Unsupervised Domain Adaptation for Semantic Segmentation with Pseudo Label Self-Refinement

Xingchen Zhao[2*], Niluthpol Chowdhury Mithun[1*], Abhinav Rajvanshi[1], Han-Pang Chiu[1], Supun Samarasekera[1]

[1]SRI International, Princeton, NJ, USA [2]Northeastern University, Boston, MA, USA
[1]firstname.lastname@sri.com [2]zhao.xingc@northeastern.edu

Abstract

Deep learning-based solutions for semantic segmentation suffer from significant performance degradation when tested on data with different characteristics than what was used during the training. Adapting the models using annotated data from the new domain is not always practical. Unsupervised Domain Adaptation (UDA) approaches are crucial in deploying these models in the actual operating conditions. Recent state-of-the-art (SOTA) UDA methods employ a teacher-student self-training approach, where a teacher model is used to generate pseudo-labels for the new data which in turn guide the training process of the student model. Though this approach has seen a lot of success, it suffers from the issue of noisy pseudo-labels being propagated in the training process. To address this issue, we propose an auxiliary pseudo-label refinement network (PRN) for online refining of the pseudo labels and also localizing the pixels whose predicted labels are likely to be noisy. Being able to improve the quality of pseudo labels and select highly reliable ones, PRN helps self-training of segmentation models to be robust against pseudo label noise propagation during different stages of adaptation. We evaluate our approach on benchmark datasets with three different domain shifts, and our approach consistently performs significantly better than the previous state-of-the-art methods.

1. Introduction

Semantic segmentation, a well-studied computer vision task, has seen significant advances with deep neural networks in the last decade [5, 14, 36, 39, 50, 58]. In practice, these models rely strongly on large-scale annotated datasets for training. However, the characteristics of the datasets used for training could be significantly different from those in the actual operational scenarios, e.g., changes in camera sensors, and lighting conditions. When there is a distribution shift between train (i.e., source) and test (i.e., target) sets, model accuracy often degrades dramatically [8, 9, 29, 44]. Creating new annotated datasets for retraining is costly, especially for per-pixel annotation.

To alleviate this issue, various UDA approaches have been developed over recent years, which focus on adapting models trained from a source domain to target domains with unlabeled data [9, 15, 19, 56]. For example, many benchmark manually annotated outdoor driving semantic segmentation datasets (e.g., Cityscapes [7] daytime driving dataset, SYNTHIA [33] synthetic driving dataset) are available to train high-performing neural network models for the source domain, but only unlabeled data is available for the target domain (e.g., nighttime driving dataset Dark Zurich [34]). State-of-the-art UDA semantic segmentation techniques often use a teacher-student self-training approach, iteratively training a student model with pseudo-labeled target data generated by a teacher model [4,15,40]. While self-training approaches have demonstrated their effectiveness, they suffer significantly from the erroneous model prediction propagation issue, i.e., confirmation bias [38,53]. Pseudo-labels are very likely to be noisy especially during the early stages of training due to the source-target domain gap. If the issue is not addressed, it consequently leads to corrupted models with degraded generalization performance.

We propose to train an auxiliary neural network model for refining the pseudo labels. Our proposed pseudo-label refinement network (PRN) is trained to serve two main purposes: it refines noisy pseudo labels, improving their quality, and localizes potential errors in pseudo labels by predicting a binary mask for challenging pixels (that are likely to have incorrectly predicted labels). The first task focuses on correcting pseudo-labels, while the second task helps in pseudo-label selection. Note that it is possible to just carry out the first task and to select pixels with maximum softmax probability (of corrected segmentation logits) below a

[*]Equal Contribution

selected threshold as the erroneous pseudo labels. However, the performance of such a naive approach would be very sensitive to the selected threshold. Moreover, it is evident that the first task is class-specific, whereas the second is class-agnostic. Therefore, we train our model specifically for the task of pseudo-label error mask prediction, which also helps to learn effective representations to correct erroneous pseudo labels. PRN minimizes confirmation bias, making self-training for semantic segmentation models more robust against noisy pseudo-labels.

Our PRN model takes the noisy segmentation logits from the teacher decoder and image features from the teacher encoder as the input and predicts the refined logits and the noise mask. This ability of our model is achieved by employing a novel training strategy using Fast Fourier Transform (FFT) based perturbations. Fourier transformation can be applied to decompose any signal (e.g., RGB image, features, logits) to amplitude (i.e., intensities or style) and phase (i.e., spatial positions or semantics) components [31]. We perturb the amplitude of source image segmentation logits using the amplitude of a random target image to effectively introduce noise while preserving object structure, facilitating effective learning of the PRN model. It allows the use of ground truth (GT) labels for the source data as supervision for the model, while style information from the target domain also acts as training inputs through perturbed logits. We also train our model using target domain data (perturbed with source style) with a similar process. Since access to target GT labels is not available, pseudo-labels of unperturbed target logits are used as supervision. This training strategy helps the model learn robust features across domains to effectively refine target pseudo-labels.

Contributions: Our work aims to tackle the propagation of noisy pseudo-labels in the training process. It has two main contributions. The first is a new pseudo-label refinement module that learns to predict the refined pseudo labels as well as the error mask containing the information about noisy labels. Our approach is different from previous approaches that rely on threshold-based selective pseudo-labeling. We also developed a novel training strategy using FFT-based perturbations that enables us to achieve the desired behavior of the refinement module. As the second contribution, our framework outperforms SOTA methods significantly in three UDA segmentation benchmarks, covering normal-to-adverse weather and synthetic-to-real adaptation.

2. Related Works

2.1. Unsupervised Domain Adaptation

A number of unsupervised domain adaptation techniques [8, 10] have been developed for reducing the domain gap between the source and target data, specifically for the semantic segmentation task. For example, Distribution Discrepancy Minimization [3] seeks to minimize the distribution discrepancy between source and target domains in some latent feature space. Curriculum learning [25, 34, 57] has also been used for domain adaptation which involves learning easier tasks before more complex tasks. Self-ensembling [6,30,43] uses an ensemble of models, and exploits the consistency between predictions under some perturbations. Adversarial training [18, 52, 54] is another popular approach that achieves the same goal by training with both clean and adversarial samples. Recently, self-training has been the most popular method for UDA [4, 15, 16, 21, 23, 24, 40, 48, 61], in which the pseudo-labels are generated for the target-data (typically by a teacher model) and then used to train the target domain model. This approach has shown SOTA performance for the UDA semantic segmentation task. However, it generally suffers from the presence of significant noise in the pseudo-labels.

