Uncertainty-aware Pseudo Label Refinery for Domain Adaptive Semantic Segmentation

Yuxi Wang[1,2,5] Junran Peng[6] Zhaoxiang Zhang[1,2,3,4,5*]
[1] Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] University of Chinese Academy of Sciences (UCAS)
[3] Centre for Artificial Intelligence and Robotics, HKISI_CAS
[4] Center for Excellence in Brain Science and Intelligence Technology, CAS
[5] National Laboratory of Pattern Recognition (NLPR) [6] Huawei Cloud&AI.
wangyuxi2016@ia.ac.cn, jrpeng4ever@gmail.com, zhaoxiang.zhang@ia.ac.cn

Abstract

Unsupervised domain adaptation for semantic segmentation aims to assign the pixel-level labels for unlabeled target domain by transferring knowledge from the labeled source domain. A typical self-supervised learning approach generates pseudo labels from the source model and then retrains the model to fit the target distribution. However, it suffers from noisy pseudo labels due to the existence of domain shift. Related works alleviate this problem by selecting high-confidence predictions, but uncertain classes with low confidence scores have rarely been considered. This informative uncertainty is essential to enhance feature representation and align source and target domains. In this paper, we propose a novel uncertainty-aware pseudo label refinery framework considering two crucial factors simultaneously. First, we progressively enhance the feature alignment model via the target-guided uncertainty rectifying framework. Second, we provide an uncertainty-aware pseudo label assignment strategy without any manually designed threshold to reduce the noisy labels. Extensive experiments demonstrate the effectiveness of our proposed approach and achieve state-of-the-art performance on two standard synthetic-2-real tasks.

1. Introduction

Semantic segmentation [18], considered as one of the fundamental problems in Computer Vision, aims to understand the image scene at the pixel level. Since the increasing numbers of images, recent advances in semantic segmentation have shown rapid progress on current datasets, such as Pascal VOC-2012 [8] and Cityscapes [5]. However, collecting large-scale real-world datasets with well-labeled dense annotations is labor-intensive and time-consuming. To overcome this limitation, one feasible solution is to train a model using synthetic and realistic datasets with free annotations, exampling as GTA5 [25] and SYNTHIA [27], and predict on the real-world dataset like Cityscapes [5]. Unfortunately, the inevitable question is that the performance will drop significantly due to the domain shift between the synthetic and the real data.

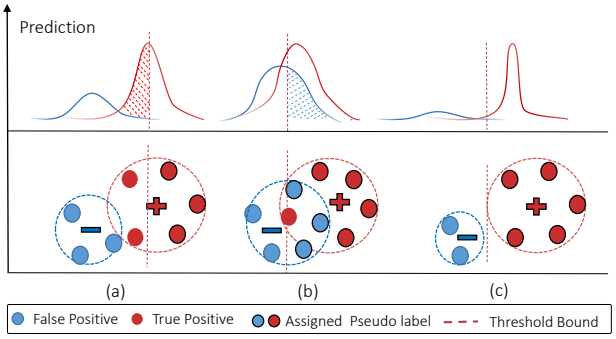

Figure 1. Problems for existing threshold-based pseudo labels generation. (a) Ignoring true positive predictions; (b) Assigning false positive predictions as labels (c) Ideal pseudo labels generation.

To address domain discrepancy, unsupervised domain adaptation methods have been proposed for semantic segmentation [12, 31, 25, 27]. Seminal practices usually exploit adversarial learning or self-supervised learning techniques. For adversarial learning methods, the dominant trend is to match the distributions of source and target domains at different levels: pixel level [11, 4, 14, 21], feature level [12, 3], output level [31, 33, 35], category level [20, 7], and patch level [32]. For self-supervised learning approaches, the key idea is to generate high-quality pseudo labels [41, 40, 30]. Although these methods significantly

*Corresponding Author

improve the adaptation performance, it still lags far behind supervised learning or semi-supervised learning.

After dissecting the domain adaptive semantic segmentation, we observe two key ingredients are ignored in previous works. First, due to the class-imbalance, different categories are prone to have distinct transferability. Some classes, such as road and building that occupy a large portions of pixel, are inherently easy to transfer across domains. Second, typical manually designed threshold methods [41, 40] generate pseudo labels according to the confidence scores, which is substantially hindered by the inevitable label noise. Incorrect pseudo labels with high confidence score can confuse the network in the target domain. The drawback of the confidence score based method is shown in Fig 1.

In this paper, we propose a target-guided uncertainty rectifying method and an uncertainty-aware pseudo labels assignment technique to address the above two issues, respectively. (1) Trained models trend to produce high-uncertainty predictions for minority classes. To remedy this issue, we resort to resampling strategies [23] to progressively refine high-uncertain predictions during the adversarial training process. We achieve this goal by resampling source data according to the uncertainty statistics of the target domain. (2) To alleviate noisy labels caused by uncertainty predictions, we propose an uncertainty-aware pseudo labels assignment strategy to generate reliable target labels. We assume the certain and uncertain predictions following different distributions and estimate them using a Gaussian Mixture Model of two modalities. We refer to our method as *UncerDA* since we heavily rely on uncertainty information for domain adaptation.

