Consistency Regularization with High-dimensional Non-adversarial Source-guided Perturbation for Unsupervised Domain Adaptation in Segmentation

Kaihong Wang, Chenhongyi Yang, Margrit Betke
Boston University
{kaiwkh, hongyi, betke}@bu.edu

Abstract

Unsupervised domain adaptation for semantic segmentation has been intensively studied due to the low cost of the pixel-level annotation for synthetic data. The most common approaches try to generate images or features mimicking the distribution in the target domain while preserving the semantic contents in the source domain so that a model can be trained with annotations from the latter. However, such methods highly rely on an image translator or feature extractor trained in an elaborated mechanism including adversarial training, which brings in extra complexity and instability in the adaptation process. Furthermore, these methods mainly focus on taking advantage of the labeled source dataset, leaving the unlabeled target dataset not fully utilized. In this paper, we propose a bidirectional style-induced domain adaptation method, called BiSIDA, that employs consistency regularization to efficiently exploit information from the unlabeled target domain dataset, requiring only a simple neural style transfer model. BiSIDA aligns domains by not only transferring source images into the style of target images but also transferring target images into the style of source images to perform high-dimensional perturbation on the unlabeled target images, which is crucial to the success in applying consistency regularization in segmentation tasks. Extensive experiments show that our BiSIDA achieves new state-of-the-art on two commonly-used synthetic-to-real domain adaptation benchmarks: GTA5-to-CityScapes and SYNTHIA-to-CityScapes. Code and pretrained style transfer model are available at: https://github.com/wangkaihong/BiSIDA.

Introduction

Deep learning methods for semantic segmentation (Long, Shelhamer, and Darrell 2015), the problem of dividing the pixels in an image into mutually exclusive and collectively exhaustive sets of class-labeled regions, have gained increasing attention. Research progress is hindered by the difficulty of creating large training datasets with accurate pixel-level annotations of these regions.

As a consequence, the use of synthetic datasets has become popular because pixel-level ground truth annotations can be generated along with the images. Unfortunately, when deep models that were trained on synthetic data are used to segment real-world images, their performance is typically limited due to the domain gap between the training and testing data. Domain adaptation methods seek to bridge the gap between the source domain training data and the target domain testing data. We here focus on unsupervised domain adaptation (UDA), the problem of adapting a model that was trained with a labeled source domain dataset, by using an unlabeled target domain dataset and optimizing its performance on the target domain.

To perform domain alignment on a pixel-level or feature-level basis, existing methods (Tsai et al. 2018; Hoffman et al. 2018; Vu et al. 2019; Luo et al. 2019; Li, Yuan, and Vasconcelos 2019; Choi, Kim, and Kim 2019) typically use adversarial training (Goodfellow et al. 2014), and training with the aligned data is then supervised by a loss computed with the annotation of the source domain dataset. However, the use of adversarial training typically comes with extra complexity and instability in training. Alternative approaches (Zou et al. 2018; Vu et al. 2019; Li, Yuan, and Vasconcelos 2019; Choi, Kim, and Kim 2019) seek to exploit information about the unlabeled target dataset by performing semi-supervised learning including entropy minimization (Grandvalet and Bengio 2004), pseudo-labeling (Lee 2013) and consistency regularization. However, these approaches either just play an auxiliary role in the training process besides supervised learning, or fail to take full advantage of the target dataset.

In this paper, we propose Bidirectional Style-induced Domain Adaptation (BiSIDA) that takes better advantage of the unlabeled dataset and optimizes the performance of a segmentation model on the target dataset. Our pipeline includes a supervised learning branch that provides supervision using annotations in the source dataset and an unsupervised branch for learning from the unlabeled target dataset without requiring its annotation. The two branches are executed in parallel. To perform domain adaptation, we constructed a non-adversarial yet effective pre-trained style-induced image generator that performs style transfer. In the supervised learning branch, the style-induced image generator transfers the style of a source domain images to become the style of a target style image. In the unsupervised branch, the image generator performs high-dimensional perturbations on target domain images with consistency regularization by transferring the target input image into a set of images that each match one of the source image styles through a self-supervised approach. Consequently, the unlabeled target dataset is utilized efficiently and the domain gap is reduced effectively.

BiSIDA performs image transfer from the source to the target domain using an image generator in the supervised phase similar to existing methods. However, to facilitate generalization, our model synthesizes images with semantic content from the source domain, and with a style that is defined by a continuous parameter that represents a "mix" of source and target domain styles, instead of transferring the style directly to the target domain. Consequently, the stochasticity of the whole process facilitates not only the training on the original images but also the gradual adaptation towards the target domain. The resulting image is then sent along with its corresponding pixel-level annotation to compute a supervised cross-entropy loss to train the segmentation model.

BiSIDA employs consistency regularization in the unsupervised phase to yield consistent predictions on randomly perturbed inputs without requiring their annotations. We apply our style-induced image generator as an augmentation method and transfer each target domain image together with a number of randomly sampled source domain images, just as in the supervised phase, but in an opposite direction. A series of images with identical content but different styles from source domain images is generated. Given that supervised learning is performed on source images that are transferred with combined styles of source images and target images, our model will be more adapted and more likely to produce correct predictions when target domain images are transferred towards the direction of the source domain images. Meanwhile, our image generator provides a high-dimensional perturbation that keeps the semantic content as indicated by (French et al. 2019) for consistency regularization in a computational affordable way. To further improve the quality of predictions, BiSIDA passes the transferred images through the self-ensemble of the trained segmentation models and obtains a pseudo-label for the unlabeled target domain image. The training of the segmentation model on the original target domain image, augmented with only brightness and contrast perturbations, is guided by its pseudo-label. Information within the unlabeled target images is learned through consistency regularization, and the model adapts to the target domain. Our method utilizes annotations from the labeled source dataset, exploits knowledge from the unlabeled target dataset, and performs gradual adaptation between the source and the target domain from both sides. In conclusion, our key contributions include:

1. A Bidirectional Style-induced Domain Adaptation (BiSIDA) framework that incorporates both target-guided supervised and source-guided unsupervised learning. We also show that domain adaptation is achievable in a bidirectional way through a continuous parameterization of the two domains, without requiring adversarial training;

2. A non-adversarial continuous style-induced image generator (CSIIG) that performs a high-dimensional source-guided perturbation on target images for consistency regularization.

