all cases by using RGB as the only input. Next, we evaluate our cross-modal consistency training in self-distillation. We initialize our model either with the weights of the learned model in stage one (*i.e.*, stage 1 initialization) or with SimCLRv2 [7] pretrained weights (*i.e.*, self-supervised initialization). In both cases, we observe a performance gain of around 1.3% over the stage one model. The qualitative evaluation of our method is illustrated in Figure 4.

Impact of the source model The majority of the source-free UDA methods are built upon DeepLab models. Here we evaluate a Transformer-based model, namely Segformer [59], as the source and target models in a source-free UDA setting. As Table 5 shows, Segformer has better generalization ability than DeepLabv2. With data augmentation only, a source Segformer model obtains an mIoU of 43.2, outperforming a source DeepLabv2 model by 4.6%. Moreover, when being adopted as the target model, Segformer achieves an mIoU of 51.3% and 50.5%, respectively. It outperforms the corresponding DeepLabv2 by 0.8% and 1.1%, when being adapted from the same source model. To verify that our method is orthogonal to previous work, we also start with a source-free model (*i.e.*, GtA w/ cPAE [26]) and a non-source-free model (*i.e.*, ProDA [67]), and apply our method on top of it. As can be seen, the mIoU has been further improved by 3.9% and 2%, respectively.

Table 6. The effect of the balancing coefficient γ.

γ	0.5	1	2	5	10
mIoU	56.0	56.4	56.2	56.7	55.7

Table 7. The mIoU obtained by the multimodal auxiliary network with varying number of SSA-Gate.

SSA-Gate no.	1	2	3	4
mIoU	43.4	49.2	53.1	56.6

Parameter sensitivity analysis Finally, we study the impact of the balancing coefficient γ in Eq. 7 on the self-training in stage one. We set γ to different values, conduct experiments on GTA5 $\rightarrow$ Cityscapes, and report the results in Table 6. The experimental results show that our proposed method is not sensitive to the balancing factor γ. In our previous experiments, we empirically set $\gamma = 1$. It shows that the mIoU can be slightly improved by setting $\gamma = 5$. We obtain the state-of-the-art mIoU of $55.7\% \sim 56.7\%$ when $\gamma \in [0.5, 10]$, which verifies and underscores the robustness of our proposed cross-modal consistency training technique. Table 7 shows the mIoU obtained by the multimodal auxiliary network with varying number of SSA-Gate. The mIoU decreases significantly to 43.4% with only one SSA-Gate, which indicates that predicting the semantic labels from depth alone is challenging without sufficient information exchange with RGB images.

6. Conclusions

We propose to enhance source-free domain adaptive semantic segmentation via cross-modal consistency training. To achieve this goal, we introduce a multimodal auxiliary network to leverage the guidance from the depth modality during training. A cross-modal consistency loss is formulated between the output of the main and the auxiliary networks, which serves as an effective regularization for source-free UDA. Our proposed approach not only outperforms the source-free prior art by a large margin, but also reduces the gap between source-free and non-source-free UDA methods in semantic segmentation.

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, pages 15384–15394, 2021. 2

[2] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-relaxed domain adaptation for image segmentation. In *MICCAI*, pages 490–499, 2020. 6

[3] Adriano Cardace, Luca De Luigi, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Plugging self-supervised monocular depth into unsupervised domain adaptation for semantic segmentation. In *WACV*, pages 1129–1139, 2022. 2

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 3, 5

[5] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. Spatial information guided convolution for real-time RGBD semantic segmentation. *IEEE Transactions on Image Processing*, 30:2313–2324, 2021. 4

[6] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*, pages 2090–2099, 2019. 2

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *NeurIPS*, 33:22243–22255, 2020. 8

[8] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *ECCV*, pages 561–577, 2020. 3, 4

[9] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, pages 1841–1850, 2019. 2

[10] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, pages 7892–7901, 2018. 2

[11] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, pages 1992–2001, 2017. 1, 2

[12] 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 *CVPR*, pages 3213–3223, 2016. 5

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, pages 702–703, 2020. 6

[14] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 6

[15] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation. In *ICCV*, pages 982–991, 2019. 2

[16] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, pages 740–756. Springer, 2016. 2

[17] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017. 2

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[19] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV*, pages 345–360, 2014. 4

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[21] 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 *ICML*, pages 1989–1998, 2018. 2

[22] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Koring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *CVPR*, pages 11130–11140, 2021. 2

[23] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, pages 9924–9935, 2022. 2

[24] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*, 2022. 2

[25] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xMUDA: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *CVPR*, pages 12605–12614, 2020. 2

[26] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *ICCV*, pages 7046–7056, 2021. 1, 2, 3, 5, 6, 7, 8