2.2. Pseudo Label Refinement

The potential existence of noisy pseudo-labels in the self-training method is likely to result in subpar performance [2]. Hence, the key concept for these methods revolves around producing dependable pseudo-labels. Some works in semi-supervised learning address this by employing a neural network module to rectify pseudo-labels or identify errors [20,22,27]. In their context, both labeled and unlabeled data stem from the same domain, allowing the refinement module to be trained using labeled data. However, it is not applicable to UDA, where labeled and unlabeled data pertain to distinct domains. Among prior UDA semantic segmentation works, most employ selective pseudo-labeling (e.g., [37, 47]). Some of the works [23, 48, 61] rely on the softmax of the model output as a confidence measure. [26] uses an adaptive confidence threshold that is updated throughout the training, while [59] explicitly estimates the prediction uncertainty during training for filtering. CBST [61] employs category confidence for generating balanced pseudo labels. MetaCor [13] models the noise distribution of the pseudo-labels to enhance the generalization ability of the model on the target domain. ProDA [56] uses online-estimated class-wise feature centroids to rectify labels, by aligning soft prototypical assignments for different views of the same target.

We present a novel FFT-based strategy for training an auxiliary PRN module to effectively refine pseudo-labels. By training the refinement module using perturbations in source and target logits, tied to style and semantics, our approach facilitates the adept refinement of noisy target predictions. This continuous learning process enhances the student model's representation using more precise labels from the target domain. In contrast, prior works lack an explicit noise-handling strategy to address domain gap, leading to potentially subpar models with low-quality pseudo-labels.

3. Methodology

We first discuss the baseline self-training UDA method. Then, we discuss UDA with our proposed pseudo-label refinement neural network model and provide the details of our FFT-based perturbation approach to train the model. Next, we discuss two additional components (i.e., contrastive learning, and Fourier-based style adaptation) of our framework which help to further improve the quality of our models. Finally, we discuss the overall loss function.

Problem Setting: Let, $\mathsf{D_S} = \{(x^i_\mathsf{S}, y^i_\mathsf{S})\}_{i=1}^{N_s}$ be the source domain dataset with N_s labeled samples where y^i_S denotes the one-hot ground-truth (GT) per pixel label for image x^i_S. Here, $x^i_\mathsf{S} \in \mathbb{R}^{H\times W}$ and $y^i_\mathsf{S} \in \{0,1\}^{H\times W\times K}$. (H,W) is the image resolution and K is the number of classes. We also have a target domain dataset $\mathsf{D_T} = \{(x^i_\mathsf{T}\}_{i=1}^{N_t}$ containing N_T images without ground-truth labels. $\mathsf{D_S}$ and $\mathsf{D_T}$ share a common set of K classes. In UDA semantic segmentation, the goal is to train a segmentation model $\mathcal{F}_\theta$ for the target domain by utilizing the labeled set $\mathsf{D_S}$ from source domain and the unlabeled set $\mathsf{D_T}$ from the target domain.

3.1. Self-Training (ST) for UDA

We can train a neural network model $\mathcal{F}_\theta$ with the available source domain images and labels employing supervised learning with a cross-entropy loss $\mathcal{L}^S_{ce}$ on source domain images. $\mathcal{L}^S_{ce}$ for i[th] sample can be written as,

$$\mathcal{L}^{S(i)}_{ce} = -\sum_{j=1}^{H\times W}\sum_{k=1}^{K} y^{(i,j,k)}_\mathsf{S} \log \mathcal{F}_\theta(x^i_\mathsf{S})^{(j,k)} \tag{1}$$

However, due to the domain gap, the model trained with only source domain images with loss $\mathcal{L}^S_{ce}$ is unlikely to generalize well to the target domain. To address the domain gap, we adopt the self-training-based UDA technique as our baseline [4, 15, 16, 40, 60]. In self-training, a teacher model $\mathcal{F}_\phi$ is used for generating the pseudo-labels $\hat{y}^i_\mathsf{T}$ for target domain images. Pseudo-labeled target data is used along with labeled source data for training the student model $\mathcal{F}_\theta$ iteratively, to adapt the model to the target domain. In general, the semantic segmentation models $\mathcal{F}_\theta$ ($\mathcal{E}_\theta$, $\mathcal{D}_\theta$) and $\mathcal{F}_\phi$ ($\mathcal{E}_\phi$, $\mathcal{D}_\phi$) consist of a feature extractor (i.e., encoder $\mathcal{E}$), followed by a classifier predicting pixel-wise labels (i.e., decoder $\mathcal{D}$).

The teacher model is updated using the exponential moving average (EMA) of weights of the student model after each training step, which helps the teacher produce stable predictions. The teacher weights ϕ_{n+1} at train step $n+1$ is,

$$\phi_{n+1} = \beta\phi_n + (1-\beta)\theta_n, \tag{2}$$

Here, β is a hyper-parameter to adjust the degree of change in the model weights. The pseudo labels $\hat{y}^i_\mathsf{T}$ for target domain images are generated using the teacher model $\mathcal{F}_\phi$.

$$\hat{y}^i_\mathsf{T} = \arg\max_k \mathcal{F}_\phi(x^i_\mathsf{T})^{(j,k)} \tag{3}$$

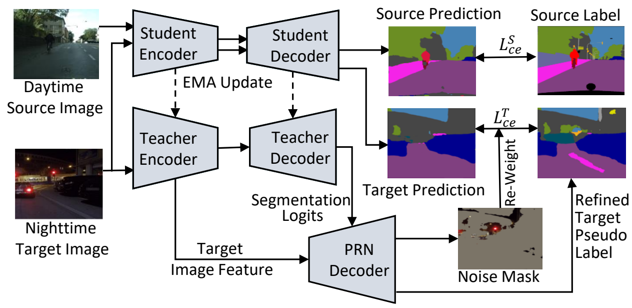

Figure 1. Overview of the UDA self-training framework with the proposed pseudo label refinement network (PRN). We consider the PRN decoder to be fixed (i.e., stop gradient flow to PRN) when calculating losses for training the student network.