The contributions of this paper can be summarized as follows:

- We propose to enhance the distribution alignment by resampling the training source images, whereas the resampling classes are designed according to the uncertainty statistics of the target domain. This tailored cross-domain setting benefits the learning of the transferable model.
- We propose to select reliable pseudo labels by fitting the predictions to certainty and uncertainty modes using GMM. Pixels belonging to the certainty mode are assigned as pseudo labels.
- Comprehensive experiments demonstrate the effectiveness of the proposed method, achieving the state-of-the-art performance on both GTA5→Cityscapes and SYNTHIA→Ciyscapes benchmarks.

2. Related Work

Domain Adaptation for Semantic Segmentation (DASS) aims to train a network that can assign pixel-level labels to unlabeled target data by learning from labeled source data. Existing methods in the literature can be roughly categorized into two groups: adversarial learning methods and self-training methods. For adversarial learning, numerous works have been explored to align source and target distributions in the pixel-level [11, 37, 36, 14], feature-level [12], output-level [31, 33, 20], and patch-level [32]. These methods examine all kinds of domain-invariant information to match the distributions between domains. *Tuan et al*. [33] leverages self-entropy maps and *Myeongjin et al*. [14] uses texture-invariant [14] information to help alignment. Besides, to achieve fine-grand matching, researchers [20, 7] propose a category-level adversarial network for each class. *Zhou* [34] has refined this category-level alignment to things and stuff. In [37, 36], the appearance of target images is transferred to source data, which prompts the images to be domain-invariant. Other works attempt to reduce discrepancy utilizing data augmentation, exampling as a GAN-based self-enhanced method has been introduced in [4]. Though these methods have explored various invariant information between source and target domain, they ignore a vital mismatching problem caused by minority categories or infrequent pixels. Thus, we focus on this challenging problem to enhance adaptation in this paper.

For self-training approaches, the essential idea is to generate reliable pseudo labels. [41, 40] utilize a class-balance self-training (CBST) for domain adaptive semantic segmentation, which generates pseudo labels depending on category-level confidence. [7] proposes a progressive strategy following a constant threshold. It may suffer from label noise, leading to incorrect alignment. In [16], researchers consider curriculum learning and offer a self-motivated pyramid framework for semantic segmentation. *Subhani* and *Ali* [30]'s dynamic entropy dependent pseudo labels generation methods could be the closest to our work. However, they only focus on label selection in the second stage while ignoring robust segmentation network training. We enhance the capability of feature representation for a model by rectifying uncertain minority classes.

Imbalanced Learning has been widely studied, including resampling and reweighting techniques. Resampling methods [28, 22, 24, 9, 23] directly balance the class distribution via modifying the training samples. Reweighting approaches, such as FocalLoss [17] and OHEM [29], assign a specific loss weight to alleviate the classifier's bias. Due to the co-occurrence problem, these resampling strategies are not suitable for our tasks, so we modify a weighted soft resampling strategy in this paper.

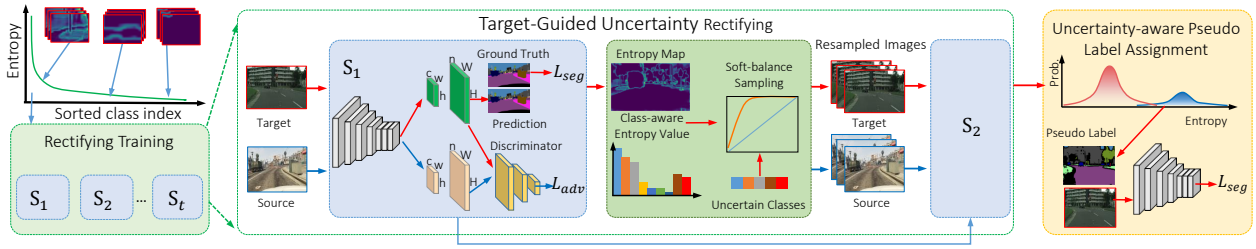

Figure 2. Example of the proposed uncertainty pseudo label refinery framework. (Best viewed in color)

3. Preliminary

Before illustrating our method, we conduct pilot experiments to demonstrate the self-training methods benefit from the initialized feature alignment model and effective pseudo labels generation method. To this end, we adopt different adaptation approaches as initialized models, such as AdaptSegNet [31], CLAN [20], ADVENT [33], MRNet [38] and SIM [34]. Then we compare conventional pseudo labels selection and the manually designed pseudo labels learning as shown in Fig 3. Note that the manually designed pseudo labels are generated from the guidance of ground truth information. The incorrect predictions with higher confidence than the threshold are rectified as ground truth, while those lower than the threshold are ignored.

It is intuitive to reason the improvement because these two aspects provide efficient supervision information for the target data. First, the model with more powerful adaptation performance can align distributions better between domains. If the feature alignment model reduces the domain gap, the ideal outputs of target features should be similar to those of source features. It will produce superior segmentation results via source domain supervised learning. Second, a high-quality pseudo labels generation strategy can improve significantly, and the optimal solution is equal to the fully supervised scenario.

In this paper, we pursue the above two observations to boost the quality of pseudo labels. We first enhance feature alignment during the adversarial adaptation process by focusing on challenging classes with high-uncertainty predictions. On the other hand, we use uncertainty information to guide the selection of pseudo labels. Contrary to previous methods, assigning pseudo labels using uncertainty can effectively reduce incorrect labels with high-confidence scores. Thus, adopting uncertainty information to boost feature alignment and pseudo labels assignment can generate reliable pseudo labels, as shown in Sec.5.