3. Extensive experiments show that our BiSIDA achieves new state-of-the-art on two commonly-used synthetic-to-real domain adaptation benchmarks: GTA5-to-CityScapes and SYNTHIA-to-CityScapes.

Related Works

Image-to-image Translation. Recent progress in image-to-image translation that transfers the style of an image while preserving its semantic content has inspired research in various related areas, including image synthesis and reducing domain discrepancy. Typical image-to-image translation approaches include CycleGAN (Zhu et al. 2017) and DualGAN (Yi et al. 2017), which keep cycle-consistency in adversarial training to preserve the semantic content of images when transferring the style of image. UNIT (Liu, Breuel, and Kautz 2017) and MUNIT (Huang et al. 2018) address the problem by mapping images into a common latent content space. *Neural style transfer* offers an alternative way to perform image-to-image translation (Gatys, Ecker, and Bethge 2016), but its optimization process is computationally impractical. Several works (Johnson, Alahi, and Fei-Fei 2016; Li and Wand 2016; Ulyanov et al. 2016; Ulyanov, Vedaldi, and Lempitsky 2017; Dumoulin, Shlens, and Kudlur 2017) proposed improvements, but these methods are limited since the style to be transferred is either fixed or the number of styles is limited. BiSIDA uses a technique from image-to-image translation, 'Adaptive Instance Normalization,' or AdaIN (Huang and Belongie 2017), in which an encoder extracts a feature map from a given input image and a decoder upsamples the feature map back to the original size of the input.

Semi-supervised Learning. When the gap between source and target domains becomes small, the problem of unsupervised domain adaptation intriguingly degenerates to a the problem of semi-supervised learning. Pseudo-labeling (Lee 2013), a commonly-used semi-supervised learning method, takes high-confidence predictions on the unlabeled dataset as one-hot labels, guiding further training. Entropy minimization (Grandvalet and Bengio 2004) can be seen as a "soft assignment" of the pseudo-label on the unlabeled dataset. Recently, consistency regularization has gained attention due to its outstanding performance as a semi-supervised learning method. The Mean-Teacher (Tarvainen and Valpola 2017) approach minimizes consistency loss on an unlabeled image between the output of a student network and the ensemble of itself, a teacher network. Fixmatch (Sohn et al. 2020) further outperforms Mean-Teacher by performing pseudo-labeling and consistency regularization between images with different degree of perturbations and achieves state-of-the-art performance on several semi-supervised learning benchmarks.

UDA for Semantic Segmentation. Current methods in UDA for segmentation can be categorized into adversarial and non-adversarial methods. "FCN in the wild" (Hoffman et al. 2016) was the first to perform a segmentation task under UDA settings and align both global and local features between domains through adversarial training. Other works (Hoffman et al. 2018; Tsai et al. 2018; Vu et al. 2019) tried to align features in one or multiple feature levels. The