These pseudo labels are also used to calculate an additional cross-entropy loss $\mathcal{L}^T_{ce}$ to train the student model $\mathcal{F}_\theta$ to adapt to the target domain. To minimize the effect of label noise, the $\mathcal{L}^{T(i)}_{ce}$ loss for target domain samples is weighted with quality estimates of the pseudo labels [15, 16, 40].

$$\mathcal{L}^{T(i)}_{ce} = -\sum_{j=1}^{H\times W}\sum_{k=1}^{K} \eta^i_T \hat{y}^{(i,j,k)}_\mathsf{T} \log \mathcal{F}_\theta(x^i_\mathsf{T})^{(j,k)} \tag{4}$$

Here, η^i_T $(0 \leq \eta^i_T \leq 1)$ is a confidence estimate of pseudo-label $\hat{y}^i_\mathsf{T}$. The labels are not always correct for the target domain samples, and the modified $\mathcal{L}^{T(i)}_{ce}$ loss takes that into consideration. On the other hand, we are fully confident about the source domain labels (i.e., η^i_s=1) as they are ground-truth and hence, $\mathcal{L}^{S(i)}_{ce}$ does not require any modification. Following prior works [15, 16], η^i_T can be calculated as the percentage of pixels in the image with maximum softmax probability exceeding a threshold τ_1.

$$\eta^i_T = \frac{\sum_{j=1}^{H\times W} \mathbb{1}[\max_k \mathcal{F}_\phi(x^i_\mathsf{T})^{(j,k)} > \tau_1]}{H\times W} \tag{5}$$

We also use augmented target data in training, which has been shown to be effective in prior works [1, 15, 40]. Data augmentation helps to learn more domain-robust features and thus improves generalization performance to unseen data. We use color jitter, gaussian blur, and ClassMix [28] as data augmentations following prior works [15, 40]. Augmented target samples are used in training the student model, while the non-augmented target samples are used by the teacher model to generate the pseudo-labels.

Despite adopting the strategies discussed above, ST approaches often suffer from the risk of training models that generalize poorly due to error propagation from the memorization of noisy pseudo labels (i.e., confirmation bias towards errors). To alleviate this issue, we propose to train an auxiliary pseudo-label refinement network (PRN) model f_σ that focuses on label refinement and label noise localization.

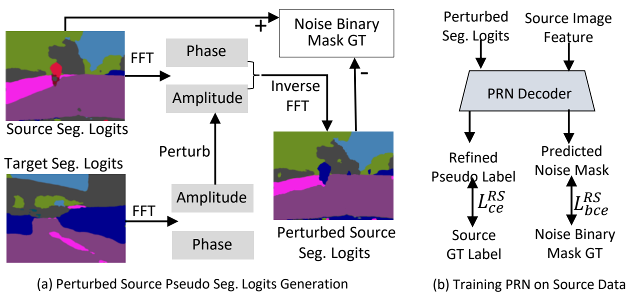

Figure 2. Perturbed label generation and training of PRN model on source data. A similar process is followed for target data as described in Sec. 3.3. In (a), we show a segmentation map instead of segmentation logits only for visualization purposes. We consider the student network to be fixed when training the PRN decoder.

3.2. UDA with Proposed Pseudo Label Refinement

An overview of the UDA self-training framework with the proposed pseudo label refinement network (PRN) is shown in Fig. 1. PRN network is a decoder model $\mathcal{D}_\sigma$ that takes in target image features from the teacher encoder $\mathcal{E}_\phi(x_{\mathsf{T}}^i)$ and segmentation logits from the teacher decoder $\mathcal{F}_\phi(x_{\mathsf{T}}^i)$. It outputs refined labels $\bar{y}_{\mathsf{T}}^i$ and noise masks μ_{T}^i. Here, $\mu_{\mathsf{T}}^{(i,j)}$ is 1 if the pseudo-label of pixel j is predicted to be noisy and 0 otherwise. Different from Eq. 5, we use noise mask μ_{T}^i to calculate the confidence estimate $\bar{\eta}_T^i$.

$$\bar{\eta}_T^i = \frac{\sum_{j=1}^{H\times W} \mathbb{1}[\mu_{\mathsf{T}}^{(i,j)} = 0]}{H \times W} \tag{6}$$

The use of noise mask instead of segmentation logits, allows us to avoid selecting a threshold (i.e., τ_1 in Eq. 5) for the calculation of quality estimate. The target cross-entropy loss $\mathcal{L}_{ce}^{T(i)}$ is modified by using $\bar{y}_{\mathsf{T}}^i$ and $\bar{\eta}_T^i$. The source cross-entropy loss $\mathcal{L}_{ce}^{S(i)}$ remains the same as the source labels are GT and do not require any refinement. Based on the predicted target noise mask μ_{T}^i, $\mathcal{L}_{ce}^{T(i)}$ calculation can avoid the difficult pixels for which the pseudo-label is predicted to be noisy. We use noise masks in creating the augmented target samples used for training the student model. For predicted difficult pixels in a target image, a randomly selected source image from the batch and corresponding GT labels are used in place of the target image pixels and pseudo labels. Next, we discuss how we train the refinement network in Sec. 3.3.

3.3. Training the PRN Network

The pseudo-label refinement network takes in source/target image features as well as perturbed segmentation logits and focuses on learning to predict higher-quality segmentation labels and noise max predicting pixels for which pseudo labels are likely to be erroneous. The model can be trained only using the labeled source data via supervised learning. However, such a model is unlikely to learn to effectively refine the target image pseudo labels. We train the PRN model using both source and target domain data using a novel FFT-based perturbation strategy. Prior work [55] showed FFT-based transfer to be effective as a pre-processing step for UDA techniques. However, we consider FFT-based perturbation as an integral part of the UDA self-training process.

First, we discuss how we perturb segmentation logits and generate pseudo noise masks for calculating losses on the source domain for training PRN. Fig. 2 provides a brief illustration of the generation of perturbed source segmentation logits and PRN on source data. For source data, we have the ground-truth (GT) source label available and we use GT labels and predicted logits from the student network to calculate the cross-entropy loss. We perturb the source image segmentation logits using segmentation logits (from the teacher network) of a randomly sampled target domain image in batch. Based on the Fast Fourier Transform, the low-level frequencies of the amplitude of source segmentation logits are replaced by that of the target domain image. The perturbed segmentation logits are reconstituted using modified amplitude and unaltered phase via the inverse FFT (iFFT). Let's assume, l_{S}^i and l_{T}^i are segmentation logits from sampled source and target images. $\mathcal{T}^A$ and $\mathcal{T}^P$ denote the amplitude and phase components of the Fourier transform $\mathcal{T}$ of segmentation logits l_{S}^i. The perturbed logits $\tilde{l}_{\mathsf{S}}^i$ calculation can be formalized as,

$$\tilde{l}_{\mathsf{S}}^i = \mathcal{T}^{-1}([\mathsf{M}_\epsilon \cdot \mathcal{T}^A(l_{\mathsf{T}}^i) + (1-\mathsf{M}_\epsilon)\cdot \mathcal{T}^A(l_{\mathsf{S}}^i), \mathcal{T}^P(l_{\mathsf{S}}^i)]) \tag{7}$$

$\mathcal{T}^{-1}$ denotes iFFT. M_ϵ is a mask which is calculated as,

$$\mathsf{M}_\epsilon(h,w) = \mathbb{1}_{(h,w)\in[-\epsilon H:\epsilon H, -\epsilon W:\epsilon W]} \tag{8}$$

To perturb the low-frequency component, the value of M_ϵ is set to 1 only for the center region and 0 otherwise. ϵ is a hyper-parameter ($\epsilon \in (0,1)$) that controls perturbation strength. In our experiment, we randomly select the ϵ value between 0.05 and 0.2. As the low-level spectrum (amplitude) encodes style characteristics and the phase encodes high-level semantics, the source logits are perturbed significantly without affecting the high-level semantics. By refining source segmentation logits perturbed in this process, we hypothesize our PRN model will learn to refine some characteristics of target pseudo-label noise.