4. Methodology

In this work, we explore the uncertainty of target predictions to denoise pseudo labels from two aspects: (1) rectifying the uncertainty-aware predictions via target-guided resampling strategy and (2) proposing uncertainty-aware pseudo labels assignment to select correct predictions. The overview pipeline is shown in Fig. 2, and we will elaborate details in the following.

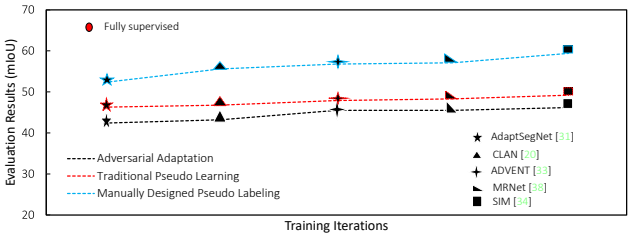

Figure 3. Results on GTA5→Cityscapes. Two different pseudo labels generation methods are adopted on several existing adaptation models.

4.1. Target-guided uncertainty rectifying

In domain adaptive semantic segmentation, the uncertain predictions of the target data correspond to the classes with infrequent pixels or small objects. These classes often occupy an insignificant portion of the image. Treating these classes equally to major classes may make the network underfitting due to insufficient training. To address this issue, we propose a specific resampling strategy to progressively reduce the target data's uncertainty. The key idea is to locate classes with high uncertainty and then compute an instance sampling probability for the source data based on these classes. The feature alignment is achieved by adversarial training on the sampled source and target batch.

Formally, given the source dataset $\mathcal{X}_s = \{x_s\}_{j=1}^{n_s}$ with the corresponding labels $\mathcal{Y}_s = \{y_s\}_{j=1}^{n_s}$ and unlabeled target dataset $\mathcal{X}_t = \{x_t\}_{j=1}^{n_t}$, we use entropy to characterize the uncertainty of predictions on the target domain. That is, the predictions with low (high) entropy are considered as certain (uncertain) samples. To locate uncertainty-aware classes, we first calculate average category-level entropy $I_{\mathcal{X}_t}^c$ on the whole target domain:

$$I_{\mathcal{X}_t}^c = \frac{1}{N_c} \sum_{x_t \in \mathcal{X}_t} \sum_i I_{x_t}^{(i)} * \mathbb{1}(\hat{y}_{x_t}^{(i,c)} = 1), \qquad (1)$$

where $\mathbb{1}$ is the indicator function and N_c indicates the

number of pixels for the c-th class in the whole dataset. $\hat{y}_{x_t}^{(i,c)} = \arg\max_c p_{x_t}^{(i,c)}$ represents the pseudo label of the i-th pixel in x_t belonging to the c-th class. $p_{x_t}^{(i,c)}$ represents the softmax probability of pixel $x_t^{(i)}$ and $p_t = F_{\theta_f}(x_t; \theta_f)$. F_{θ_f} is the initialized segmentation model with parameters θ_f. I_{x_t} refers to the normalized pixel-wised entropy map overall C classes:

$$I_{x_t}^{(i)} = -\frac{1}{\log(C)} \sum_{c=1}^{C} p_{x_t}^{(i,c)} \log p_{x_t}^{(i,c)}. \tag{2}$$

Then, we rank $I_{\mathcal{X}_t}^c$ and obtain a subset S_k with top-k high-uncertain classes. Classes in S_k are hard to transfer due to the minor proportion or class imbalance. We remedy this issue via an instance-level resampling strategy to provide sufficient training samples for these rare classes on the next adversarial learning process.

Soft-balance sampling for uncertainty-ware classes. To refine uncertainty-aware classes, we compute an image-level sampling probability for the source data according to target predictions. At first, we consider the whole C classes to calculate the class-balance sampling probability p_c for the source data $\mathcal{X}_s$ pursuing [23]:

$$p_c(\mathcal{X}_s) = \frac{N_c(\mathcal{X}_s)^\lambda}{\sum_{c=1}^{C} N_c(\mathcal{X}_s)^\lambda} \frac{1}{N_c(\mathcal{X}_s)}, \tag{3}$$

where $N_c(\mathcal{X}_s)$ indicates the number of pixels for the c-th class in $\mathcal{X}_s$ and λ is a hyper-parameter to soften the discrepancy between frequent and infrequent categories. It should be noticed that λ controls the sampling strategy from no-balance ($\lambda = 1$) to class-aware balance ($\lambda = 0$). Second, considering label co-occurrence that one image can contain multiple categories, a given image x_s with a label y_s could be repeatedly sampled by each category it contains. Thus the image-level sampling probability can be estimated as a weighted summation for contained categories:

$$p_i(x_s) = \sum_{\hat{c}} \frac{N_{\hat{c}}(x_s)^\lambda \mathbb{1}(y_s^{\hat{c}} = 1)}{\sum_{\hat{c}} N_{\hat{c}}(x_s)^\lambda \mathbb{1}(y_s^{\hat{c}} = 1)} p_{\hat{c}}(\mathcal{X}_s), \tag{4}$$

where $N_{\hat{c}}(x_s)$ denotes the number of pixels for the $\hat{c}$-th category contained in the image x_s and $\hat{c} \in S_k$. We only consider uncertainty-aware classes in S_k to rectify uncertainty during this process. Due to the obtained probability usually closes and sometimes goes towards zero, so we design a smoothing function for $p_i(x_s)$ as Eq.(5) shows,

$$\hat{p}_i(x_s) = 0.1 + \frac{1}{1 + \exp(-\alpha \times (p_i(x_s) - \mu))}. \tag{5}$$

Here, α and μ control the smooth shape of sampling probability, and the selected values are studied in subsection 5.8.

