DAST: Unsupervised Domain Adaptation in Semantic Segmentation Based on Discriminator Attention and Self-Training

Fei Yu [1], Mo Zhang [1,2], Hexin Dong [1], Sheng Hu [1], Bin Dong [1,2,3], Li Zhang [1,2]

[1] Center for Data Science, Peking University, Beijing, China
[2] Center for Data Science in Health and Medicine, Peking University, Beijing, China
[3] Beijing International Center for Mathematical Research(BICMR), Peking University, Beijing, China
{yufei1900, zhangmo007, donghexin, hs95, zhangli_pku}@pku.edu.cn dongbin@math.pku.edu.cn

Abstract

Unsupervised domain adaption has recently been used to reduce the domain shift, which would ultimately improve the performance of semantic segmentation on unlabeled real-world data. In this paper, we follow the trend to propose a novel method to reduce the domain shift using strategies of discriminator attention and self-training. The discriminator attention strategy contains a two-stage adversarial learning process, which *explicitly* distinguishes the well-aligned (domain-invariant) and poorly-aligned (domain-specific) features, and then guides the model to focus on the latter. The self-training strategy adaptively improves the decision boundary of the model for target domain, which *implicitly* facilitates the extraction of domain-invariant features. By combining the two strategies, we find a more effective way to reduce the domain shift. Extensive experiments demonstrate the effectiveness of our proposed method on numerous benchmark datasets.

Introduction

Semantic segmentation is a classic computer vision task that aims to predict a semantic label for each pixel in an image. Despite the notable progress in this field driven by the rapid development of deep learning (Chen et al. 2018; Long, Shelhamer, and Darrell 2015; Ronneberger, Fischer, and Brox 2015; Wei et al. 2018), it remains challenging to apply segmentation model trained on the labeled source data to the unlabeled target/*real-world* data, which vary substantially in their illumination, style, and the context in different domains. One possible solution draws on supervised learning to retrain or fine-tune the pre-trained model, which however requires expensive and time-consuming pixel-level manual annotations. An alternative way is to use unsupervised domain adaptation (UDA) to reduce the domain shift, so as to train a model that is able to segment the target images without labels.

The key component of semantic segmentation using UDA methods is to align the features from different domains (Chen et al. 2019; Hoffman et al. 2018; Tsai et al. 2018; Vu et al. 2019). Although the main idea is straightforward — matching the overall feature-level distributions of the source and the target domains, the difficulty of implementation varies with adapting features for different regions in an image. For example, the adaption is easier in the case of sky than in the cases of buildings, traffic lights, and sidewalks, because the regions of the sky are similar regardless of the images while the latter group is characterized with different architectural styles or traffic rules. (Luo et al. 2019b) believes that aligning the source domain and the target domain globally leads to negative transfer of information and undermines the performance of the model in the originally well-aligned regions. Therefore, they propose to generate a local alignment score map and allow different weights for regions with different local alignment scores.

Following the spirit of (Luo et al. 2019b), in this work, we propose a strategy called discriminator attention (DA), to directly evaluate whether the local features are hard-adapted. The proposed DA strategy includes two stages of adversarial learning — discovering and correcting. In the discovering phase, a discriminator network (also known as the discoverer, D) aligns the intermediate features of the segmentation network and uses the confidence of local alignment to form an attention map that reweights the feature maps for label prediction. In the correcting stage, another discriminator network (called corrector, C) further aligns the output of the segmentation network based on the previous attention map. As illustrated in Figure 1 (b), the model pays more attention to hard-adapted regions for domain alignment.

Considering that the distribution of *real-world* data (target domain) is over-complex, we further introduce a self-training strategy to guarantee that the decision boundary of the model is suitable for the target domain. As shown in Figure 1 (c), the decision boundary of the segmentation network after UDA still tends to favor the distribution of source domain data, but the tendency is corrected after we apply the self-training strategy. Specifically, we adaptively improve the model's decision boundary by training the segmentation network with pseudo labels generated from the previous predictions.

In summary, we propose an effective and intuitive unsupervised domain adaptation method for semantic segmentation, combining the strategies of discriminator attention and self-training (DAST). The main contributions can be summarized as follows:

- We propose a novel two-stage adversarial learning (DA),

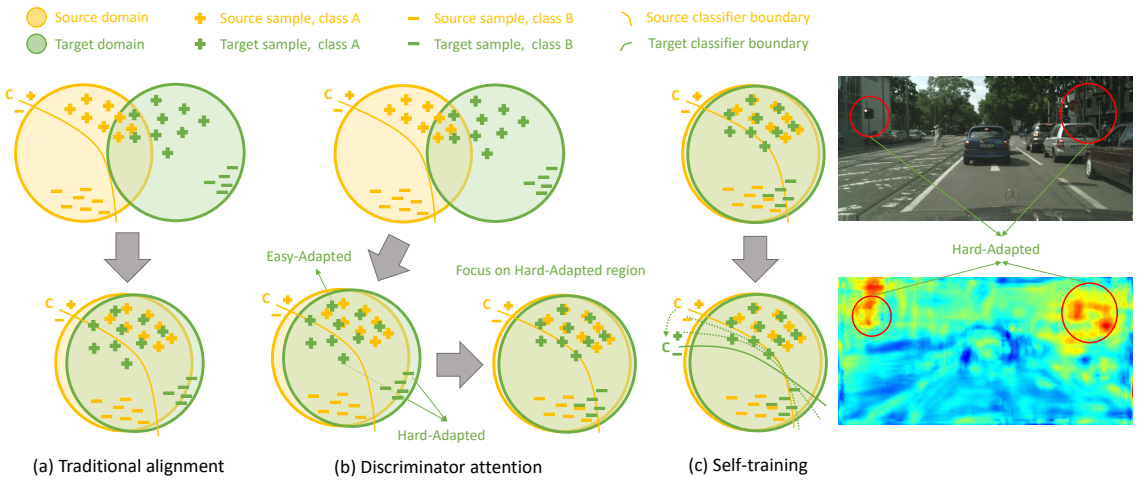

Figure 1: Illustration of the traditional adversarial learning methods and ours. (a) Traditional adversarial learning methods equally align the distribution among the entire images. (b) Our proposed discriminator attention can focus on the hard-adapted regions and achieve a better alignment. (c) We introduce a self-training module to learn an adaptive classifier boundary for the target domain, which can further improve the performance. (Best view in color)

which utilizes attention mechanism to attach higher weights to hard-adapted regions and simultaneously align the feature space and the output space.

- Our method is complementary to existing domain adaptation techniques, such as self-training.
- Our method achieves superior performance on the adaptation from SYNTHIA (Ros et al. 2016)/GTA5 (Richter et al. 2016) dataset to *real-world* dataset, Cityscapes (Cordts et al. 2016).

Related Works

In this section, we briefly summarize the methods related to the key ideas of the proposed DAST, including adversarial learning, discriminator confidence and self-training.

Adversarial Learning. Being the most explored approach in the field of unsupervised domain adaptation for semantic segmentation, adversarial learning mostly aligns information either on feature level or pixel level to reduce the domain shift[1].

For the feature-level alignment, Hoffman et al. (Hoffman et al. 2016) first apply adversarial learning to align the feature distributions between different domains to train a semantic segmentation model for the real-world images. Later on, Tsai et al. (Tsai et al. 2018) find that to align the output space distribution is more effective than to align the distribution of the intermediate feature space. Luo et al. (Luo et al. 2019b) utilize co-training to keep semantic consistency from multiple views of the features, which encourages the category-level alignment of different domains. Vu et al. (Vu et al. 2019) use adversarial learning to match the entropy of output predictions in source and target domains, which provides an alternative way of output space alignment. Tsai et al. (Tsai et al. 2019) construct different modes of images through patch-level clustering, and obtain a discriminator that pays more attention to high-level patterns, so as to optimize the domain alignment.

The pixel-level alignment is also known as image-to-image translation or style transfer. Benefiting from generative adversarial network (GAN), pixel-level alignment translates source images to the target domain, or vice versa, to facilitate learning a segmentation model across different domains, such as classic CycleGAN (Zhu et al. 2017). Recently, researchers have added the flavor of feature-level alignment to the pixel-level alignment, in order to achieve more accurate segmentation. Hoffman et al. (Hoffman et al. 2018) and Chen et al. (Chen et al. 2019) align the intermediate features to optimize image-to-image translation. Li et al. (Li, Yuan, and Vasconcelos 2019) construct a two-way learning process, which iteratively improves the segmentation and image-translation. Chang et al. (Chang et al. 2019) use pixel-level adversarial learning to disentangle the image features and train the segmentation model with content-only information. Choi et al. (Choi, Kim, and Kim 2019) adopt AdaIN (Huang and Belongie 2017) to embed the image information of unlabeled target images into the training process, making the model suitable to segment the target images.

Discriminator Confidence. Discriminator confidence is the output of the discriminator network in a fully convolutional manner. Hung et al. (Hung et al. 2019) use discriminator confidence to select the regions with small differences between the segmentation prediction and the label to form a pseudo label for the model training. (Kurmi, Kumar, and Namboodiri 2019; Wang et al. 2019) explore it in the intermediate layer as an attention map for domain adaptation in image classification. Inspired by these works, we implement the attention map in the proposed DA module using the dis-

[1]We take output-level alignment as a special case of feature-level alignment

criminator confidence scores from the intermediate features.

Self-Training. Self-training or self-distillation has shown impressive results in recent years (Dong et al. 2019; Zhai et al. 2019; Zhang et al. 2019). In the field of UDA for semantic segmentation, CNN based self-training methods mainly fine-tune a trained segmentation model using the target images and the pseudo labels, which implicitly forces the model to extract the domain-invariant features. Zou et al. (Zou et al. 2018) perform self-training by adjusting class weights to generate more accurate pseudo labels to train the segmentation model. French et al. (French, Mackiewicz, and Fisher 2018) adopt the mean-teacher framework, which introduces a consistency regularization to realize domain adaptation between the mean-teacher (target domain) and the student (source domain). In the proposed method, we find that self-training could be combined with the DA module to further improve the decision boundary of the segmentation model for unlabeled target images.

Method

In this section, we first provide an overview of our method. Then, we describe the overall objective function. Finally, we discuss the proposed discriminator attention in more detail.