After applying the FFT-based perturbation discussed above, the image's semantic content remains the same and we can use the available GT source label for training. The binary mask GT μ_{S}^i can be created by comparing the original and perturbed logits (transformed to labels using $\arg\max$): Set to 1 for identical labels and 0 for others. PRN uses this perturbed source segmentation logits and source student encoder feature to generate refined segmentation logits $\bar{l}_{\mathsf{S}}^i$ and noise masks $\bar{\nu}_{\mathsf{S}}^i$. Let, $(\bar{l}_{\mathsf{S}}^i, \bar{\nu}_{\mathsf{S}}^i)$ =

$\mathcal{D}_\sigma(\mathcal{E}_\theta(x^i_{\mathsf{S}}), \tilde{l}^i_{\mathsf{S}})$. To refine source labels, PRN is trained with two loss components (i.e., (1) cross-entropy loss $\mathcal{L}^{RS}_{ce}$ between refined segmentation label $\bar{l}^i_{\mathsf{S}}$ with GT source label y^i_{S} and (2) binary cross-entropy loss $\mathcal{L}^{RS}_{bce}$ between predicted noise mask $\bar{\nu}^i_{\mathsf{S}}$ and GT binary noise mask μ^i_{S}).

$$\mathcal{L}^{RS(i)}_{ce} = -\sum_{j=1}^{H\times W}\sum_{k=1}^{K} y^{(i,j,k)}_{\mathsf{S}} \log(\bar{l}^i_{\mathsf{S}})^{(j,k)} \tag{9}$$

$$\begin{aligned}\mathcal{L}^{RS(i)}_{bce} = -\sum_{j=1}^{H\times W} \Big(\mu^{(i,j)}_{\mathsf{s}} \log(\bar{\nu}^i_{\mathsf{S}})^{(j)} \\ +(1-\mu^{(i,j)}_{\mathsf{s}}) \log(1-\bar{\nu}^i_{\mathsf{S}})^{(j)}\Big)\end{aligned} \tag{10}$$

Next, we discuss the losses for training PRN on target domain data. For target data, we do not have access to ground-truth labels. For reference, we use pseudo-label generated from the refinement decoder with unperturbed target logits from the teacher model. Since the target pseudo label at the early stage of training can be noisy, we set a high threshold τ_2 on selecting the pseudo labels. After the learning rate warm-up period is over, we assume the model to have some ability to localize the noise in the pseudo labels. Then, we avoid the noisy parts of the target pseudo-label based on the predicted error mask μ^i_{T} when calculating the cross-entropy loss for the target. Similar to Eq. 7, the perturbed target segmentation logits $\tilde{l}^i_{\mathsf{T}}$ are constituted by replacing low-frequency part of its amplitude by the amplitude of segmentation logits of a source image randomly sampled from the batch. $\tilde{l}^i_{\mathsf{T}}$ can be written as,

$$\tilde{l}^i_{\mathsf{T}} = \mathcal{T}^{-1}([\mathsf{M}_\epsilon \cdot \mathcal{T}^A(l^i_{\mathsf{S}}) + (1-\mathsf{M}_\epsilon)\cdot \mathcal{T}^A(l^i_{\mathsf{T}}), \mathcal{T}^P(l^i_{\mathsf{T}})]) \tag{11}$$

PRN network takes in perturbed target logits and target feature from the teacher encoder, to output refined target logits $\bar{l}^i_{\mathsf{T}}$ and noise mask $\bar{\nu}^i_{\mathsf{T}}$. Here, $(\bar{l}^i_{\mathsf{T}}, \bar{\nu}^i_{\mathsf{T}}) = \mathcal{D}_\sigma(\mathcal{E}_\theta(x^i_{\mathsf{T}}), \tilde{l}^i_{\mathsf{T}})$. The binary mask ground truth $\mu^i_{\mathsf{T}}$is again created based on differences between the original and perturbed logits. Now, the cross-entropy loss $\mathcal{L}^{RT}_{ce}$ and binary cross-entropy loss $\mathcal{L}^{RT}_{bce}$ for training PRN for target data refinement and error localization can be written as,

$$\mathcal{L}^{RT(i)}_{ce} = -\sum_{j=1}^{H\times W}\sum_{k=1}^{K} \mathbb{1}[\mu^{(i,j)}_{\mathsf{T}} = 0]\ \bar{y}^{(i,j,k)}_{\mathsf{T}} \log(\bar{l}^i_{\mathsf{T}})^{(j,k)} \tag{12}$$

$$\begin{aligned}\mathcal{L}^{RT(i)}_{bce} = -\sum_{j=1}^{H\times W} \Big(\mu^{(i,j)}_{\mathsf{T}} \log(\bar{\nu}^i_{\mathsf{T}})^{(j)} \\ +(1-\mu^{(i,j)}_{\mathsf{T}}) \log(1-(\bar{\nu}^i_{\mathsf{T}})^{(j)})\Big)\end{aligned} \tag{13}$$

The student and refinement networks are trained simultaneously, but we consider one network fixed when calculating loss for the other. We discuss training loss in Sec. 3.5.

3.4. Contrastive Learning and Fourier Adaptation

To further stabilize the adaptation performance of model, we take two additional measures, i.e., pixel-wise contrastive loss (CL), and Fourier based style adaptation (FA).