Progressively rectifying scheme. We refine the uncertain-aware classes in S_k by resampling source images to enhance feature alignment. After an adversarial adaptation process, classes in S_k should be updated. Then the S_k is used to calculate a new image-level sampling probability according to Eq.(4). A new adversarial adaptation on the sampled source and target batches is construted based on the latest sampling probability. We repeat this rectifying scheme until the classes in S_k don't change, or the adaptation performance doesn't improve. We enhance feature alignment from this iterative refinery by involving more source samples belonging to the hard classes. The pipeline is shown in Fig.2.

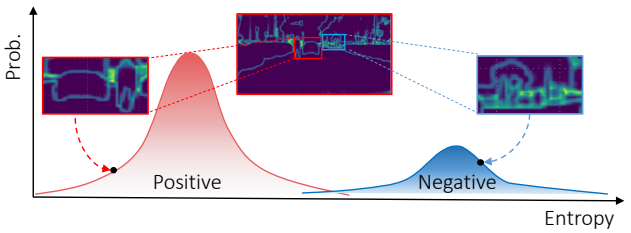

Figure 4. Illustration of uncertainty-aware pseudo label assignment. The entropy is modeled as a probability distribution using Gaussian Mixture Model of negative and positive predictions. Pixels belong to positive part are assigned as pseudo labels.

4.2. Uncertainty-aware pseudo label assignment

The rectified model can produce reliable predictions because the feature extractor F_{θ_f} generates enhanced target features. However, due to the domain gap, the noisy labels still exist because of the incorrect predictions with high confidence, which leads to poor results and poor generalization. To remedy this problem, we resort to uncertainty information to generate pseudo labels. We observe that the predictions with low/high uncertainty usually correspond to correct/incorrect pseudo labels. Therefore, the uncertainty-aware information can significantly separate pseudo labels into correct (positive) and incorrect (negative) parts.

Motivated by this, we propose a novel uncertainty-aware pseudo labels assignment strategy according to the target uncertainty predictions. Intuitively, we can infer from the uncertainty if a labeled sample is more likely to be positive (correct) or negative (incorrect). We achieve this goal by fitting a mixture distribution model, as shown in Fig. 4. Specifically, we use a Gaussian Mixture Model (GMM) with two components to fit positive and negative distributions. Samples belonging to the positive distribution are selected as pseudo labels. Furthermore, considering class-imbalance, we fit the distribution for the category-level entropy $I_{x_t}^c$. The probabilistic distribution of the c-th class can be obtained as:

$$\begin{aligned} P_c(I_{x_t}^c) = & \, w_{neg} \mathcal{N}_{neg}(I_{x_t}^c; \mu_{neg}, \sigma_{neg}) \\ & + w_{pos} \mathcal{N}_{pos}(I_{x_t}^c; \mu_{pos}, \sigma_{pos}), \end{aligned} \tag{6}$$

where $w_{neg}, \mu_{neg}, \sigma_{neg}$ and $w_{pos}, \mu_{pos}, \sigma_{pos}$ denote the

Table 1. Comparison to state-of-the-art methods of adaptation from GTA5 to Cityscapes based on ResNet-101 backbone. The top group is for adversarial adaptation ("AA"), while the bottom represents performance using self-training learning ("ST").

Method	road	side.	build.	wall	fence	pole	light	sign	vege.	terr.	sky	person	rider	car	truck	bus	train	motor.	bike	**mIoU**
AdaptSeg [31]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
SIBAN [19]	88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	**40.0**	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
CyCADA [11]	86.7	35.6	80.1	19.8	17.5	**38.0**	39.9	**41.5**	82.7	27.9	73.6	**64.9**	19.0	65.0	12.0	28.6	4.5	31.1	42.0	42.7
CLAN [20]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	**6.7**	31.9	31.4	43.2
DISE [1]	91.5	47.5	82.5	31.3	25.6	33.0	33.7	25.8	82.7	28.8	**82.7**	62.4	30.8	85.2	27.7	34.5	6.4	25.2	24.4	45.4
ADVENT [33]	89.4	33.1	81.0	26.6	**26.8**	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
PatchAlign [32]	**92.3**	**51.9**	82.1	29.2	25.1	24.5	33.8	33.0	82.4	32.8	82.2	58.6	27.2	84.3	33.4	**46.3**	2.2	29.5	32.3	46.5
MRNet [38]	89.1	23.9	82.2	19.5	20.1	33.5	42.2	39.1	85.3	33.7	76.4	60.2	33.7	**86.0**	36.1	43.3	5.9	22.8	30.8	45.5
Ours (AA)	88.7	31.2	**83.7**	**34.1**	24.1	37.6	**42.9**	33.0	**85.8**	38.9	80.3	63.7	**34.2**	85.9	**41.2**	42.5	3.4	**33.8**	**42.5**	**48.8**
LSE [30]	90.2	40.0	83.5	31.9	26.4	32.6	38.7	37.5	81.0	34.2	84.6	61.6	33.4	82.5	32.8	45.9	6.7	29.1	30.6	47.5
PLCA [13]	84.0	30.4	82.4	35.3	24.8	32.2	36.8	24.5	85.5	37.2	78.6	**66.9**	32.8	85.5	40.4	48.0	8.8	29.8	41.8	47.7
BDL [15]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	**43.6**	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
SIM [34]	90.6	44.7	84.8	34.3	28.7	31.6	35.0	37.6	84.7	43.3	85.3	57.0	31.5	83.8	**42.6**	48.5	1.9	30.4	39.0	49.2
TextDA [14]	**92.9**	**55.0**	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	**87.1**	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2
FDA [36]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	**53.1**	**16.9**	27.7	46.4	50.5
Zhe *et al*. [39]	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
Ours (ST)	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**