Method Overview

In this work, we focus on the problem of unsupervised domain adaptation for semantic segmentation, where we have the access to the labeled source dataset $\{\mathbf{x}_s, \mathbf{y}_s\}$ and unlabeled target dataset $\{\mathbf{x}_t\}$. As shown in Figure 2, the overall network architecture is mainly composed of a segmentation network (segmentor S), and two discriminator networks (discoverer D and corrector C). The network backbone of the segmentor S can be any fully-convolutional network for semantic segmentation. For better description and discussion, S is divided into a feature extractor E and a label predictor P, where $S = E \circ P$. Discriminators (D and C) are CNN-based classifiers with a fully convolutional output, which could provide confidence scores for all output locations to evaluate the local alignment of the different domains.

In the source flow, E extracts a feature map $\mathbf{f}_s$ from a source domain image $\mathbf{x}_s$, where $\mathbf{f}_s = E(\mathbf{x}_s)$. The predictor P then takes $\mathbf{f}_s$ as an input to form a pixel-level semantic segmentation $\mathbf{p}_s$, where $\mathbf{p}_s = P(\mathbf{f}_s)$, which will be used to calculate a segmentation loss $\mathcal{L}_{seg}$ under the supervision of the source label $\mathbf{y}_s$. On the other hand, $\mathbf{f}_s$ and $\mathbf{p}_s$ will be input into the discoverer D and the corrector C for feature-level and output-level adversarial learning, respectively.

In the target flow, for a given image $\mathbf{x}_t$, E outputs a feature map $\mathbf{f}_t$ which is first input to the discoverer D. By optimizing the adversarial loss $\mathcal{L}^D_{adv}$, D aligns the feature distribution of $\mathbf{f}_t$ and $\mathbf{f}_s$ and provides a confidence score of alignment for each location in $\mathbf{f}_t$ to form an attention map α, where $\alpha = |(D(\mathbf{f}_t)|$. α reweights $\mathbf{f}_t$ into a new feature map $\hat{\mathbf{f}}_t = \alpha(\mathbf{f}_t)$, which is input to P to yield the pixel-level prediction $\mathbf{p}_t$ with more focus on poorly-aligned regions, where $\mathbf{p}_t = P(\hat{\mathbf{f}}_t)$. The corrector C is then introduced to perform an adversarial learning between $\mathbf{p}_t$ and $\mathbf{p}_s$. To further enhance the adaptation of poorly-aligned regions, we reweight the adversarial loss $\mathcal{L}^C_{adv}$ with the attention map α.

In addition, we apply a self-training strategy to improve the decision boundary of the segmentation model. Similar to (Li, Yuan, and Vasconcelos 2019), we introduce a super parameter q of the pixel portion. We generate the pseudo label $\hat{\mathbf{p}}_t$ using the top q of pixels in $\mathbf{p}_t$ with higher probability values and mask out other pixels which will not participate in gradient back-propagation.

The training process of the proposed method is summarized in **Algorithm 1**. In practice, we set the initial q to 50% and the maximum iteration K of self-training to 3 (the performance converges).

Algorithm 1 Training process of proposed method

Input:
The source domain sample, (x_s, y_s)
The target domain sample, x_t
The initial network, S(segmenter), D(discoverer), C(corrector)
Output:
The trained network, $S'_{K+1}, D'_{K+1}, C'_{K+1}$
1: train $S'_0 \leftarrow S, D'_0 \leftarrow D, C'_0 \leftarrow C$ with loss $\mathcal{L}_{seg}$ and $\mathcal{L}_{adv}$
2: **for** $k = 0$ to K **do**
3: input x_t into S'_k and generate pseudo label $p_t$with a fixed portion q_k
4: train $S'_{k+1} \leftarrow S'_k, D'_{k+1} \leftarrow D'_k, C'_{k+1} \leftarrow C'_k$ with loss $\mathcal{L}_{seg}$, $\mathcal{L}_{adv}$ and $\mathcal{L}_{p\_seg}$
5: **end for**
6: return $S'_{K+1}, D'_{K+1}, C'_{K+1}$

Objective Functions

The overall loss function mainly consists of four loss terms:

$$\mathcal{L}_{overall}(E,P,D,C) = L_{seg}(E,P) + \lambda_d \mathcal{L}^D_{adv}(E,D) + \lambda_c \mathcal{L}^C_{adv}(E,P,C) + \mu \mathcal{L}_{p\_seg}(E,P).$$

The first term $\mathcal{L}_{seg}(E,P)$ guides the segmenter S ($S = P \circ E$) to perform a dense prediction of the segmentation in the source domain,

$$\min_{E,P} \mathcal{L}_{seg}(E,P) = \min_{E,P} \mathbb{E}_{\mathbf{x}_s,\mathbf{y}_s \sim p(\mathbf{X}_s,\mathbf{Y}_s)}(\ell_{ce}(P \circ E(\mathbf{x}_s), \mathbf{y}_s))$$

where $\ell_{ce}(\cdot,\cdot)$ indicates the multi-class cross entropy loss used in this work.