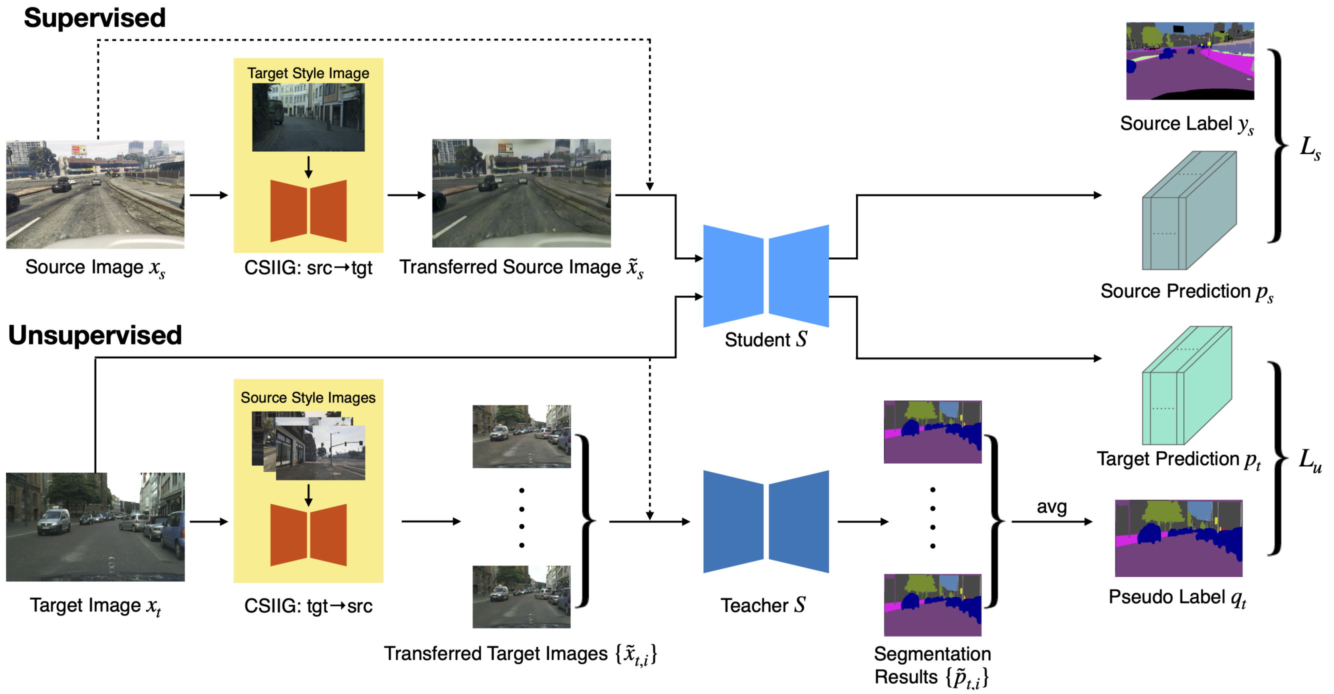

Figure 1: The proposed training framework BiSIDA: It consists of two branches that are executed in parallel: 1) The supervised branch (top) augments a source-domain image x_s through our continuous style-induced image generator (CSIIG) with the style of a target-domain image. A supervised segmentation loss L_s is computed with the corresponding annotation of the source image. 2) In the unsupervised learning branch (bottom), a target domain image and a series of source domain images are used to produce the corresponding transferred images $\tilde{x}_{t,i}$. Then each of these images passes through the teacher network to generate a set of probability maps $\tilde{p}_{t,i}$. BiSIDA computes an average of these maps to generate a pseudo-label q_t, which is then used to compute the loss L_u for consistency regularization. In both branches, BiSIDA allows for the option to skip the image generator for some of the input images (dotted line).

adversarial alignment process of each category between domains can be treated adaptively (Luo et al. 2019; Wang et al. 2020). Choi, Kim, and Kim, 2019, trained an image translator in an adversarial way and took its output to perform consistency regularization. Li, Yuan, and Vasconcelos, 2019, applied bidirectional learning in which an image translator and a segmentation model guide each other's training in a mutual way. Pseudo-labeling is also performed to enhance performance.

Non-adversarial methods include a variety of techniques. Curriculum DA (Zhang, David, and Gong 2017) and PyCDA (Lian et al. 2019), for example, adopt the concept of curriculum learning and align label distributions over images, landmark superpixels, or regions. CBST (Zou et al. 2018) utilizes self-training to exploit information from the target domain images. DCAN (Wu et al. 2018) applies channel-wise alignment to merge the domain gap from both the pixel level and the feature level. Recently, Yang and Soatto, 2020, proposed to align pixel-level discrepancy by performing a Fourier transformation. Combined with entropy minimization, pseudo-labeling and model ensemble, their method achieves current state-of-the-art performance.

The work that maybe most resembles ours is by Choi, Kim, and Kim, 2019, However, our method does not rely on a strong image translator that needs to be trained in an adversarial way as theirs. Furthermore, our method of adopting consistency regularization is able to exploit information more efficiently and effectively from target images by virtue of our high-dimensional perturbation method.

Method

In the UDA setting, the dataset from the source domain $\mathcal{S}$ consists of a set of images $X^{\mathcal{S}} = \{x_{s,i}\}_{i=1,\dots,N^{\mathcal{S}}}$ with their corresponding pixel-level annotations $Y^{\mathcal{S}} = \{y_{s,i}\}_{i=1,\dots,N^{\mathcal{S}}}$, and the dataset from the target domain $\mathcal{T}$ consists of a set of images $X^{\mathcal{T}} = \{x_{t,i}\}_{i=1,\dots,N^{\mathcal{T}}}$ without annotations. The task is, given a segmentation model, to optimize the performance of this model on the target domain using the annotated source dataset $(X^{\mathcal{S}}, Y^{\mathcal{S}})$ and the target images $X^{\mathcal{T}}$ without annotations. The architecture of BiSIDA is shown in Figure 1. Our approach uses *self-ensembling* (Tarvainen and Valpola 2017), which consists of two segmentation networks, a *student network* F^s and a *teacher network* F^t, both having the same architecture (blue in Fig. 1). Given source and target domain images, x_s and x_t, during one iteration of the training process, losses L_s and L_u are computed that in turn are used to update the weights of F^s and F^t.

Continuous Style-induced Image Generator

We propose a continuous style-induced image generator (CSIIG) G that is built on top of the AdaIN image-to-image style transfer generator g, which combines content image c and style image s using a content-style trade-off parameter α. We can combine source and target styles by controlling this parameter α in a continuous manner from 0 to 1,

$$G(c, s, \alpha) = g(\alpha \hat{t} + (1 - \alpha) t_c), \tag{1}$$

where t_c is the feature map describing the content image c, and $\hat{t}$ is a combination of content and style feature maps t_c and t_s, which are both extracted by a fixed pretrained encoder. Finally, BiSIDA ensures that the generated output yields RGB values in the range of $[0, 255]$ by clipping.

Target-guided Supervised Learning

The supervised branch of BiSIDA performs an iterative process, in which images $x_{s,i}$ and $x_{t,i}$ are used to compute a loss value $L_{s,i}$, for $i = 1, \dots, \max\{N^S, N^T\}$ iterations. For notation simplification, we drop the index i in the following. In each iteration, BiSIDA first performs a brightness and contrast perturbation $\mathcal{A}$ on each source domain image x_s. The perturbation is governed by a uniformly-distributed random variable. The resulting perturbed image $\mathcal{A}(x_s)$ is then passed through our style-induced image generator G to perform style transfer using the target domain image x_t. The random perturbation is computed by sampling a uniform distribution $\mathcal{U}(0, 1)$ to produce the content-style trade-off parameter α that controls the style of the transferred image $\tilde{x}_s = G(\mathcal{A}(x_s), x_t, \alpha)$. Generated image $\tilde{x}$ is then passed into the student segmentation network F^s, which computes a probability map $p_s = F^s(\tilde{x}_s)$ that describes, for each pixel, the probability of the pixel belonging to a category j.

Finally, BiSIDA computes the cross-entropy loss for the supervised branch of its model, L_s, between the probability map p_s and the pixel-level annotation y_s belonging to x_s:

$$L_s = -\frac{1}{HW} \sum_{m=1}^{H \times W} \sum_{j=1}^{J} y_{s,m,j} \log(p_{s,m,j}), \tag{2}$$

where H and W are the height and width of the image, and J is the number of categories of the segmentation. Loss L_s is then used to train F^s. BiSIDA has an alternative way to compute L_s that skips the image generation and passes a source image, after the brightness and contrast perturbation, directly into F^s . It is advised to use this alternative for at least some of the input images (with probability $1 - p_{s \to t}$) if BiSIDA is used with an image generator that creates blurry output images, which in turn make it difficult for F^s to segment properly.

