Label-Driven Reconstruction for Domain Adaptation in Semantic Segmentation

Jinyu Yang[0000−0002−7004−3570], Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan[0000−0003−1237−8978], and Junzhou Huang[0000−0002−9548−1227]

University of Texas at Arlington, Texas, USA
{jinyu.yang, weizhi.an, sheng.wang, xinliang.zhu, chaochao.yan}@mavs.uta.edu, jzhuang@uta.edu

Abstract. Unsupervised domain adaptation enables to alleviate the need for pixel-wise annotation in the semantic segmentation. One of the most common strategies is to translate images from the source domain to the target domain and then align their marginal distributions in the feature space using adversarial learning. However, source-to-target translation enlarges the bias in translated images and introduces extra computations, owing to the dominant data size of the source domain. Furthermore, consistency of the joint distribution in source and target domains cannot be guaranteed through global feature alignment. Here, we present an innovative framework, designed to mitigate the image translation bias and align cross-domain features with the same category. This is achieved by 1) performing the target-to-source translation and 2) reconstructing both source and target images from their predicted labels. Extensive experiments on adapting from synthetic to real urban scene understanding demonstrate that our framework competes favorably against existing state-of-the-art methods.

Keywords: Image-to-image Translation, Image Reconstruction, Domain Adaptation, Semantic Segmentation

1 Introduction

Deep Convolutional Neural Networks (DCNNs) have demonstrated impressive achievements in computer vision tasks, such as image recognition [20], object detection [17], and semantic segmentation [32]. As one of the most fundamental tasks, semantic segmentation predicts pixel-level semantic labels for given images. It plays an extremely important role in autonomous agent applications such as self-driving techniques.

Existing supervised semantic segmentation methods, however, largely rely on pixel-wise annotations which require tremendous time and labor efforts. To overcome this limitation, publicly available synthetic datasets (e.g., GTA [41] and SYNTHIA [42]) which are densely-annotated, have been considered recently. Nevertheless, the most obvious drawback of such a strategy is the poor knowledge generalization caused by domain shift issues (e.g., appearance

and spatial layout differences), giving rise to dramatic performance degradation when directly applying models learned from synthetic data to real-world data of interest. In consequence, domain adaptation has been exploited in recent studies for cross-domain semantic segmentation, where the most common strategy is to learn domain-invariant representations by minimizing distribution discrepancy between source and target domains [56, 29], designing a new loss function [61], considering depth information [7, 49], or alternatively generating highly confident pseudo labels and re-training models with these labels through a self-training manner [62, 28, 22, 38, 51, 25, 54]. Following the advances of Generative Adversarial Nets (GAN) [19], adversarial learning has been used to match cross-domain representations by minimizing an adversarial loss on the source and target representations [21, 34, 35, 13], or adapting structured output space across two domains [45, 28]. Recent studies further consider the pixel-level (e.g., texture and lighting) domain shift to enforce source and target images to be domain-invariant in terms of visual appearance [58, 1, 52, 36, 9, 53]. This is achieved by translating images from the source domain to the target domain by using image-to-image translation models such as CycleGAN [60] and UNIT [30].

Despite these painstaking efforts, we are still far from being able to fully adapt cross-domain knowledge mainly stemming from two limitations. First, adversarial-based image-to-image translation introduces inevitable bias to the translated images, as we cannot fully guarantee that the translated source domain $\mathcal{F}(\mathcal{X}_s)$ is identical to the target domain $\mathcal{X}_t$ ($\mathcal{X}_s$ and $\mathcal{X}_t$ denote two domains, and $\mathcal{F}$ indicates an image-to-image translation model). This limitation is especially harmful to the source-to-target translation [58, 1, 52, 36, 28], since the data size of the source domain is much larger than the target domain in most of domain adaptation problems. Moreover, source-to-target translation is more computationally expensive than target-to-source translation. Second, simply aligning cross-domain representations in the feature space [21, 1, 45] ignores the joint distribution shift (i.e., $\mathcal{P}(G(\mathcal{X}_s), Y_s) \neq \mathcal{P}(G(\mathcal{X}_t), Y_t)$, where G is used for feature extraction, while Y_s and Y_t indicate ground truth labels). These limitations give rise to severe false positive and false negative issues in the target prediction. This problem can get even worse when there is a significant discrepancy in layout or structure between the source and target domains, such as adapting from synthetic to real urban traffic scenes.

In this paper, we propose an innovative domain adaptation framework for semantic segmentation. The key idea is to reduce the image translation bias and align cross-domain feature representations through image reconstruction. As opposed to performing source-to-target translation [1, 52, 28], for the first time, we conduct the target-to-source translation to make target images indistinguishable from source images. This enables us to substantially reduce the bias in translated images and allows us to use original source images and their corresponding ground truth to train a segmentation network. Compared to the source-to-target translation, our method is also much more efficient. Besides, a reconstruction network is designed to reconstruct both source and target images from their predicted labels. It is noteworthy that we reconstruct images directly from the

Fig. 1. An example of our method on synthetic-to-real urban scene adaptation. Given a target-domain (or real) image (a), we first make target-to-source translation to obtain source-like (or synthetic) image (b), and then perform segmentation on these translated images. Our method improves the segmentation accuracy in the target domain by reconstructing both source and target images from their predicted labels (c). (d) illustrates the image reconstructed from (c), while (e) indicates the ground truth label.

label space, rather than the feature space as reported in previous studies. This is essential to guide the segmentation network by penalizing the reconstructed image that semantically deviates from the corresponding input image. Most importantly, this strategy enforces cross-domain features with the same category close to each other.