weights, means and variances of the negative and positive Gaussians, respectively. $I^c_{x_t}$ is an image-level entropy for the c-th class. To estimate these parameters of this GMM in Eq. (6), we use the *Expectation-Maximization* (EM) algorithm to optimize the distributions and weights (w_{neg}, w_{pos}) following a uniform distribution. Once the distribution is estimated, the correct pseudo labels can be easily selected from the positive distribution.

Compared to [26], our method denoises pseudo labels by estimating uncertainty to a bimodal distribution without any manually designed thresholds. Moreover, the proposed method can address the class-imbalance problem. Pseudo labels are selected at class-level, which treats different classes equally regardless of their occurrence frequency.

5. Experiments

5.1. Datasets

We evaluate the proposed method on two challenge adaptation tasks from the synthetic to the real domain. Synthetic datasets that have abundant pixel-level annotations act as source domains, including GTA5 [25] and SYNTHIA [27]. At the same time, the real-world dataset Cityscapes [5] that has zero label is considered as the target domain.

GTA5 is selected from a video computer game based on the urban scenery of Los Angeles city, along with pixel-level labels for 33 different categories. It contains 24,966 images with a resolution of 1,914×1,052. During training, we resize images to 1,280 × 720 and then random crop them to 1,024 × 512. We only consider 19 categories in common with Cityscapes [5], similar to previous methods.

SYNTHIA is another synthetic dataset that contains 9,400 annotated images with a resolution of 1,280×760. We also random crop the images to 1,024×512, and 16 standard categories with the Cityscapes dataset are considered for training. The evaluation is performed on both the 16- and 13- class subsets following the standard protocol.

Cityscapes is a real-world dataset collected from urban street scenes including 18 different cities around Germany and neighboring countries. It has 2,975 training images and 500 validation images with a resolution of 1,024×512. The two domain adaptation scenarios are constructed as GTA5→Cityscapes and SYNTHIA→Cityscapes.

5.2. Implementation Details

In the experiments, we utilize the multi-level adaptation framework similar to [31] and [33] to train the segmentation network F_{θ_f} and adversarial discriminator. The architecture of F_{θ_f} is DeepLab-v2 [2], which adopts the ResNet-101 [10] pre-trained on ImageNet [6] as the backbone model. We first train the adaptation model using a conventional sampling strategy for 25,000 iterations, which is sampling source and target images with an equivalent sampling probability. Then we attempt to align the distributions of source and target domains utilizing the proposed target-guided uncertainty rectifying strategy progressively.

5.3. Comparisons with state-of-the-art methods

We comprehensively compare the proposed method with state-of-the-art domain adaptation approaches in Table 1 and Table 2. The compared methods can be divided into 1) domain alignment through adversarial adaptation and 2) self-training approaches. We demonstrate the effectiveness of our method on both the adversarial adaptation stage and

Table 2. Comparison to state-of-the-art methods of adaptation from SYNTHIA to Cityscapes based on ResNet-101 backbone. The top group is for adversarial adaptation ("AA"), while the bottom group represents performance using self-training learning ("ST"). mIoU and mIoU* are averaged over 16 and 13 categories.