Source-guided Unsupervised Learning

To start with, we introduce the generation of the pseudo-label that guides the self-learning on the target dataset. Given that our model is more adapted to the source domain where our supervised learning is performed, the quality of produced pseudo-label is generally higher. Consequently, pseudo-label will be computed from target images transferred to the direction of the appearance of the source domain in our framework. Similar to the supervised learning branch, BiSIDA first performs a random brightness and contrast perturbation $\mathcal{A}$ on a target domain image x_t. The resulting image $\mathcal{A}(x_t)$ is style-transferred using k randomly-sampled source images $\{x_{s,i}\}_{i=1}^k$ as style images by the style-induced image generator G with probability $p_{t \to s}$, yielding $\tilde{x}_{t,i} = G(\mathcal{A}(x_t), x_{s,i}, \alpha)$. With probability $1 - p_{t \to s}$, the image generator is skipped, and $\{\tilde{x}_{t,i}\}_{i=1}^k = \{\mathcal{A}(x_t)\}$ with $k = 1$ (this is only needed if an image generator is used with BiSIDA that produces low-resolution images).

BiSIDA passes transformed images $\{\tilde{x}_{t,i}\}_{i=1}^k$ through the teacher model F^t individually to acquire stable predictions $\tilde{p}_{t,i} = F^t(\tilde{x}_{t,i})$. BiSIDA then averages these predictions to compute the probability map $p_l = \frac{1}{k} \sum_{i=1}^k \tilde{p}_{t,i}$ that is used to define the pseudo-label of x_t as follows. BiSIDA employs a sharpening function, which is widely adopted in various semi-supervised learning methods (Berthelot et al. 2019), to re-arrange the distribution of the probability map p_l:

$$p_{l,i}^{\triangle} = \text{Sharpening}(p_l, T)_i = \frac{p_{l,i}^{\frac{1}{T}}}{\sum_{j=1}^{J} p_{l,j}^{\frac{1}{T}}}, \tag{3}$$

where the ‘temperature’ T is a control parameter.

Finally BiSIDA computes the pseudo-label $q_t = \text{argmax}(p_l^{\triangle})$, which it then uses to compute the loss L_u of its unsupervised learning branch. Concretely, BiSIDA perturbs the target image x_t using the random brightness and contrast perturbation $\mathcal{A}$ and passes it through the student network F^s to compute the probability map $p_t = F^s(\mathcal{A}(x_t))$. In this way, BiSIDA performs consistency regularization between two images perturbed in different ways from an identical target image x_t.

In practice, the categories present in a segmentation may be represented by a wide range of pixel numbers (imbalance) or pixel values (complexity). The imbalance and complexity found in training datasets causes the model to bias its prediction toward popular or easier categories, especially if it is trained by relying on labels that are not guaranteed to be accurate, as may apply to pseudo-labels learned in a semi-supervised manner. To address this problem, BiSIDA employs a class-balanced reweighting mechanism in computing the unsupervised loss L_u as follows. It first determines the prior distribution of the categories, by computing, for each category j, the proportion d_j of pixels labeled j in all images of the dataset. BiSIDA then computes the reweighting factor w for each class j as:

$$w_j = \frac{1}{\lambda \, d_j^{\gamma}}, \tag{4}$$

where λ and γ are hyper-parameters. This yields the final unsupervised loss $L_u =$

$$-\frac{1}{HW} \sum_{m=1}^{H \times W} [\mathbb{1}(\max(p_{l,m}) \geq \tau) \sum_{j=1}^{J} w_j q_{t,m,j} \log(p_{t,m,j})], \tag{5}$$

where τ is a confidence threshold that has the following function: For each pixel, we assume that the pseudo-label

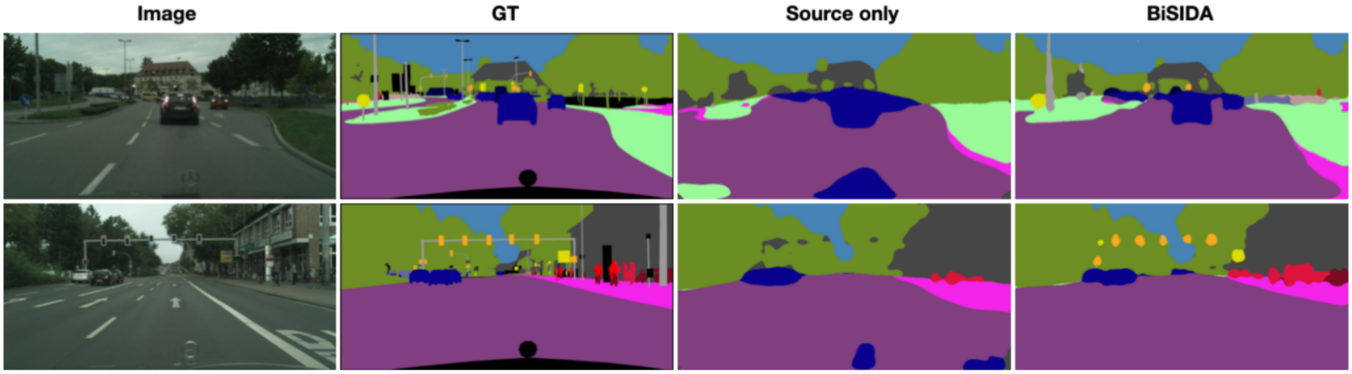

Figure 2: Sample results (column “BiSIDA”). Target-domain testing images from the CityScapes dataset (column “Images”) were segmented by BiSIDA on the GTA5 dataset. The ground truth segmentation (“GT”) and results of a model trained only with source domain images (“Source only”) are also shown. Note that our method is capable of capturing rare and difficult categories, such as traffic lights and signs.

is correct if the maximum probability among all categories exceeds the confidence threshold τ, and we thus should include it in the computation of the cross-entropy loss (indicator function $\mathbb{1}$ returns 1). If the maximum probability is not higher than τ, i.e., the pixel has a relatively flat category distribution, we do not take the pixel into account in computing the loss L_u (indicator function $\mathbb{1}$ returns 0).