The second and third terms are the adversarial losses of the discoverer D and the corrector C, respectively. Following LSGAN (Mao et al. 2017), we use the least square loss to replace the sigmoid cross entropy in the vanilla GAN, because the sigmoid-based loss usually stops updating when the discriminator reaches the optimum (Hong et al. 2019). $\mathcal{L}^D_{adv}(E,D)$ and $\mathcal{L}^C_{adv}(E,P,C)$ correspond to the two-stage adversarial learning in our discriminator attention module.

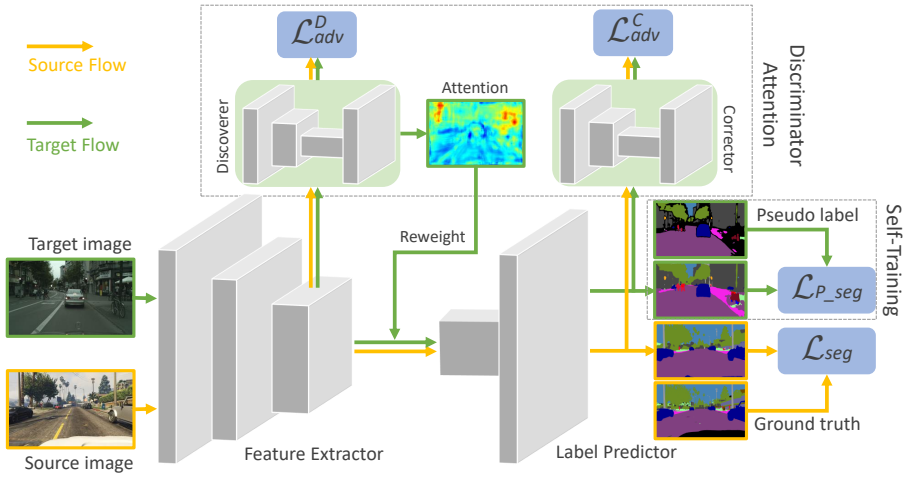

Figure 2: Overview of our proposed method. The randomly selected images from the source and target domain are used to train a cross-domain segmentation network by adversarial training. Two fully convolutional discriminator network named discover and corrector are used to obtain a better alignment. The discoverer could align the intermediate features of the segmentation network and form an attention map using the confidence of local alignment. The corrector can focus on the hard-adapted regions and optimize the overall adversarial learning based on the previous attention map. In addition, we utilize self-training to train an adaptive classifier boundary for the target domain. (Best view in color)

In the first stage, domain-invariant features extracted by E are expected to confuse the discoverer D, which aims to minimize the loss $\mathcal{L}_{adv}^{D}(E, D)$ by alternatively optimizing D and E,

$$\min_{D} \mathcal{L}_{adv}^{D}(D) = \min_{D} \mathbb{E}_{\mathbf{x}_s \sim p(\mathbf{X}_s)}[(D(\mathbf{f}_s) - 0)^2] + \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{X}_t)}[(D(\mathbf{f}_t) - 1)^2]$$

$$\min_{E} L_{adv}^{D}(E) = \min_{E} \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{X}_t)}[(D(E(\mathbf{x}_t)) - 0)^2]$$

After D is optimized, for a given target image $\mathbf{x}_t$, an attention map is generated to distinguish the easy-adapted and hard-adapted regions, $\alpha = |D(\mathbf{f}_t)|$.

In the second stage, we expect that $P \circ E$ outputs segmentation predictions that are able to confuse C.

$$\min_{C} \mathcal{L}_{adv}^{C}(C) = \min_{C} \mathbb{E}_{\mathbf{x}_s \sim p(\mathbf{X}_s)}[(C(\mathbf{p}_s) - 0)^2] + \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{X}_t)}[(C(\mathbf{p}_t) - 1)^2]$$

$$\min_{E,P} L_{adv}^{C}(E, P) = \min_{E,P} \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{X}_t)}[(C(P \circ E(\mathbf{x}_t)) - 0)^2]$$

The fourth loss term is related to self-training strategy, which adaptively improves the decision boundary of the segmenter S ($S = P \circ E$) to fit the target distribution,

$$\min_{E,P} \mathcal{L}_{p\_seg}(E, P) = \min_{E,P} \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{X}_t)}(\ell_{ce}(\mathbf{p}_t, \hat{\mathbf{p}}_t)).$$

In the overall loss, $\lambda_d, \lambda_c, \mu$ are the hyper-parameters used to balance the relative importance of different terms. During training, we set $\lambda_d = 0.01, \lambda_c = 0.01, \mu = 1$.

Design of Attention Mechanism

For the target image feature $\mathbf{f}_t$, the confidence scores of the discoverer $\alpha = |D(\mathbf{f}_t)|$ show whether $\mathbf{f}_t$ locally matches the distribution of $\mathbf{f}_s$. A low α_{ij} represents a well-aligned region in $\mathbf{x}_t$ and a high α_{ij} represents a poorly-aligned region. Therefore, we use α as an attention map of $\mathbf{f}_t$ to encourage the model to focus on matching features of those poorly-aligned regions. Moreover, to prevent the gradient explosion at the early stage of the experiment, we add a $tanh$ activation to α as a normalization layer. Finally, we expand $tanh(\alpha)$ to fit the dimension of $\mathbf{f}_t$ for the subsequent element-wise multiplication,

$$\alpha' = expand(tanh(\alpha))$$

Since the magnitude of $tanh(\alpha)$ is less than 1, its gradient may disappear in the late stage of the training process. We thus adopt the residual attention mechanism (Wang et al. 2017) to calculate the new feature map,

$$\mathbf{f}_t' = \mathbf{f}_t + \mathbf{f}_t \odot \alpha'$$

Experiments and Results

In this section, we will present our experiments and results. We first describe the benchmark datasets and experimental setups. Then, we report our main results and compare them with the state-of-the-art methods on the benchmark datasets.