Method	road	side.	build.	wall	fence	pole	light	sign	vege.	sky	person	rider	car	bus	motor.	bike	**mIoU**	**mIoU***
AdaptSeg [31]	79.2	37.2	78.8	**10.5**	0.3	25.1	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.3	39.5	45.9
PatchAlign [32]	82.4	38.0	78.6	8.7	0.6	26.0	3.9	11.1	75.5	**84.6**	53.5	21.6	71.4	32.6	19.3	31.7	40.0	46.5
CLAN [20]	81.3	37.0	80.1	-	-	-	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	**22.6**	30.7	-	47.8
ADVENT [33]	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	**80.4**	84.1	57.9	23.8	73.3	**36.4**	14.2	33.0	41.2	48.0
DISE [1]	**91.7**	**53.5**	77.1	2.5	0.2	27.1	6.2	7.6	78.4	81.2	55.8	19.2	82.3	30.3	17.1	34.3	41.5	48.8
Ours (AA)	81.2	35.6	**81.5**	9.9	**0.8**	**35.9**	**29.6**	**19.9**	78.9	78.1	**62.8**	**27.1**	**83.7**	27.9	16.8	**53.1**	**45.2**	**52.0**
TextDA [14]	**92.6**	**53.2**	79.2	-	-	-	1.6	7.5	78.6	**84.4**	52.6	20.0	82.1	34.8	14.6	39.4	-	49.3
LSE [30]	82.9	43.1	78.1	9.3	0.6	28.2	9.1	14.4	77.0	83.5	58.1	25.9	71.9	38.0	29.4	31.2	42.6	49.4
BDL [15]	86.0	46.7	80.3	-	-	-	14.1	11.6	79.2	81.3	54.1	27.9	73.7	**42.2**	25.7	45.3	-	51.4
SIM [34]	83.0	44.0	80.3	-	-	-	17.1	15.8	**80.5**	81.8	59.9	**33.1**	70.2	37.3	28.5	45.8	-	52.1
FDA [36]	79.3	35.0	73.2	-	-	-	19.9	24.0	61.7	82.6	61.4	31.1	83.9	40.8	**38.4**	51.1	-	52.5
Ours (ST)	79.4	34.6	**83.5**	**19.3**	**2.8**	**35.3**	**32.1**	**26.9**	78.8	79.6	**66.6**	30.3	**86.1**	36.6	19.5	**56.9**	**48.0**	**54.6**

self-training stage, denoted as *UncerDA (AA)* and *UncerDA (ST)*. Specifically, the *UncerDA (AA)* represents the performance of the proposed target-guided uncertainty rectifying strategy, and *UncerDA (ST)* illustrates the uncertainty-aware pseudo label assignment method. Recent works [1, 11, 15, 32] have revealed that pixel-wise translation from source to target can enhance performance. We follow this practice in the final model. The difference is that we treat translated source images with target style and original contents as target data.

Observing Table 1 and Table 2, we can conclude that: (1) The proposed method achieves state-of-the-art results with 52.6% and 54.6% mIoU on GTA5/SYNTHIA→Cityscapes, respectively, which significantly outperforms other methods. (2) Compared to existing approaches, our method gains improvement not only on the adversarial adaptation (AA) stage (*i.e*., 48.8% *vs*. 45.5% in Table 1), but also on the self-training (ST) stage (*i.e*., 52.6% *vs*. 50.3% in Table 1). It demonstrates the effectiveness of the proposed target-guided uncertainty rectifying and uncertainty-aware pseudo labels assignment for matching the distributions of source and target domains. (3) *UncerDA* achieves evident advantage in hard classes, *e.g. fence, pole, motor, and bike*. The results revels that our method can handle challenging small or rare objects thanks to rectifying strategy and reliable pseudo labels. We provide visual examples of prediction results in Fig. 5.

5.4. Influence of Different Components

We dissect the contributions of each component to the overall performance. In Table 3, the first group indicates the baseline model with adversarial adaptation (AA) and image translation (IT), increasing the performance from 36.6% to 45.3%. SR represents applying the proposed soft-balance resampling to the source domain ignoring target uncertainty, which achieves performance to 46.1%. Through target-guided resampling, the performance of model TGAA is improved to 48.8%. It verifies that the proposed target-guided uncertainty rectifying has two advantages: 1) It surpasses naive resampling techniques thanks to considering cross-domain information; 2) It is effective to refine uncertainty-aware prediction for adaptation. Finally, the self-training of proposed uncertainty-aware pseudo labels (UPST) provides significant improvement to 52.6%.

Table 3. Ablation study on GTA5→Cityscapes. AA + IT acts as the baseline model with adversarial adaptation and image translation techniques; SR indicates the proposed soft-balance resampling strategy on source domain; TGAA is target-guided uncertainty rectifying adversarial adaptation; UPST stands for the proposed uncertainty-aware pseudo labels self-training process.

Method	AA	IT	SR	TGAA	UPST	mIoU
Source Only						36.6
+AA [31]	✓					42.9
+IT [15]	✓	✓				45.3
+SR	✓	✓	✓			46.1
+TGAA	✓	✓	✓	✓		48.8
+UPST	✓	✓	✓	✓	✓	**52.6**

5.5. The influence of soft-balance sampling

To testify the effectiveness of the proposed soft-balance sampling strategy, we compare it to other balance methods in Table 4. We utilize the conventional multi-level adaptation framework [31, 20] with no-balance training as our baseline model, achieving 42.9% mIoU on GTA5 → Cityscapes task. The class-balance technique boosts the adaptation performance to 45.7%, with sampling all categories data equally. It reveals that the balanced training strategy remedy for domain bias by enhancing the sam-