Optimization

To summarize, our BiSIDA framework comprises two processes, a supervised learning process, performed on the labeled source dataset, and an unsupervised learning process, performed on the unlabeled target dataset, using the techniques consistency regularization and pseudo-labeling. BiSIDA computes the final loss L, given a weight λ_u, as the sum of the two losses L_s and the weighted L_u:

$$L = L_s + \lambda_u L_u. \tag{6}$$

During the training process, the weights of the student network F_s are updated toward the direction of the weight gradient computed via back-propagation of the loss L, while the weights of the teacher network are updated as the exponential moving averages of the corresponding weights of the student network.

Concretely, the weight θ_i^t of the teacher network F^t at the ith iteration of the training process is updated as the exponential moving average of the weight θ_i^s of the student network F^s as follows:

$$\theta_i^t = \eta\theta_{i-1}^t + (1-\eta)\theta_i^s \tag{7}$$

given an exponential moving average decay η.

Experiments

We conducted extensive experiments on two commonly used synthetic-to-real segmentation benchmarks. Comparisons with several state-of-the-art (SOTA) methods and ablation studies are presented to show the effectiveness of our BiSIDA framework. We visualize some segmentation results in Figure 2.

Datasets

We used two synthetic-to-real benchmarks, GTA5-to-CityScapes and SYNTHIA-to-CityScapes. The CityScapes dataset (Cordts et al. 2016) consists of images of real street scenes of spatial resolution of 2048×1024 pixels. It includes 2,975 images for training, 500 images for validation, and 1,525 images for testing. In our experiments, we used the 500 validation images as a test set. The GTA5 dataset (Richter et al. 2016) includes 24,966 synthetic images with a resolution of 1914×1052 pixels that are obtained from the video game GTA5 along with pixel-level annotations that share all 19 common categories of CityScapes. For the SYNTHIA dataset (Ros et al. 2016), we used the SYNTHIA-RAND-CITYSCAPES subset, which contains 9,400 rendered images of size 1280×760 and shares 16 common categories with the CityScapes dataset.

Network Architecture

Image generator: We use the encoder/decoder architecture of AdaIN (Huang and Belongie 2017). As in AdaIN, we keep our continuous style-induced image generator lightweighted and computationally affordable, and therefore adopted the first several layers up to relu4_1 of a fixed pre-trained VGG-19 network as the encoder in our experiments. For the decoder, we reversed the order of layers in the encoder and replaced the pooling layers by nearest upsampling (Huang and Belongie 2017).

Segmentation network: We chose the classical semantic segmentation network FCN-8s (Long, Shelhamer, and Darrell 2015) with a VGG16 backbone network, pre-trained with ImageNet, for both the student and teacher network.

Training Protocol

We trained the **continuous style-induced image generator** using randomly-cropped 640 × 320 images, and a batch size of 4. The ADAM optimizer was used with a learning rate of 1×10^{-5} and momentum of 0.9 and 0.999. To balance the reconstruction of the content image and the extraction from the style image (Huang and Belongie 2017), we used a style

Method	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*light*	*sign*	*vegetation*	*terrain*	*sky*	*person*	*rider*	*car*	*truck*	*bus*	*train*	*motocycle*	*bicycle*	mIoU
(Zhang, David, and Gong 2017)	74.9	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	16.6	29.0
(Zou et al. 2018)	66.7	26.8	73.7	14.8	9.5	28.3	25.9	10.1	75.5	15.7	51.6	47.2	6.2	71.9	3.7	2.2	5.4	18.9	32.4	30.9
(Tsai et al. 2018)	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
(Hoffman et al. 2018)	85.2	37.2	76.5	21.8	15.0	23.8	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
(Vu et al. 2019)	86.9	28.7	78.7	28.5	25.2	17.1	20.3	10.9	80.0	26.4	70.2	47.1	8.4	81.5	26.0	17.2	18.9	11.7	1.6	36.1
(Wu et al. 2018)	82.3	26.7	77.4	23.7	20.5	20.4	30.3	15.9	80.9	25.4	69.5	52.6	11.1	79.6	24.9	21.2	1.3	17.0	6.7	36.2
(Luo et al. 2019)	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
(Sankaranarayanan et al. 2018)	88.0	30.5	78.6	25.2	23.5	16.7	23.5	11.6	78.7	27.2	71.9	51.3	19.5	80.4	19.8	18.3	0.9	20.8	18.4	37.1
(Li, Yuan, and Vasconcelos 2019)	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
(Yang and Soatto 2020)	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
(Wang et al. 2020)	88.1	35.8	83.1	25.8	23.9	29.2	28.8	28.6	83.0	36.7	82.3	53.7	22.8	82.3	26.4	38.6	0.0	19.6	17.1	42.4
(Choi, Kim, and Kim 2019)	90.2	51.5	81.1	15.0	10.7	37.5	35.2	28.9	84.1	32.7	75.9	62.7	19.9	82.6	22.9	28.3	0.0	23.0	25.4	42.5
Ours	89.3	40.9	82.5	30.9	24.7	20.9	26.9	32.1	81.8	33.1	81.6	53.4	20.3	83.0	24.8	29.4	0.0	28.6	36.6	**43.2**

Table 1: Comparison of our BiSIDA model with other methods on the GTA5-to-CityScapes benchmark using models with VGG-16 as backbone. The mIoU represents the average of individual mIoUs among all 19 categories between GTA5 and CityScapes.