Datasets

We evaluate the proposed DAST method on the challenging *synthetic-2-real* setups: SYNTHIA (Ros et al. 2016) and GTA5 (Richter et al. 2016) datasets are used as the source domain dataset and Cityscapes (Cordts et al. 2016) is used

Methods	Arch.	Mech.	road	sidewalk	building	wall	fence	pole	light	sign	vege.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Source only	V	-	60.7	13.7	56.9	12.9	20.1	19.0	15.4	6.5	77.7	16.2	56.8	40.0	3.3	63.6	15.3	9.5	0.0	8.1	0.1	26.1
CLAN	V	A	88.0	30.6	79.2	23.4	20.5	26.1	23.0	14.8	81.6	34.5	72.0	45.8	7.9	80.5	26.6	29.9	0.0	10.7	0.0	36.6
BDL	V	RSA	89.2	40.9	81.2	29.1	19.2	14.2	29.0	19.6	83.7	35.9	80.7	54.7	23.3	82.7	25.8	28.0	2.3	25.7	19.9	41.3
FDA-MBT	V	RSA	86.1	35.1	80.6	30.8	20.4	27.5	30.0	26.0	82.1	30.3	73.6	52.5	21.7	81.7	24.0	30.5	29.9	14.6	24.0	42.2
Baseline	V	A	87.3	29.8	78.6	21.1	18.2	22.5	21.5	11.0	79.7	29.6	71.3	46.8	6.5	80.1	23.0	26.9	0.0	10.6	0.3	35.0
Ours(DA)	V	A	89.3	40.4	79.2	34.9	22.8	23.1	24.0	16.8	79.9	28.7	67.9	45.1	17.8	82.1	25.7	31.9	4.1	19.7	3.1	38.8
Ours(DAST)	V	SA	90.5	49.2	81.9	34.0	27.0	26.5	26.6	21.5	83.0	37.3	76.3	52.0	23.1	83.5	29.9	42.0	12.1	19.8	25.8	44.3
Source only	R	-	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
CLAN	R	A	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
SIBAN	R	A	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
BDL	R	RSA	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
DPR	R	RSA	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
UIDA	R	SA	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3
DTST	R	RSA	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
FDA-MBT	R	RSA	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
Baseline	R	A	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
Ours(DA)	R	A	92.3	54.2	81.9	27.3	25.3	33.4	39.1	23.2	84.0	34.2	71.1	58.7	29.7	85.2	28.1	34.7	4.8	25.6	19.6	44.8
Ours(DAST)	R	SA	92.2	49.0	84.3	36.5	28.9	33.9	38.8	28.4	84.9	41.6	83.2	60.0	28.7	87.2	45.0	45.3	7.4	33.8	32.8	49.6

Table 1: Experimental results for GTA5 $\rightarrow$ Cityscapes. "Source only" denotes the model only trained on source data without adaptation. The architecture "V" and "R" represent the VGG-16 and ResNet-101 backbones, respectively. The mechanism "R", "S", and "A" means image-to-image translation, self-training, and adversarial training, respectively. Our baseline model is AdaptSegNet (Tsai et al. 2018). Other previous state-of-the-art methods include CLAN (Luo et al. 2019b), SIBAN (Luo et al. 2019a), BDL (Li, Yuan, and Vasconcelos 2019), DPR (Tsai et al. 2019), UIDA (Pan et al. 2020), DTST (Wang et al. 2020), FDA-MBT (Yang and Soatto 2020).

as the target domain dataset. The details of these datasets are described as follows: **1) GTA5 dataset.** GTA5 dataset consists of 24966 labeled urban scene images with a resolution of 1914×1052 collected from the video game Grand Theft Auto V. **2) SYNTHIA dataset.** SYNTHIA is another more challenging synthetic image dataset. For this dataset, we only use the SYNTHIA-RAND-CITYSCAPES subset, which has 16 common categories with Cityscapes. It contains 9400 labeled urban scene images with a resolution of 1280×760. **3) Cityscapes dataset.** Cityscapes consists of 2975 real-world images in the training set and 500 in the validation set with a resolution of 2048×1024. In all the experiments of this work, we use the 2975 images in the Cityscapes training set as the unlabeled target images and test the model with mean Intersection-over-Union (mIoU) on the 500 validation images. We only used the labels of the Cityscapes images to evaluate the segmentation performance and not for the training process.

Network Architecture

Inspired by (Choi, Kim, and Kim 2019; Tsai et al. 2018), we adopt the DeepLab (Chen et al. 2017) framework with VGG-16 (Simonyan and Zisserman 2014) and ResNet-101(He et al. 2016) backbone as our segmentation network. The initial weight is pretrained on ImageNet (Deng et al. 2009). After the last convolutional layer, the Atrous Spatial Pyramid Pooling (ASPP) module is applied with the sampling rates of $\{6, 12, 18, 24\}$. Finally, we utilize an upsampling layer to rescale the final segmentation output to match the dimension of the input image.