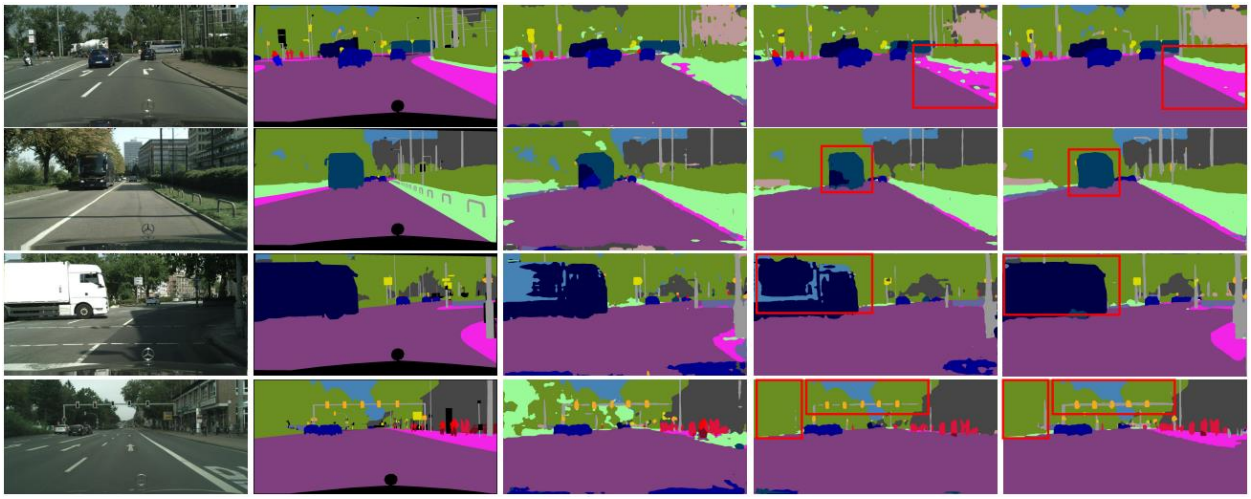

Figure 5. Visualization of the segmentation results. We perform our results from adversarial adaptation ("AA") and self-training learning ("ST"), respectively. The "baseline" model is achieved with adversarial learning and image transfer.

Table 4. The comparison of different sampling methods. Models are evaluated on GTA5→Cityscapes task.

Methods	λ	mIoU
No-Balance	-	42.9
Class-Balance	-	45.7
Focal Loss [17]	-	44.5
Soft-Balance	0.3	44.7
	0.5	44.8
	0.7	45.9
	0.9	**46.1**
	1.0	45.7

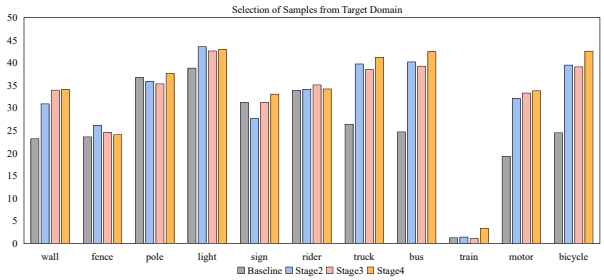

Figure 6. The influence of uncertainty-aware rectifying for selected infrequent categories on different sampling stages.

pling of infrequent categories. Focal loss [17] is used to eliminate the category imbalance problem by controlling the gradient contribution from different categories. Compared to the no-balance training, the focal loss can provide 1.6 points improvement (while it is worse than the class-balance sampling). The proposed soft-balance strategy with hyper-parameter λ provides an extended range of sampling probabilities. The λ controls the sampling strategies from no-balance ($\lambda = 0$) to class-balance ($\lambda = 1.0$), and we achieve a peak point at $\lambda = 0.9$, with the performance of 46.1% mIoU, outperforming class-balance sampling by 0.4 points.

5.6. The influence of uncertainty rectifying

In this subsection, we further investigate the effectiveness of the proposed target-guided uncertainty rectifying technique on different stages. The target-guided sampling can affect the infrequent categories at each rectifying stage, as shown in Fig 6. From this graph, we can observe that: 1) The performance of minority classes has a significant improvement compared to the baseline model. 2) The target-guided resampling strategy can provide sufficient training data and avert the over-fitting problem.

We also present some visualization results of the segmentation prediction probability for infrequent categories in Fig. 7. For selected categories, the likelihood of the baseline model (the third row) is low, revealing that these categories are challenging for alignment and usually lead to uncertainty predictions. After our rectifying process, the predictions of these imbalanced categories become confident (the fourth row) due to sufficient samples to train.

5.7. The influence of uncertainty-aware pseudo labels assignment

We construct comparison experiments based on different models and strategies to generate pseudo labels to verify the scalability of the proposed uncertainty-aware pseudo label

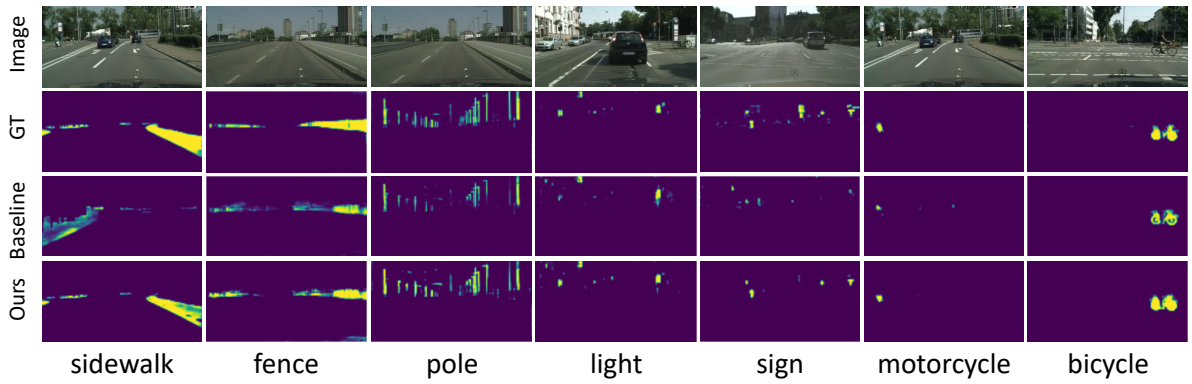

Figure 7. Visualization of category-wise segmentation prediction probability for the selected imbalance classes.