Method	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*light*	*sign*	*vegetation*	*sky*	*person*	*rider*	*car*	*bus*	*motocycle*	*bicycle*	mIoU	mIoU*
(Zhang, David, and Gong 2017)	65.2	26.1	74.9	0.1	0.5	10.7	3.5	3.0	76.1	70.6	47.1	8.2	43.2	20.7	0.7	13.1	29.0	34.8
(Vu et al. 2019)	67.9	29.4	71.9	6.3	0.3	19.9	0.6	2.6	74.9	74.9	35.4	9.6	67.8	21.4	4.1	15.5	31.4	36.6
(Tsai et al. 2018)	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
(Luo et al. 2019)	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
(Zou et al. 2018)	69.6	28.7	69.5	12.1	0.1	25.4	11.9	13.6	82.0	81.9	49.1	14.5	66.0	6.6	3.7	32.4	35.4	40.7
(Wu et al. 2018)	79.9	30.4	70.8	1.6	0.6	22.3	6.7	23.0	76.9	73.9	41.9	16.7	61.7	11.5	10.3	38.6	35.4	41.7
(Sankaranarayanan et al. 2018)	80.1	29.1	77.5	2.8	0.4	26.8	11.1	18.0	78.1	76.7	48.2	15.2	70.5	17.4	8.7	16.7	36.1	42.1
(Chen, Li, and Gool 2018)	77.7	30.0	77.5	9.6	0.3	25.8	10.3	15.6	77.6	79.8	44.5	16.6	67.8	14.5	7.0	23.8	36.2	41.7
(Chen et al. 2019)	78.3	29.2	76.9	11.4	0.3	26.5	10.8	17.2	81.7	81.9	45.8	15.4	68.0	15.9	7.5	30.4	37.3	43.0
(Choi, Kim, and Kim 2019)	90.1	48.6	80.7	2.2	0.2	27.2	3.2	14.3	82.1	78.4	54.4	16.4	82.5	12.3	1.7	21.8	38.5	46.6
(Li, Yuan, and Vasconcelos 2019)	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
(Yang and Soatto 2020)	84.2	35.1	78.0	6.1	0.4	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
Ours	87.4	42.4	79.0	17.0	0.1	23.9	2.8	22.9	82.0	80.4	51.1	19.1	76.7	33.3	14.4	41.2	**42.1**	**48.7**

Table 2: Comparison of our framework with other methods on SYNTHIA to CityScapes benchmark using models with VGG-16 as backbone. The mIoU represents the average of individual mIoUs among all 16 categories between SYNTHIA and CityScapes while the mIoU^* represents that among 13 common categories excluding wall, fence and pole.

weight of 0.1 in calculating the loss function of the image generator.

The **segmentation model** was trained on images randomly cropped to 960×480 pixels with batch size of 1. On the GTA5 dataset, we applied the ADAM optimizer with a learning rate of 1×10^{-5}, weight decay of 5×10^{-4} and momentum of 0.9 and 0.999. For the SYNTHIA dataset, we adopted the SGD optimizer with a learning rate of 1×10^{-5}, momentum of 0.99 and and weight decay of 5×10^{-4}. We follow conventions from previous works and set the exponential moving average decay for the teacher model to 0.999 (Choi, Kim, and Kim 2019) and the confidence threshold τ in the pseudo-label generation process to 0.9 (French, Mackiewicz, and Fisher 2018). The probability of performing target-guided image transfer $p_{s \to t}$ and source-guided image transfer $p_{t \to s}$ is 0.5. The number k of source images for style transfer, the unsupervised weight λ_u, and the sharpening temperature T are set to 4, 1, and 0.25, respectively, based on pilot experiments. We also determined, in a pilot trial, the reweighting parameters γ and λ, by inspecting the distribution of the number d_j of pixels per category j and its inverse, ensuring that the weight w_j serves to emphasize categories with smaller number of pixels sufficiently (Eq. 4. We fixed $\gamma = 1/3$ and $\lambda = 5$ for the experiments on SYNTHIA and around $1/6$ and 1 for the experiments on GTA V.

The CSIIG and two segmentation networks F^s and F^t are trained on a NVIDIA Tesla V100 GPU.

Image generator	mIoU	mIoU*
MUNIT (480 × 480)	36.7	43.4
MUNIT (960 × 480)	36.5	43.2
FDA ($\beta = 0.01$)	29.0	34.0
FDA ($\beta = 0.03$)	32.1	37.9
FDA ($\beta = 0.1$)	32.3	38.6
FDA ($\beta = 0.3$)	30.0	35.7
Ours (AdaIN)	**42.1**	**48.7**

Table 3: Comparison with two different architectures of image generators. 480×480 means input images to the generator are cropped to 480×480 pixels, while 960×480 means images are cropped to 960×480. mIoU represents averaged mIoU over 16 classes and mIoU* over 13 classes.

Comparisons with SOTA Methods

We first compare the performance of BiSIDA on the GTA5-to-CityScapes benchmark with that of other methods with VGG-16 as the backbone (Table 1), using the mean Intersection-over-Union (mIoU) metric, where the mean is taken over all categories. Our results reveal that BiSIDA outperforms most competitive methods, especially TGCF-DA+SE (which employs adversarial training as augmentation), by 1.6% (i.e., 0.7 pp).

We present the performance of our and other methods on the SYNTHIA-to-CityScapes benchmark using two metrics (Table 2). Due to fewer training data and less realistic-looking images, this task is more difficult than the previous one. However, BiSIDA outperforms the current state-of-the-art method by a margin of 3.8% (or 1.6 pp) over 16 categories.

Ablation Studies

Effect of our continuous style-induced image generator and domain parameterization: We substitute our continuous style-induced image generator from our BiSIDA paradigm with two high-performing image generating methods, namely MUNIT (Huang et al. 2018) and FDA (Yang and Soatto 2020), to evaluate its influence on BiSIDA performance when tested on the SYNTHIA-to-CityScapes benchmark with VGG-16 as the backbone. In contrast to ours, both these two methods generate images without loss of resolution: MUNIT is trained in a complicated adversarial manner through cycle consistency, while FDA computes the transferred images through Fourier transformation in a deterministic way without any pretraining process. The experiment revealed that neither method was able to produce a mix of styles between source and target images and parameterize domains as well as our continuous style-induced image generator.