The discriminators (the discoverer D and the corrector C) are fully convolutional networks that retain the spatial information. Furthermore, D consists of 4 convolutional layers with channel numbers of $\{256, 128, 64, 1\}$ and the values of kernel size, padding size, and stride are 3, 1, and 1, respectively. C aligns the semantic predictions of different domains. Following (Tsai et al. 2018), it consists of 5 convolutional layers with kernel size, padding size, and stride of 4, 1, 2, respectively, and its channel numbers are $\{64, 128, 256, 512, 1\}$. Instead of regular ReLU, C uses Leaky ReLU as the activation with a fixed negative slope of 0.2.

Implementation Details

The model is implemented using the PyTorch toolbox and runs on a single Titan V GPU with 12 GB memory[2]. We introduce different settings for the segmentation network and the discriminators. 1) For the segmentation network, we use the stochastic gradient descent (SGD) algorithm as the optimizer. The initial learning rate is set as 2.5×10^{-4}, momentum is 0.9 and weight decay is 5×10^{-4}. 2) For the discriminators, we use the Adam algorithm as the optimizer. The initial learning rate is set as 10^{-4} and $\beta_1 = 0.9, \beta_2 = 0.99$. We also adopt the same polynomial decay with a power of 0.9 to update learning rate as mentioned in (Tsai et al. 2018).

We train the discriminator attention module only for 150k iterations and after that, we add self-training to fine-tune the model with pseudo labels for an additional 20k in several

[2]Code: https://github.com/yufei1900/DAST_segmentation

Methods	Arch.	Mech.	road	sidewalk	building	wall	fence	pole	light	sign	vege.	sky	person	rider	car	bus	mbike	bike	mIoU	mIoU*
Source Only	V	-	6.4	16.1	47.5	6.2	0.2	18.9	0.3	8.1	68.6	75.2	46.6	7.1	57.2	12.9	2.9	7.1	23.8	27.4
CLAN	V	A	80.4	30.7	74.7	-	-	-	1.4	8.0	77.1	79.0	46.5	8.9	73.8	18.2	2.2	9.9	-	39.3
BDL	V	RSA	72.0	30.3	74.5	0.1	0.3	24.6	10.2	25.2	80.5	80.0	54.7	23.2	72.7	24.0	7.5	44.9	39.0	46.1
FDA-MBT	V	RSA	84.2	35.1	78.0	6.1	0.44	27.0	8.5	22.1	77.2	79.6	55.5	19.9	74.8	24.9	14.3	40.7	40.5	47.3
Baseline	V	A	78.9	29.2	75.5	-	-	-	0.1	4.8	72.6	76.7	43.4	8.8	71.1	16.0	3.6	8.4	-	37.6
Ours(DA)	V	A	82.5	33.3	76.8	2.1	0.9	20.2	1.9	8.1	76.4	77.9	42.9	13.6	69.9	17.5	7.7	15.1	34.2	40.3
Ours(DAST)	V	SA	86.1	35.7	79.9	5.2	0.8	23.1	0.0	6.9	80.9	82.5	50.6	19.8	79.7	21.9	21.3	38.8	39.6	46.5
Source Only	R	-	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	38.6
CLAN	R	A	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
SIBAN	R	A	82.5	24.0	79.4	-	-	-	16.5	12.7	79.2	82.8	58.3	18.0	79.3	25.3	17.6	25.9	-	46.3
BDL	R	RSA	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
DPR	R	RSA	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
UIDA	R	SA	84.3	37.7	79.5	5.3	0.4	24.9	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	41.7	48.9
DTST	R	RSA	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-MBT	R	RSA	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
Baseline	R	A	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
Ours(DA)	R	A	83.2	40.6	80.3	10.2	0.3	27.5	7.9	11.2	79.4	84.6	54.1	20.9	73.4	33.2	18.1	27.3	40.8	47.2
Ours(DAST)	R	SA	87.1	44.5	82.3	10.7	0.8	29.9	13.9	13.1	81.6	86.0	60.3	25.1	83.1	40.1	24.4	40.5	45.2	52.5

Table 2: Experimental results for SYNTHIA $\rightarrow$ Cityscapes. The mIoU* denotes the mean IoU of 13 common classes. "Source only" denotes the model only trained on source data without adaptation. The architecture "V" and "R" represent the VGG-16 and ResNet-101 backbones, respectively. The mechanism "R", "S", and "A" means image-to-image translation, self-training, and adversarial training, respectively. Our baseline model is AdaptSegNet (Tsai et al. 2018).

Methods	mIoU(VGG16)	mIoU(Res101)
Baseline	35.0	41.4
+MSE	35.9	43.4
+MSE+2D	37.2	43.9
+MSE+2D+Atten.	38.8	44.8

Table 3: The ablation study results of discriminator attention adapted from GTA5 to Cityscapes. "MSE" denotes the mean square error loss function. "2D" denotes the feature distribution is aligned by 2 discriminators without attention.

rounds until the performance converges. During training, the images from GTA5 are resized to 1280×720 resolution, the images from SYNTHIA are resized to 1280×760 resolution, and the images from Cityscapes are resized to 1024×512. During validation, we upsample the segmentation predictions to 2048×1024 to calculate evaluation metrics. We train our model without any data augmentation steps.