We add the pixel-pixel contrastive loss $\mathcal{L}_{con}$ in training our student network. We hypothesize this addition will complement the source and target cross-entropy losses, for further improving the quality of our learned representations. Pixel-pixel contrastive loss has been shown in prior works to improve the training of semantic segmentation models [46]. $\mathcal{L}_{con}$ attempts to pull the features of pixels of the same object class (i.e., positive pairs) close and push away the features of pixels of different object classes (i.e., negative pairs). We randomly select pairs of source and test domain samples from the input batch to calculate the loss. For a pixel a, let $\mathbf{X}^a_{\mathsf{P}}$ denote the set of all positive samples (i.e., pixel collection belonging to the same class of pixel a). Similarly, let $\mathbf{X}^a_{\mathsf{N}}$ denote the set of all negative samples (i.e., pixels not belonging to the same class of pixel a).

$$\mathcal{L}_{con} = -\frac{1}{|\mathbf{X}^a_{\mathsf{P}}|}\sum_{a^+\in \mathbf{X}^a_{\mathsf{P}}} \log \frac{\mathbf{s}(f_a, f_{a^+})}{\mathbf{s}(f_a, f_{a^+})) + \sum_{a^-\in\mathbf{X}^a_{\mathsf{N}}} \mathbf{s}(f_a, f_{a^-})} \tag{14}$$

Here similarity function, $\mathbf{s}(f_a, f_{a^+})$ is calculated as $\exp(\cos(f_a, f_{a^+})/\zeta)$. ζ is the temperature hyper-parameter that controls the similarity magnitude. We utilize GT labels for source samples and refined labels from the PRN network for target samples to find the positive and negative samples. The noise binary mask is used to avoid the target image pixels predicted to have incorrect labels.

We adopt Fourier adaptation (FA) module following [55] to generate a synthetic source image with a target image style (without changing semantic content). As this module does not require any learning, it can be easily integrated into our pipeline with minimal additional load. The approach is similar to how we performed perturbation of logits (Eq. 7). However, we now focus on transforming the source image to the target style to reduce the perceptual gap between domains. These generated synthetic source images are then used in training instead of the original source images.

3.5. Overall Loss

The student network and PRN network are trained simultaneously. However, we consider the PRN network $\mathcal{F}_\sigma$ fixed when training the student network $\mathcal{F}_\theta$, i.e., the losses used for training the student network do not affect the PRN network. In this regard, we stop the gradient from flowing back in the PRN network. Similarly, we train the PRN network $\mathcal{F}_\sigma$ considering the student network $\mathcal{F}_\theta$ is fixed. Finally, the overall optimization problem can be written as follows,

$$\min_\theta(\mathcal{L}^T_{ce} + \mathcal{L}^S_{ce} + \lambda_1\mathcal{L}_{con}) + \min_\sigma(\lambda_2(\mathcal{L}^{RS}_{ce} + \mathcal{L}^{RS}_{bce}) + \mathcal{L}^{RT}_{ce} + \mathcal{L}^{RT}_{bce}) \tag{15}$$

Here, λ_1 and λ_2 are loss weight coefficients.

Table 1. Evaluation on **GTA→Cityscapes**. We report mean IoU (mIoU) over 19 categories on the Cityscapes validations set.

Method		Road	S.Walk	Build.	Wall	Fence	Pole	T.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.Bike	Bike	mIoU
CBST [61]	ResNet-Based	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
CCM [23]	ResNet-Based	93.5	57.6	84.6	39.3	24.1	25.2	35.0	17.3	85.0	40.6	86.5	58.7	28.7	85.8	49.0	56.4	5.4	31.9	43.2	49.9
MetaCor [13]	ResNet-Based	92.8	58.1	86.2	39.7	33.1	36.3	42.0	38.6	85.5	37.8	87.6	62.8	31.7	84.8	35.7	50.3	2.0	36.8	48.0	52.1
DACS [40]	ResNet-Based	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.2
UAPLR [47]	ResNet-Based	90.5	38.7	86.5	41.1	32.9	40.5	48.2	42.1	86.5	36.8	84.2	64.5	38.1	87.2	34.8	50.4	0.2	41.8	54.6	52.6
CorDA [45]	ResNet-Based	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 [56]	ResNet-Based	87.8	56.0	79.7	45.3	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
DACS (w/ PRN)	ResNet-Based	**92.7**	**48.6**	**88.9**	**43.2**	**33.3**	**43.8**	**49.0**	**38.0**	**88.4**	**44.0**	**86.5**	**70.1**	**45.0**	**90.0**	**41.4**	**50.6**	**42.0**	**45.3**	**58.7**	**57.9**
DAFormer [15]	SegFormer	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.2
MIC-DAFormer [17]	SegFormer	96.7	75.0	90.0	58.2	50.4	51.1	56.7	62.1	90.2	51.3	92.9	72.4	47.1	92.8	78.9	83.4	75.6	54.2	62.6	70.6
Ours	SegFormer	**95.8**	**73.3**	**92.8**	**56.2**	**51.9**	**51.6**	**59.6**	**62.8**	**93.1**	**49.9**	**96.3**	**76.1**	**47.0**	**96.3**	**77.7**	**81.7**	**68.2**	**59.9**	**64.3**	**71.3**
DAFormer (w/ HRDA) [16]	SegFormer	96.4	74.4	91.0	61.6	51.5	57.1	63.9	69.3	91.3	48.4	94.2	79.0	52.9	93.9	84.1	85.7	75.9	63.9	67.5	73.8
Ours (w/ HRDA)	SegFormer	**96.4**	**76.2**	**90.9**	**66.6**	**53.6**	**58.9**	**63.3**	**68.9**	**92.3**	**50.4**	**95.2**	**78.3**	**54.8**	**95.3**	**84.8**	**87.4**	**74.7**	**65.3**	**70.8**	**75.0**

4. Experiments

4.1. Experimental Setup

Datasets and Metrics. The proposed method is evaluated on several UDA semantic segmentation benchmarks, i.e., CityScapes→Dark Zurich, GTA→Cityscapes, SYNTHIA→Cityscapes. CityScapes (CS) contains daytime driving scenes from 50 different cities [7]. Following prior works [15, 40, 55], we use $2,975$ training and 500 validation images. Dark Zurich [34] is another driving dataset containing $8,779$ images (with GPS) captured at nighttime, twilight, and daytime with a resolution of 1080p. Dark Zurich (DarkZ) contains $2,416$ unlabeled nighttime images for training. It also contains 201 labeled nighttime images (50 validation, and 151 test) for evaluation. The evaluation images have pixel-level annotations for the 19 classes of Cityscapes. GTA [32] is a synthetic dataset containing $24,966$ images with resolution 1914×1052 collected from GTA5 video game. The 19 classes common with CityScapes are used by SOTA methods for evaluation. SYNTHIA (SYN) is another synthetic dataset with $9,400$ images with a resolution of 1280×760 [33]. The 16 classes common with CityScapes are used for evaluation.

Network Architecture. Our default semantic segmentation network model is based on the SegFormer architecture [50] following recent state-of-the-art works [4, 15]. The model consists of Transformer based MiT-B5 encoder (that generates hierarchical feature representation) and an MLP decoder (that aggregates information from multiple layers) [50]. We also train baselines with a ResNet101-based DeepLabV2 model. The Refine decoder utilizes the same MLP decoder architecture as described above. It utilizes aggregated information from the encoder output and segmentation logits. The number of channels in the input layers is increased to enable channel-wise concatenation of the two. It includes two output layers: one for predicting segmentation maps and another for binary noise masks.