[27] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017. 3

[28] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *ECCV*, pages 440–456, 2020. 2

[29] Junjie Li, Zilei Wang, Yuan Gao, and Xiaoming Hu. Exploring high-quality target domain information for unsupervised

domain adaptive semantic segmentation. In *ACM Multimedia*, pages 5237—5245, 2022. 6, 7

[30] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, pages 6936–6945, 2019. 2, 6

[31] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, pages 1215–1224, 2021. 1, 2, 3, 6, 7

[32] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *CVPR*, pages 525–534, 2021. 1

[33] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. Towards natural and accurate future motion prediction of humans and animals. In *CVPR*, pages 10004–10012, 2019. 1

[34] Adrian Lopez-Rodriguez and Krystian Mikolajczyk. Desc: Domain adaptation for depth estimation via semantic consistency. *International Journal of Computer Vision*, 131(3):752–771, 2023. 2

[35] Yulei Lu, Yawei Luo, Li Zhang, Zheyang Li, Yi Yang, and Jun Xiao. Bidirectional self-training with multiple anisotropic prototypes for domain adaptive semantic segmentation. In *ACM Multimedia*, pages 1405—-1415, 2022. 6, 7

[36] 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 *ICCV*, pages 2507–2516, 2019. 2

[37] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, 2020. 2, 6, 7

[38] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *CVPR*, pages 12435–12445, 2021. 2, 4

[39] S Prabhu Teja and François Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *CVPR*, pages 9613–9623, 2021. 1, 2, 3, 6, 7

[40] Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. In *IJCAI*, 2021. 2

[41] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, pages 102–118, 2016. 5

[42] 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 *CVPR*, pages 3234–3243, 2016. 5

[43] Suman Saha, Anton Obukhov, Danda Pani Paudel, Menelaos Kanakis, Yuhua Chen, Stamatios Georgoulis, and Luc Van Gool. Learning to relate depth and semantics for unsupervised domain adaptation. In *CVPR*, pages 8197–8207, 2021. 2

[44] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *ECCV*, pages 687–704, 2018. 1, 5, 6

[45] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *ECCV*, pages 532–548, 2020. 2

[46] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 33:596–608, 2020. 4

[47] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017. 4

[48] 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. 1, 2

[49] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *PAMI*, 2021. 2

[50] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, pages 2517–2526, 2019. 1, 2

[51] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, pages 7364–7373, 2019. 2, 3, 7

[52] 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 *ECCV*, pages 642–659, 2020. 6, 7

[53] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *ICCV*, pages 8515–8525, 2021. 2, 3, 5, 6, 7

[54] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, pages 322–330, 2019. 5

[55] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *CVPR*, 2021. 5, 6

[56] Quanliang Wu and Huajun Liu. Unsupervised domain adaptation for semantic segmentation using depth distribution. In *Advances in Neural Information Processing Systems*. 3

[57] Zuxuan Wu, Xin Wang, Joseph E Gonzalez, Tom Goldstein, and Larry S Davis. ACE: Adapting to changing environments for semantic segmentation. In *ICCV*, pages 2121–2130, 2019. 2

[58] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. SePiCo: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *arXiv preprint arXiv:2204.08808*, 2022. 2

[59] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34:12077–12090, 2021. 8

[60] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 33:6256–6268, 2020. 4

[61] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In *ICCV*, pages 854–863, 2021. 2, 7

[62] 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. 1, 2

[63] Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *CVPR*, pages 4085–4095, 2020. 2

[64] Mucong Ye, Jing Zhang, Jinpeng Ouyang, and Ding Yuan. Source data-free unsupervised domain adaptation for semantic segmentation. In *ACM Multimedia*, pages 2233—-2242, 2021. 2, 6, 7

[65] Yifang Yin, Harsh Shrivastava, Ying Zhang, Zhenguang Liu, Rajiv Ratn Shah, and Roger Zimmermann. Enhanced audio tagging via multi-to single-modal teacher-student mutual learning. In *AAAI*, volume 35, pages 10709–10717, 2021. 2

[66] Fuming You, Jingjing Li, Lei Zhu, Zhi Chen, and Zi Huang. Domain adaptive semantic segmentation without source data. In *ACM Multimedia*, pages 3293—-3302, 2021. 2, 6, 7

[67] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*, pages 12414–12424, 2021. 1, 2, 4, 5, 6, 7, 8

[68] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *NeurIPS*, 32, 2019. 2, 6, 7

[69] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, pages 1106–1120, 2021. 1, 2, 4, 6, 7

[70] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Jiangmiao Pang, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Context-aware mixup for domain adaptive semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 2

[71] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017. 2, 3

[72] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, pages 289–305, 2018. 2

[73] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, pages 5982–5991, 2019. 2