Table 5. The influence of different pseudo labels. The model named in the "Pseudo Label" column denotes that we deploy the corresponding model's pseudo label.

Models	Pseudo Label	mIoU
AdaptSegNet [31]	-	42.4
(a)	CBST [41]	47.1
(b)	Ours	**47.8**
MRNet [38]	-	45.5
(c)	CBST [41]	49.7
(d)	Ours	**51.2**
UncerDA (AA)	-	48.8
(e)	CBST [41]	51.2
(f)	Ours	**52.6**

assignment. The target model is refined by pseudo labels following the variance restraint [39]. First, we adopt the proposed method to AdaptSegNet [31] and MRNet [38] to generate pseudo labels. As shown in Table 5, the proposed method improves performance from 42.2% to 47.8% and from 45.5% to 51.2% with the same trends. Second, we also compare the other pseudo-label generation method such as CBST [41]. The results demonstrate that our method is also superior to CBST on these three different models. Meanwhile, the performance based on our UncerDA (AA) model is the best with achieving 52.6%, which surpasses AdaptSegNet and MRNet with 4.8 and 1.4 points.

5.8. Parameters analysis

In this subsection, we analyze the hyper-parameters introduced in our work. First, to ensure the smooth shape of sampling probability, we adjust α and μ in Eq. (5) as Fig. 8 (a) shows. We choose $\alpha = 80$ and $\mu = 0.02$ to increase probability rapidly near 0 and tend to flat near 1. Second, for parameter k that indicates top-k uncertainty classes in the target data, we vary it and plot the performance in Fig. 8 (b) on GTA5/SYNTHIA→Cityscapes. The result shows

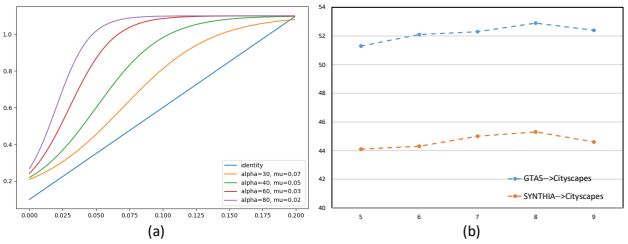

Figure 8. (a) shows the parameters selection for α and μ. (b) shows the parameters selection for k.

that our model is stable to k in the range of 5 to 9.

6. Conclusion

We identify that the challenge of pseudo label generation in self-supervised domain adaptive semantic segmentation lies in the uncertainty-aware alignment related to the class-imbalance distribution. The proposed target-guided uncertainty rectifying method effectively enhances representation for minority classes by applying a soft-balance resampling strategy for classes with high uncertainty in the target domain. The proposed pseudo label assignment method reduces label noise by estimating two different distributions for negative and positive predictions, which can effectively model the correct and incorrect pseudo labels. These two strategies provide robust and reliable pseudo labels for training.

7. Acknowledgements

This work was supported in part by the National Key R&D Program of China(No. 2018YFB1402605),the National Natural Science Foundation of China (No. 61836014, No. 61773375, No. 62072457), and in part by the TuSimple Collaborative Research Project.

References

[1] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1900–1909, 2019.

[2] 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] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018.

[4] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6830–6840, 2019.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[7] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 982–991, 2019.

[8] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[9] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks*, pages 1322–1328, 2008.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[12] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[13] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *Advances in Neural Information Processing Systems*, 2020.

[14] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12975–12984, 2020.

[15] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.

[16] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6758–6767, 2019.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[19] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6778–6787, 2019.

[20] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[21] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.

[22] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 864–873, 2016.

[23] Junran Peng, Xingyuan Bu, Ming Sun, Zhaoxiang Zhang, Tieniu Tan, and Junjie Yan. Large-scale object detection in the wild from imbalanced multi-labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9709–9718, 2020.

[24] Minlong Peng, Qi Zhang, Xiaoyu Xing, Tao Gui, Xuanjing Huang, Yu-Gang Jiang, Keyu Ding, and Zhigang Chen. Trainable undersampling for class-imbalance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4707–4714, 2019.

[25] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision*, pages 102–118, 2016.

[26] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2020.

[27] 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.

[28] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision*, pages 467–482, 2016.

[29] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[30] M Naseer Subhani and Mohsen Ali. Learning from scale-invariant examples for domain adaptation in semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 290–306, 2020.

[31] 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.

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

[33] 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 Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[34] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12635–12644, 2020.

[35] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 518–534, 2018.

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

[37] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6810–6818, 2018.

[38] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 1076–1082, 2020.

[39] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021.

[40] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019.

[41] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision*, pages 289–305, 2018.