Implementation Details and Metrics. We follow DAFormer [15] training parameters in training our models. Our model is trained using AdamW. A learning rate of 6×10^{-5} is used for the encoder. A learning rate of 6×10^{-4} is used for both the student decoder and refinement decoder. The AdamW betas are set to $(0.9, 0.999)$, and weight decay is set to 0.01. We use a batch size of 2, crop of 512X512 and train for $40K$ iterations. We use linear learning rate warmup with a warmup rate of 10^{-6} for the first $1.5K$ iterations. The EMA weight parameter β is set to 0.999. τ_2 is set to 0.968. The loss weights λ_1 is set to 0.1 and λ_2 is set to 25. We set mask parameter ϵ to 0.005 for the FA module. As the metric for semantic segmentation performance, we use mean intersection over union (mIoU). The results of our method are reported averaging over 3 random seeds.

Computational Load. We find our Framework increases training time by about 24.7% compared to the standard self-training (Sec. 3.2) baseline, DAFormer. Please note that the increase is only in training, while the inference time remains the same. The computation increase due to applying FFT is minimal (i.e., about 4.9% increase in computation time compared to without it). Overall, we find the training of our model was completed in a reasonable time with limited resources (e.g., using a single 2080Ti GPU in about 19 hours for CityScapes→Dark Zurich).

4.2. Experimental Results

We provide GTA→CS and CS→DarkZ quantitative results in this section. We also provide GTA→CS ablation study and some qualitative examples. The SYN→CS experiments and more qualitative examples are in the supplementary. Also, ablation studies (i.e., CS→DarkZ, SYN→CS, refinement loss weights) are in the supplementary.

4.2.1 GTA→Cityscapes Results

In Table 1, we compare the performance of our approach on GTA→Cityscapes adaptation, against several state-of-the-

Table 2. Evaluation on **Cityscapes→Dark-Zurich**. We report mean IoU (mIoU) over 19 common categories between these datasets.

Method		Ref.	Road	S.Walk	Build.	Wall	Fence	Pole	T.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.Bike	Bike	mIoU
Source-Only [5]	ResNet-Based	x	79.0	21.8	53.0	13.3	11.2	22.5	20.2	22.1	43.5	10.4	18.0	37.4	33.8	64.1	6.4	0.0	52.3	30.4	7.4	28.8
AdaptSegNet [41]		x	86.1	44.2	55.1	22.2	4.8	21.1	5.6	16.7	37.2	8.4	1.2	35.9	26.7	68.2	45.1	0.0	50.1	33.9	15.6	30.4
ADVENT [42]		x	85.8	37.9	55.5	27.7	14.5	23.1	14.0	21.1	32.1	8.7	2.0	39.9	16.6	64.0	13.8	0.0	58.8	28.5	20.7	29.7
BDL [24]		x	85.3	41.1	61.9	32.7	17.4	20.6	11.4	21.3	29.4	8.9	1.1	37.4	22.1	63.2	28.2	0.0	47.7	39.4	15.7	30.8
DACS [40]		x	83.1	49.1	67.4	33.2	16.6	42.9	20.7	35.6	31.7	5.1	6.5	41.7	18.2	68.8	76.4	0.0	61.6	27.7	10.7	36.7
DACS (w/ PRN)		x	**75.8**	**43.1**	**54.5**	**16.6**	**15.0**	**36.2**	**35.9**	**38.1**	**59.0**	**28.6**	**26.4**	**52.4**	**45.8**	**68.7**	**34.3**	**1.5**	**47.0**	**30.7**	**16.5**	**38.2**
DMAda [11]		✓	75.5	29.1	48.6	21.3	14.3	34.3	36.8	29.9	49.4	13.8	0.4	43.3	50.2	69.4	18.4	0.0	27.6	34.9	11.9	32.1
MGCDA [35]		✓	80.3	49.3	66.2	7.8	11.0	41.4	38.9	39.0	64.1	18.0	55.8	52.1	53.5	74.7	66.0	0.0	37.5	29.1	22.7	42.5
CDAda [51]		✓	90.5	60.6	67.9	37.0	19.3	42.9	36.4	35.3	66.9	24.4	79.8	45.4	42.9	70.8	51.7	0.0	29.7	27.7	26.2	45.0
DANIA [49]		✓	91.5	62.7	73.9	39.9	25.7	36.5	35.7	36.2	71.4	35.3	82.2	48.0	44.9	73.7	11.3	0.1	64.3	36.7	22.7	47.0
CCDistill [12]		✓	89.6	58.1	70.6	36.6	22.5	33.0	27.0	30.5	68.3	33.0	80.9	42.3	40.1	69.4	58.1	0.1	72.6	47.7	21.3	47.5
Source-Only [50]	SegFormer-Based	x	84.2	39.2	60.2	33.3	6.7	35.9	33.7	32.1	49.1	20.7	11.0	51.5	46.0	73.1	10.8	0.6	73.9	28.1	23.3	37.5
DAFormer [15]		x	93.5	65.5	73.3	39.4	19.2	53.3	44.1	44.0	59.5	34.5	66.6	53.4	52.7	82.1	52.7	9.5	89.3	50.5	38.5	53.8
MIC-DAFormer [17]		x	88.2	60.5	73.5	53.5	23.8	52.3	44.6	43.8	68.6	34.0	58.1	57.8	48.2	78.7	58.0	13.3	91.2	46.1	42.9	54.6
Refign [4]		✓	91.8	65.0	80.9	37.9	25.8	56.2	45.2	51.0	78.7	31.0	88.9	58.8	52.9	77.8	51.8	6.1	90.8	40.2	37.1	56.2
Ours		x	**94.3**	**74.8**	**82.5**	**53.2**	**26.4**	**62.3**	**43.6**	**49.9**	**66.3**	**37.1**	**69.3**	**67.9**	**61.9**	**81.2**	**53.9**	**13.8**	**90.5**	**44.7**	**35.0**	**58.4**
DAFormer (w/ HRDA) [16]		x	90.4	56.3	72.0	39.5	19.5	57.8	52.7	43.1	59.3	29.1	70.5	60.0	58.6	84.0	75.5	11.2	90.5	51.6	40.9	55.9
Ours (w/ HRDA)		x	**92.9**	**55.8**	**74.5**	**40.2**	**21.3**	**61.9**	**53.9**	**45.4**	**63.9**	**35.6**	**76.9**	**63.2**	**64.3**	**89.3**	**71.2**	**14.4**	**89.5**	**52.8**	**47.3**	**58.6**

art UDA semantic segmentation approaches and baselines. We divide the table into 3 parts to aid our study.