Concretely, we pretrained a MUNIT model using images from both the source and target domains (cropped to 480×480 pixels due to memory limitations) for 340,000 iterations and fixed it during the following experiments. For the sake of a fair comparison, we ran two separate experiments with the pretrained MUNIT model using images cropped to 480×480 and 960×480, respectively. For the experiments on FDA, we also used input images of size 960×480 and applied various β values ranging from 0.01 to 0.3. For the MUNIT and FDA experiments, we set $p_{s \to t}$ and $p_{t \to s}$ to 1 because of the improved quality of generated images. We observed that the BiSIDA performance with MUNIT or FDA as the image generator is significantly lower than the BiSIDA performance with CSIIG (see Table 3). We suggest that the results with MUNIT are inferior to the results with our CSIIG due to the lack of domain parameterization by MUNIT. The results with FDA are worse than with MUNIT, most likely because the Fourier transformation cannot generate high-dimensional image perturbations for unsupervised learning. We suggest that the significantly lower accuracy of both methods is also due to the continuous parameterization between two domains of our method, which facilitates the generation of high-quality pseudo-labels and enables gradual domain adaptation. Moreover, we found that the computational burden of BiSIDA with MUNIT is more than ten times larger than BiSIDA with CSIIG (138 h versus 13 h training time).

S2T	T2S	PL	SE	GTA	SYN
				29.3	28.9
✓				34.7	32.0
	✓			31.8	31.4
✓	✓			35.1	40.2
✓	✓	✓		35.4	40.8
✓	✓		✓	39.4	41.8
✓	✓	✓	✓	**43.2**	**42.1**

Table 4: Ablation study on the style-induced image transfer and unsupervised modules. S2T stands for Source-domain-to-Target-domain image transfer, T2S stands for Target-domain-to-Source-domain image transfer, PL stands for pseudo-labeling and SE stands for self-ensembling. GTA represents the mIoU (16 classes) from the GTA5-to-CityScapes dataset while SYN represents the mIoU from the SYNTHIA-to-CityScapes dataset.

Effect of Source-to-Target (S2T) and Target-to-Source (T2S) Style-induced image transfer, unsupervised learning via pseudo-labeling (PL), and self-ensembling (SE): In this ablation study, we removed all four techniques, S2T, T2S, PL, and SE, from BiSIDA, as well as all meaningful combinations of these techniques. Results are presented in Table 4. Removing all four techniques yields a baseline model that does not include any transfer learning but simply consists of a single semantic segmentation network that is trained on the labeled source images and then tested on the target images (see 3rd column in Fig. 2). The ablation of the target-guided image transfer CSIIG means we deactivated the unsupervised learning branch of BiSIDA and report results on supervised style transfer learning. We then tested three more cases that include variations of the unsupervised learning branch of BiSIDA, ablating PL, SE, and PL & SE.

As we can observe from the results in Table 4 rows 2 and 3, the source-guided and target-guided image transfer

cbPert	T2S	mIoU	mIoU*
		32.2	38.6
✓		32.1	38.5
	✓	41.8	48.3
✓	✓	**42.1**	**48.7**

Table 5: Experiments on augmentation methods on SYNTHIA-to-CityScapes. cbPert represents contrast and brightness perturbation, T2S represents source-guided image translation performed on target-domain images. mIoU represents averaged mIoU over 16 classes and mIoU* over 13 common classes.

each improved the performance on both benchmarks when applied without the other techniques. It is also worth noting that the improvement over the baseline brought by the target-guided image transfer is smaller than that of the source-guided imaged transfer since the target domain images translated with styles from source domain cannot provide better self-guidance without having the source domain aligned to the intermediate continuous space. A more significant performance leap is shown when these two transfers are performed simultaneously, especially on the SYNTHIA-to-CityScapes benchmark where the domain gap is larger than that of the GTA5-to-CityScapes, showing the advantage of our bidirectional style-induced image translation method.

As for the modules in the unsupervised learning branch of BiSIDA, when pseudo-labeling is disabled, we use the probability maps to compute the unsupervised loss, and the problem is transformed to entropy minimization. When self-ensembling is disabled, the probability maps are generated by the segmentation model itself. From the results in Table 4 rows 5 and 6, we conclude that both pseudo-labeling and self-ensembling contribute to a similar degree in enhancing the performance of BiSIDA. Additionally, we also observe that most of the improvement on GTA5-to-CityScapes comes from the application of the unsupervised learning modules (PL, SE) while the improvement on SYNTHIA-to-CityScapes, on the other hand, comes from the style-induced image translation process. We thus suggest that the challenge in the GTA5-to-CityScapes benchmark is to perform feature-level alignment while the challenge for the SYNTHIA-to-CityScapes benchmark is to perform pixel-level alignment.

Effect of brightness and contrast pertubation and source-guided image transfer: To learn more about the effectiveness of the brightness-and-contrast perturbation and source-guided image transfer performed on target images in the unsupervised learning branch of BiSIDA, we conducted an ablation study that removes either or both pertubation techniques with all other settings fixed. We used the SYNTHIA-to-CityScapes benchmark, where pixel-level alignment plays a more important role. The results, shown in Table 5, reveal that the introduction of source-guided image transfer significantly improves performance by a large margin. The experiment also shows that the brightness and contrast perturbation is not a sufficiently strong perturbation method for consistency regularization. It only helps slightly when the source-guided image transfer is also applied, presumably because it enhances the stochasticity in the high-dimensional perturbation process.

weight	0.1	0.5	1.0	5.0	10.0
mIoU	37.8	41.9	**42.1**	39.8	38.6
mIoU*	44.3	48.1	**48.7**	46.3	45.2

Table 6: Comparison with different unsupervised loss weights λ_u. mIoU represents averaged mIoU over 16 classes of SYNTHIA-to-CityScapes and mIoU* represents that over 13 classes.

Discussion

Unsupervised learning weight: In our BiSIDA, the unsupervised loss weight λ_u is a crucial hyperparameter to balance the focus of our model between the supervised learning on the labeled source dataset and the unsupervised learning on the unlabeled target dataset. We investigate the effect of using different unsupervised loss weights on our method, we conducted an experiment on the SYNTHIA-to-CityScapes benchmark with five different unsupervised loss weights.