Experimental Results

The previous methods of UDA for semantic segmentation can be roughly divided into image-to-image translation (R), self-training (S), adversarial training (A), and their combinations. The experimental results compared with these methods are shown in Table 1 (GTA5 to Cityscapes) and Table 2 (SYNTHIA to Cityscapes). The proposed DAST achieves the superior performance on two benchmark datasets.

GTA5 $\rightarrow$ Cityscapes. As shown in Table 1, our proposed method with DA only significantly outperforms the baseline by 3.8% and 3.4% in the mean IoU for two architectures and exceed all the other models using a single strategy. Although CLAN also aims to tackle the equally global alignment problem, our method is more effective and outperforms them by 2.2% and 1.6% for two architectures. Compared with the composite methods, our DAST also achieves the state-of-the-art performance.

SYNTHIA $\rightarrow$ Cityscapes. In the SYNTHIA dataset, the spatial layout or local context differs substantially from that in Cityscapes dataset. Despite the large domain shift between SYNTHIA and Cityscapes, our proposed method with DA only outperforms other adversarial learning-based methods and brings 2.7% and 1.3% improvement compared to the baseline over the 13 common classes for two architectures. Among the composite methods, our DAST also achieves the state-of-the-art performance.

Analysis

Ablation Studies

To verify the effectiveness of each part in discriminator attention, we perform 3 ablation tests. As shown in Table 3, the MSE loss function brings 0.9% and 2.0% improvement compared to the baseline. The combination of 2 discriminators brings another 1.3% and 0.5% improvement. The introduction of attention contributes 1.6% and 0.9% mIoU gain. Our final model exceeds the baseline for most of the categories in terms of segmentation accuracy.

Hyper-parameters Analysis

As shown in Table 4, when $q = 50\%$, the model achieves 49.6 of mIoU as the best performance on Cityscapes validation set.

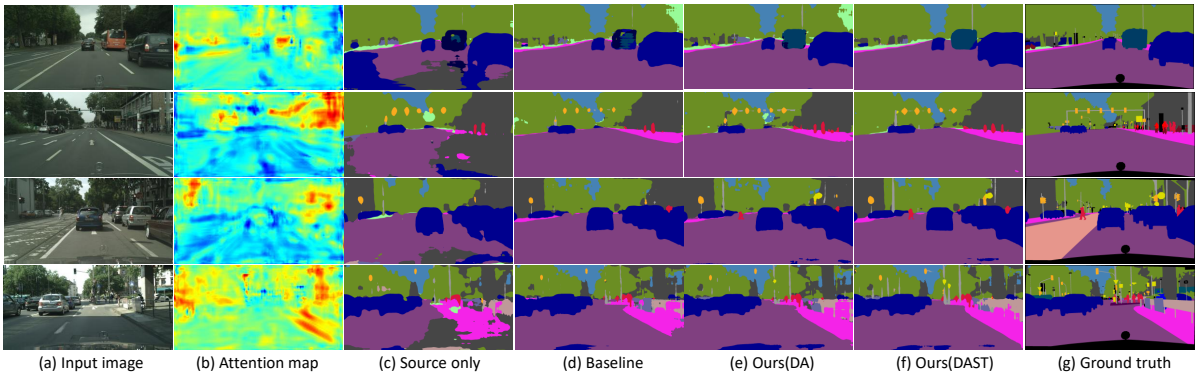

Figure 3: Visual results of segmentation map and attention map on VGG-16 backbone. (a) Input images. (b) Attention map generated by our discriminator attention module. (c) Segmentation map predicted by source only model. (d) Segmentation map predicted by baseline model. (e) Segmentation map predicted by our DA model. (f) Segmentation map predicted by our DAST model. (g) Ground truth.

q	40%	50%	60%
mIoU	49.0	49.6	48.7

Table 4: The results of hyper-parameter q from 40% to 60% adapted from GTA5 to Cityscapes on ResNet-101 backbone.

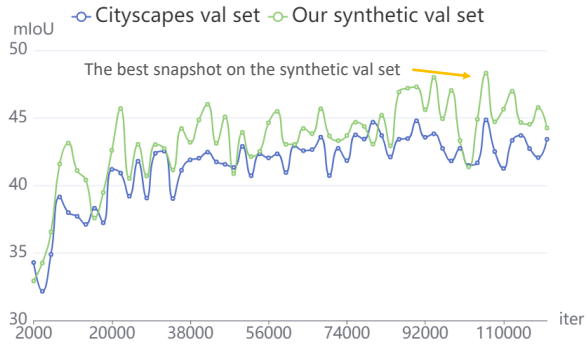

Figure 4: mIoU comparison on the Cityscapes val set and our synthetic val set by adapting from GTA5 to Cityscapes.