SOTA Performance. From the second part of the table, we see our method outperforms the best-performing prior state-of-the-art methods (i.e., DAFormer, MIC) by significant margins. We see $+3.1\%$ absolute improvement compared to DAFormer and $+0.7\%$ absolute improvement compared to MIC in mIOU (i.e., 68.2% with DAFormer and 70.6% with MIC compared to 71.3% with ours). Encouragingly, the mIoU improves consistently over most classes. We believe the masked image consistency module from MIC can be easily integrated into our method to further improve performance. We leave it as a future work.

Performance with HRDA Training. We report the performance of our model with HRDA-based training in the last row of Table 1, i.e., Ours (w/ HRDA). HRDA [16] is a recent UDA training approach that allows training with high-resolution images and has been shown to be able to boost UDA performance. We see incorporating HRDA training further improves our performance (75.0% vs. 71.3%). We also see the improvement is consistent with that of DAFormer with HRDA. Ours (w/ HRDA) outperforms DaFormer (w/ HRDA) by $+1.2\%$ absolute mIoU.

Pseudo-Label Refinement Methods. In the first part of Table 1, we compared with several prior SOTA PL refinement methods for ResNet CNN-based UDA semantic segmentation (e.g., CBST [70], CCM, MetaCor, UAPLR, ProDA). Among all the methods, we find incorporating our proposed PRN module with DACS, i.e., DACS (w/ PRN), performs the best. For example, Ours DACS (w/ PRN) achieves $+0.4\%$ improvement over ProDa and $+1.3\%$ over CorDA. Our explicit noise-handling approach with a learned auxiliary module helps us better address the domain gap compared to prior works, leading to potentially more resilient models with superior pseudo-labels.

4.2.2 Cityscapes→Dark Zurich Results

We report the performance of our approach on Cityscapes→Dark-Zurich adaptation in Table 2. We compare our approach with several state-of-the-art approaches. Most of the existing adverse-weather adaptation methods use additional reference images (depicting the same scene in the source domain) from the target domain, as auxiliary data to improve domain adaptation performance. Although these approaches show promising results, the requirement of collecting images of the same scene from both source and target domains, limits the applicability of these approaches. Note that our approach does not utilize any additional reference images, and still outperforms all these state-of-the-art approaches significantly.

In the first and second parts of the table, we report several CNN-based methods. In the third and fourth parts of the table, we report transformer-based methods. As expected, CNN-based models perform worse than the Trasformer-based models. From table 2, the best-performing state-of-the-art method is Refign [4], which also uses additional reference images to boost adaptation performance. Our method does not use any additional reference data, but still shows $+2.2\%$ absolute improvement in mIoU over Refign (i.e., 56.2% with Refign compared to 58.4% with ours). We also observe that our model performs the best in almost all the categories. Recent SOTA MIC-DAFormer [17] performs best among the methods that do not use reference images. Our model achieves an absolute improvement of $+3.8\%$ over MIC-DAFormer (i.e., 54.6% with MIC compared to 58.4% with ours). In the last part of the table, we also report the performance of our method and DAFormer incorporating HRDA training. We again observe that our method with HRDA performs significantly better ($+2.7\%$) than DAFormer with HRDA.

Table 3. Ablation study with different components of our proposed method on **GTA→Cityscapes**.

#	ST	PL-R	NM	CL w/o R	CL w/ R	FA	HRDA	mIoU
3.1	x	x	x	x	x	x	x	45.6
3.2	✓	x	x	x	x	x	x	68.3
3.3	✓	✓	x	x	x	x	x	69.7
3.4	✓	✓	✓	x	x	x	x	70.1
3.5	✓	✓	✓	✓	x	x	x	70.1
3.6	✓	✓	✓	x	✓	x	x	70.5
3.7	✓	✓	✓	x	x	✓	x	70.8
3.8	✓	✓	✓	x	✓	✓	x	**71.3**
3.9	✓	✓	✓	x	x	x	✓	74.3
3.10	✓	✓	✓	x	✓	x	✓	74.6
3.11	✓	✓	✓	x	x	✓	✓	74.8
3.12	✓	✓	✓	x	✓	✓	✓	**75.0**

4.2.3 Ablation Studies

In Table 3, We perform an ablation study with different components of the method, i.e., Self-Training (ST), Pseudo Label Refinement (PL-R), Noise Mask (NM), Contrastive Learning without or with using the output of PRN (CL w/o R, CL w/ R), Fourier Adaptation (FA) and HRDA training.

PRN Module: Comparing rows 3.1 and 3.2 in Table 3, it is evident that the baseline self-training approach is effective in adapting from source to target domain (i.e., mIoU improves $+22.7\%$). Based on rows 3.2, and 3.4, we see that our pseudo-label refinement model leads to $+1.8\%$ absolute improvement in mIoU and hence, playing a vital role in achieving state-of-the-art accuracy. Comparing 3.3 and 3.4, we see that including noise mask (NM) prediction in PRN along with pseudo-label refinement (PL-R) leads to significant improvement in performance.

CL and FA: Rows 3.5 and 3.6 indicate the performance change by adding contrastive learning (CL) module. Comparing row 3.4 with row 3.5, we see no performance change when the CL module does not use the PRN model output. However, comparing row 3.4 with row 3.6, we see a performance improvement of 0.4% when PRN model output is used. Based on this, we find that our proposed pseudo-label refinement module is critical for the effective use of pixel-wise contrastive learning. It is because its success relies on the quality of pseudo labels to select positive and negative pairs. Based on row 3.7 and row 3.8, we see adding FA style adaptation module helps to improve adaptation performance. Overall, we see CL and FA components are complementary to our main contribution. With the proposed PRN, they show consistent improvement across benchmarks.

Overall Framework: The difference between rows 3.2 and 3.8 shows that our approach helps significantly to improve performance over the baseline self-training approach ($+3.0\%$ absolute improvement in mIoU). Going further, we also show how HRDA-based training can boost our UDA performance. Rows 3.9, 3.10, and 3.11 show the effect of using HRDA training for experiments 3.4, 3.6, and 3.7 respectively, and all of them show a consistent improvement in absolute mIoU. Finally, comparing row 3.12 and row 3.8, we can see that HRDA-training provides a significant boost of 3.7% to our UDA pipeline. Please see the supplementary for more ablation studies.