The results in Table 6 reveal that when the weight is too small, the benefit of unsupervised learning is limited and consistency regularization cannot be performed effectively. When the weight is too large, the model fails to achieve satisfying performance. A reason may be that the model becomes bias prone and prefers an easier category in the early stage of training. We found that our model reaches a peak in performance when the weight λ_u is set to 1.

Number of style images used in source-guided image transfer: We also explore the effect of the number of source domain style images k in the unsupervised phase. The results indicate that when the number of style images is small (fewer than 4), the model achieves suboptimal performance, probably because the quality of the generated pseudo-label is undermined, while increasing the number of style images (more than 4) is also not beneficial since it increases the computational costs without improving performance.

Conclusion

We proposed a Bidirectional Style-induced Domain Adaptation (BiSIDA) framework for training a semantic segmentation model via target-guided supervised learning and source-guided unsupervised learning. The framework is particularly useful for solving segmentation problems where the labels for the target domain images are expensive to acquire but labels for source domain images are are not, e.g., because the source domain images might be synthesized. With the employment of our continuous style-induced image generator, we show the effectiveness of learning from unlabeled target datasets by providing high-dimensional perturbations for consistency regularization. Furthermore, our work also reveals that the alignment of source and target domains from both directions is achievable without requiring adversarial training.

Acknowledgements

We acknowledge grants by NSF (1838193) and ONR (MURI N00014-19-1-2571, associated with AUS-MURIB000001) and the Boston University Hariri Institute for Computing for partial support.

References

Berthelot, D.; Carlini, N.; Goodfellow, I. J.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *NeurIPS*.

Chen, Y.; Li, W.; Chen, X.; and Gool, L. V. 2019. Learning Semantic Segmentation From Synthetic Data: A Geometrically Guided Input-Output Adaptation Approach. In *CVPR*.

Chen, Y.; Li, W.; and Gool, L. V. 2018. ROAD: Reality Oriented Adaptation for Semantic Segmentation of Urban Scenes. In *CVPR*.

Choi, J.; Kim, T.; and Kim, C. 2019. Self-Ensembling With GAN-Based Data Augmentation for Domain Adaptation in Semantic Segmentation. In *ICCV*.

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 *CVPR*.

Dumoulin, V.; Shlens, J.; and Kudlur, M. 2017. A Learned Representation For Artistic Style. In *ICLR*.

French, G.; Aila, T.; Laine, S.; Mackiewicz, M.; and Finlayson, G. 2019. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. *arXiv preprint arXiv:1906.01916* .

French, G.; Mackiewicz, M.; and Fisher, M. H. 2018. Self-ensembling for visual domain adaptation. In *ICLR*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image Style Transfer Using Convolutional Neural Networks. In *CVPR*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Networks. *CoRR* abs/1406.2661.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised Learning by Entropy Minimization. In *NIPS*.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.; Isola, P.; Saenko, K.; Efros, A. A.; and Darrell, T. 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *ICML*.

Hoffman, J.; Wang, D.; Yu, F.; and Darrell, T. 2016. FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation. *CoRR* abs/1612.02649. URL http://arxiv.org/abs/1612.02649.

Huang, X.; and Belongie, S. J. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *ICCV*.

Huang, X.; Liu, M.; Belongie, S. J.; and Kautz, J. 2018. Multimodal Unsupervised Image-to-Image Translation. In *ECCV*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *ECCV*.

Lee, D.-H. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop, ICML*.

Li, C.; and Wand, M. 2016. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. In *ECCV*.

Li, Y.; Yuan, L.; and Vasconcelos, N. 2019. Bidirectional Learning for Domain Adaptation of Semantic Segmentation. In *CVPR*.

Lian, Q.; Duan, L.; Lv, F.; and Gong, B. 2019. Constructing Self-Motivated Pyramid Curriculums for Cross-Domain Semantic Segmentation: A Non-Adversarial Approach. In *ICCV*.

Liu, M.; Breuel, T.; and Kautz, J. 2017. Unsupervised Image-to-Image Translation Networks. In *NeurIPS*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; and Yang, Y. 2019. Taking a Closer Look at Domain Shift: Category-Level Adversaries for Semantics Consistent Domain Adaptation. In *CVPR*.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for Data: Ground Truth from Computer Games. In *ECCV*.

Ros, G.; Sellart, L.; Materzynska, J.; Vázquez, D.; and López, A. M. 2016. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *CVPR*.

Sankaranarayanan, S.; Balaji, Y.; Jain, A.; Lim, S.; and Chellappa, R. 2018. Learning From Synthetic Data: Addressing Domain Shift for Semantic Segmentation. In *CVPR*.

Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685* .

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*.

Tsai, Y.; Hung, W.; Schulter, S.; Sohn, K.; Yang, M.; and Chandraker, M. 2018. Learning to Adapt Structured Output Space for Semantic Segmentation. In *CVPR*.

Ulyanov, D.; Lebedev, V.; Vedaldi, A.; and Lempitsky, V. S. 2016. Texture Networks: Feed-forward Synthesis of Textures and Stylized Images. In *ICML*.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. S. 2017. Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis. In *CVPR*.

Vu, T.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. In *CVPR*.

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 *CVPR*.

Wu, Z.; Han, X.; Lin, Y.; Uzunbas, M. G.; Goldstein, T.; Lim, S.; and Davis, L. S. 2018. DCAN: Dual Channel-Wise Alignment Networks for Unsupervised Scene Adaptation. In *ECCV*.

Yang, Y.; and Soatto, S. 2020. FDA: Fourier domain adaptation for semantic segmentation. In *CVPR*.

Yi, Z.; Zhang, H. R.; Tan, P.; and Gong, M. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *ICCV*.

Zhang, Y.; David, P.; and Gong, B. 2017. Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. In *ICCV*.

Zhu, J.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *ICCV*.

Zou, Y.; Yu, Z.; Kumar, B. V. K. V.; and Wang, J. 2018. Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-training. In *ECCV*.