When to Start the Self-Training

Most previous studies suffer from biased models and overly-optimistic estimates, because they often select the best result from all evaluations of the intermediate snapshots on the Cityscapes val set. To address the overoptimistic issue, we establish a synthetic val set which consists of 500 randomly selected GTA5 images transferred to Cityscapes style by the method of (Yang and Soatto 2020). The performance of the model on the synthetic val set can guide the choice of the best snapshot and when to start the self-training. We use the mean of the two styles ($\beta = 0.5$ and $\beta = 0.9$) as the final results of the synthetic val set. As shown in Figure 4, the synthetic val set can approximately fit the Cityscapes val set.

Visualization

We use four qualitative examples in Figure 3 to illustrate that the attention map serves as a good indicator of hard-adapted regions. For example, the hard-adapted regions —the“bus”, “sidewalk” and “traffic light” that are wrongly predicted in the source-only models (col (c))— are assigned with higher confidence values in the attention map. Whereas the easy-adapted regions—the ”road”, ”sky” and ”car” regions that are correctly predicted in the source-only models—are characterized with lower confidence values in the attention map. This consistency proves that the attention map successfully distinguishes hard-adapted regions and easy-adapted regions.

Figure 3 also reveals that the attention mechanism can improve the segmentation results. Cols (d) and (e) demonstrate the segmentation map predicted by the baseline model and our model that introduces the attention mechanism, respectively. It is obvious that our results improved by the attention mechanism are visually closer to the ground truth (col (g)) than the baseline's. Cols (f) demonstrates that self-training could be combined with the DA module to further improve the performance.

Conclusions

In this paper, we have proposed a novel method that combines the discriminator attention and the self-training to realize the unsupervised domain adaptation for semantic segmentation. The discriminator attention module includes two stages of adversarial learning, which utilize the attention map to attach higher weights to hard-adapted regions and performs the feature-level and the output-level alignments between different domains. The self-training module dynamically generates pseudo labels to adapt the decision boundary of the segmentation network to fit the distribution of unlabeled target images. The experimental results and the qualitative examples prove that our method outperforms the previous state-of-the-art methods on the benchmark datasets.

Acknowledgments

This work was supported in part by National Key R&D Program of China (No. 2018YFC0910700); National Natural Science Foundation of China (NSFC) grant No. 11831002, 81801778, 71704024; Beijing Natural Science Foundation (No. 180001) and Beijing Academy of Artificial Intelligence (BAAI).

References

Chang, W.-L.; Wang, H.-P.; Peng, W.-H.; and Chiu, W.-C. 2019. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1900–1909.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. 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.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, Y.-C.; Lin, Y.-Y.; Yang, M.-H.; and Huang, J.-B. 2019. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1791–1800.

Choi, J.; Kim, T.; and Kim, C. 2019. Self-Ensembling with GAN-based Data Augmentation for Domain Adaptation in Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 6830–6840.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dong, B.; Hou, J.; Lu, Y.; and Zhang, Z. 2019. Distillation $\approx$ Early Stopping? Harvesting Dark Knowledge Utilizing Anisotropic Information Retrieval For Overparameterized Neural Network. *arXiv preprint arXiv:1910.01255* .

French, G.; Mackiewicz, M.; and Fisher, M. 2018. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 6.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *International Conference on Machine Learning*, 1989–1998.

Hoffman, J.; Wang, D.; Yu, F.; and Darrell, T. 2016. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649* .

Hong, Y.; Hwang, U.; Yoo, J.; and Yoon, S. 2019. How generative adversarial networks and their variants work: An overview. *ACM Computing Surveys (CSUR)* 52(1): 1–43.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Hung, W. C.; Tsai, Y. H.; Liou, Y. T.; Lin, Y. Y.; and Yang, M. H. 2019. Adversarial learning for semi-supervised semantic segmentation. In *29th British Machine Vision Conference, BMVC 2018*.

Kurmi, V. K.; Kumar, S.; and Namboodiri, V. P. 2019. Attending to discriminative certainty for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 491–500.

Li, Y.; Yuan, L.; and Vasconcelos, N. 2019. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6936–6945.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Luo, Y.; Liu, P.; Guan, T.; Yu, J.; and Yang, Y. 2019a. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 6778–6787.

Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; and Yang, Y. 2019b. 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*, 2507–2516.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

Pan, F.; Shin, I.; Rameau, F.; Lee, S.; and Kweon, I. S. 2020. Unsupervised Intra-domain Adaptation for Semantic Segmentation through Self-Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3764–3773.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *European conference on computer vision*, 102–118. Springer.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. 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*, 3234–3243.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; and Chandraker, M. 2018. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7472–7481.

Tsai, Y.-H.; Sohn, K.; Schulter, S.; and Chandraker, M. 2019. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE International Conference on Computer Vision*, 1456–1465.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2517–2526.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Wang, X.; Li, L.; Ye, W.; Long, M.; and Wang, J. 2019. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5345–5352.

Wang, Z.; Yu, M.; Wei, Y.; Feris, R.; Xiong, J.; Hwu, W.-m.; Huang, T. S.; and Shi, H. 2020. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12635–12644.

Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; and Huang, T. S. 2018. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7268–7277.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4085–4095.

Zhai, X.; Oliver, A.; Kolesnikov, A.; and Beyer, L. 2019. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, 1476–1485.

Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, 3713–3722.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zou, Y.; Yu, Z.; Vijaya Kumar, B.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, 289–305.