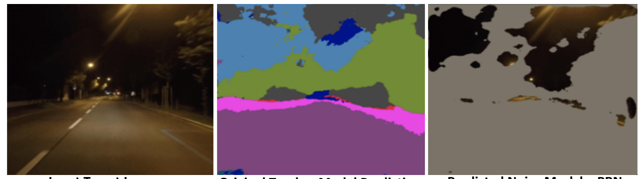

Figure 3. An example to analyze PRN module prediction quality.

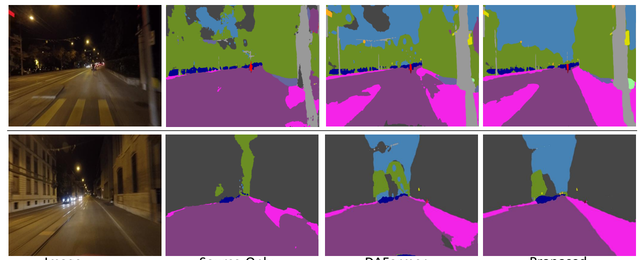

Figure 4. Qualitative comparison of the proposed on CS→DarkZ

4.2.4 Qualitative Results

We provide an example in Fig. 3 showing teacher prediction and predicted noise mask to qualitatively evaluate the PRN module. Another example of the original teacher model prediction and noise mask is shown in Fig. 1. We see noise mask generally corresponds well with the noisy part of pseudo labels in both cases. We also show two qualitative examples showing a semantic map generated from our model comparing source-only and DAFormer models in Fig. 4. We observe that the proposed method performs significantly better qualitatively than other approaches. Please see the supplementary for more qualitative results.

5. Conclusion

In this work, we present a novel self-training-based framework for the unsupervised adaptation of semantic segmentation models. We propose training auxiliary pseudo-label refinement network that helps self-training to be less susceptible to erroneous pseudo-label predictions by localizing and refining them. We also introduce two additional components, contrastive learning and Fourier-based style adaptation in our framework to further improve the quality of the trained model. Our proposed approach shows significant performance improvement compared to the previous state-of-the-art approaches in both normal-to-adverse and synthetic-to-real adaptation benchmarks.

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. 3

[2] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 2

[3] Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Sch¨olkopf, and Alex J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22:49–57, 2006. 2

[4] David Brüggemann, Christos Sakaridis, Prune Truong, and Luc Van Gool. Refign: Align and refine for adaptation of semantic segmentation to adverse conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3174–3184, 2023. 1, 2, 3, 6, 7

[5] 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. 1, 7

[6] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[7] 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. 1, 6

[8] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. *Domain adaptation in computer vision applications*, pages 1–35, 2017. 1, 2

[9] Gabriela Csurka, Riccardo Volpi, and Boris Chidlovskii. Unsupervised domain adaptation for semantic image segmentation: a comprehensive survey. *arXiv preprint arXiv:2112.03241*, 2021. 1

[10] Gabriela Csurka, Riccardo Volpi, and Boris Chidlovskii. Unsupervised domain adaptation for semantic image segmentation: a comprehensive survey. In *arXiv:2112.03241*, 2021. 2

[11] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE, 2018. 7

[12] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9913–9923, 2022. 7

[13] Xiaoqing Guo, Chen Yang, Baopu Li, and Yixuan Yuan. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3927–3936, 2021. 2, 6

[14] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu. Qiao. Adaptive pyramid context network for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[15] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 1, 2, 3, 6, 7

[16] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 372–391. Springer, 2022. 2, 3, 6, 7

[17] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023. 6, 7

[18] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[19] Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24120–24131, 2023. 1

[20] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 429–445. Springer, 2020. 2

[21] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[22] Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9957–9967, 2022. 2

[23] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2, 6

[24] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6936–6945, 2019. 2, 7

[25] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach.

In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[26] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *European conference on computer vision*, pages 415–430. Springer, 2020. 2

[27] Robert Mendel, Luis Antonio De Souza, David Rauber, Joao Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 141–157. Springer, 2020. 2

[28] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021. 3

[29] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015. 1

[30] Christian S. Perone, Pedro Ballester, Rodrigo C. Barros, , and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019. 2

[31] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346, 1982. 2

[32] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 6

[33] 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. 1, 6

[34] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383, 2019. 1, 2, 6

[35] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3139–3153, 2020. 7

[36] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[37] M Naseer Subhani and Mohsen Ali. Learning from scale-invariant examples for domain adaptation in semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 290–306. Springer, 2020. 2

[38] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1

[39] Junjiao Tian, Niluthpol Chowdhury Mithun, Zachary Seymour, Han-Pang Chiu, and Zsolt Kira. Striking the right balance: Recall loss for semantic segmentation. In *International Conference on Robotics and Automation (ICRA)*, pages 5063–5069. IEEE, 2022. 1

[40] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 1, 2, 3, 6, 7

[41] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 7

[42] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 7

[43] Kaihong Wang, Chenhongyi Yang, and Margrit Betke. Consistency regularization with high-dimensional non-adversarial source-guided perturbation for unsupervised domain adaptation in segmentation. In *Conference on Artificial Intelligence (AAAI)*, 2021. 2

[44] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 1

[45] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021. 6

[46] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021. 5

[47] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 9092–9101, 2021. 2, 6

[48] Zhonghao Wang, Mo You, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen mei Hwu, Thomas S. Huang, and Humphrey Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[49] Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. A one-stage domain adaptation network with image alignment for unsupervised nighttime semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):58–72, 2021. 7

[50] 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*, 34:12077–12090, 2021. 1, 6, 7

[51] Qi Xu, Yinan Ma, Jing Wu, Chengnian Long, and Xiaolin Huang. Cdada: A curriculum domain adaptation for nighttime semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2021. 7

[52] Jinyu Yang, Chunyuan Li, Weizhi An, Hehuan Ma, Yuzhi Guo, Yu Rong, Peilin Zhao, and Junzhou Huang. Exploring robustness of unsupervised domain adaptation in semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[53] Jianfei Yang, Xiangyu Peng, Kai Wang, Zheng Zhu, Jiashi Feng, Lihua Xie, and Yang You. Divide to adapt: Mitigating confirmation bias for domain adaptation of black-box predictors. *arXiv preprint arXiv:2205.14467*, 2022. 1

[54] 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 *Conference on Artificial Intelligence (AAAI)*, 2020. 2

[55] 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. 4, 5, 6

[56] 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 *Proc. IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 1, 2, 6

[57] Yang Zhang, Philip David, Hassan Foroosh, and Boqing Gong. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 42(8):1823–1841, 2020. 2

[58] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[59] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision (IJCV)*, 129(1):1106–1120, 2021. 2

[60] 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. 3

[61] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proc. European Conference on Computer Vision (ECCV)*, pages 289–305, 2018. 2, 6