The performance of our method is evaluated on synthetic-to-real scenarios of urban scene understanding, i.e., GTA5 to Cityscapes and SYNTHIA to Cityscapes. Our results demonstrate that the proposed method achieves significant improvements compared with existing methods. Figure 1 demonstrates an example of our model in adapting cross-domain knowledge in semantic segmentation tasks and reconstructing the input image from its output label. We also carry out comprehensive ablation studies to analyze the effectiveness of each component in our framework.

The contribution of this paper is threefold.

- For the first time, we propose and investigate the target-to-source translation in domain adaptation. It reduces the image translation bias and is more computationally efficient compared to the widely-used source-to-target translation.
- To enforce semantic consistency, we introduce a label-driven reconstruction module that reconstructs both source and target images from their predicted labels.
- Extensive experiments show that our method achieves the new state-of-the-art performance on adapting synthetic-to-real semantic segmentation.

2 Related Work

Semantic Segmentation Recent achievements in semantic segmentation mainly benefit from the technical advances of DCNNs, especially the emergence of Fully Convolutional Network (FCN) [32]. By adapting and extending contemporary deep classification architectures fully convolutionally, FCN enables pixel-wise semantic prediction for any arbitrary-sized inputs and has been widely recognized

as one of the benchmark methods in this field. Numerous methods inspired by FCN were then proposed to further enhance segmentation accuracy, which have exhibited distinct performance improvement on the well-known datasets (e.g., PASCAL VOC 2012 [14] and Cityscapes [11]) [5, 31, 59, 6, 4].

However, such methods heavily rely on human-annotated, pixel-level segmentation masks, which require extremely expensive labeling efforts [11]. In consequence, weakly-supervised methods, which are based on easily obtained annotations (e.g., bounding boxes and image-level tags), were proposed to alleviate the need for effort-consuming labeling [12, 40]. Another alternative is to resort to freely-available synthetic datasets (e.g., GTA5 [41] and SYNTHIA [42]) with pixel-level semantic annotations. However, models learned on synthetic datasets suffer from significant performance degradation when directly applied to the real datasets of interest, mainly owing to the domain shift issue.

Domain Adaptation Domain adaptation aims to mitigate the domain discrepancy between a source and a target domain, which can be further divided into supervised adaptation, semi-supervised adaptation, and unsupervised adaptation, depending on the availability of labels in the target domain. The term unsupervised domain adaptation refers to the scenario where target labels are unavailable and have been extensively studied [33, 48, 15, 46, 47, 55, 57].

Recent publications have highlighted the complementary role of pixel-level and representation-level adaptation in semantic segmentation [1, 37, 58, 52, 7], where the pixel-level adaptation is mainly achieved by translating images from the source domain to the target domain (source-to-target translation). Specifically, unpaired image-to-image translation is used in CyCADA [1] to achieve pixel-level adaptation by restricting cycle-consistency. Similarly, FCAN achieves the image translation by combining the image content in the source domain and the "style" from the target domain [58]. I2IAdapt [37] further considers to align source and target representations based on an image-to-image translation strategy, attempting to adapt domain shift. Instead of using the adversarial learning for image translation, DCAN performs source-to-target translation by leveraging target images for channel-wise alignment [52]. Driven by the fact that geometry and semantics are coordinated with each other, GIO-Ada augments the standard image translation network by integrating geometric information [7]. However, source-to-target translation introduces substantial bias to the translated images, given that the size of the source domain is usually much larger than the target domain. To address this problem, we propose the first-of-its-kind target-to-source image translation to reduce pixel-level domain discrepancy. Compared to the source-to-target translation, it is more computationally efficient and enables us to remove the uncertainty by training the segmentation network with original source images and their corresponding labels.

Motivated by the observation that cross-domain images (e.g., GTA5 and Cityscapes) often share tremendous structural similarities, ASN [45] adapts structured output based on the adversarial learning. The strength of this method is its ability to provide weak supervision to target images by enforcing target outs to be indistinguishable from source outputs. However, it is limited to the scenario

where two domains have a huge layout discrepancy, resulting in meaningless predictions for target images. To address this limitation, we further enforce the semantic consistency between target images and their predicted labels through a reconstruction network.

Inspired by the self-training, [28, 22, 38, 51, 25, 54] generate pseudo labels for target images and then re-training the segmentation model with these labels. It outperforms the existing methods by a large margin. However, such a strategy underestimates the side effect of pseudo labels that are incorrectly predicted. As a consequence, the segmentation model fails to increasingly improve itself using these wrong ground truth. Instead, our method reconstructs source and target input images from the label space to ensure these outputs are semantically correct. The image-to-image translation network in [28] uses a reconstruction loss and a perceptual loss to maintain the semantic consistency between the input image and the reconstruction from the translated image. Different from [28], we design a cycle-reconstruction loss in our reconstruction network to enforce the semantic consistency between the input image and the reconstruction from the predicted label.

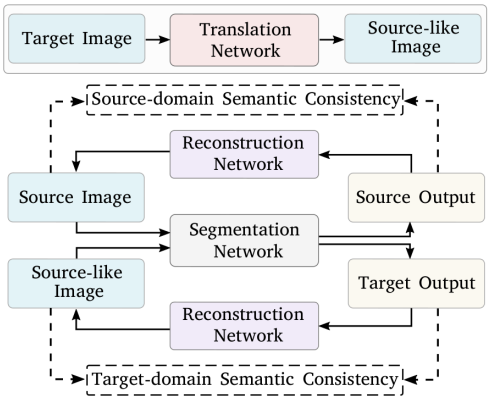

Fig. 2. An overview of our framework.

Reconstruction-based strategy for unsupervised domain adaptation has received considerable attention recently [16, 2]. The key idea is to reconstruct input images from their feature representations to ensure that the segmentation model can learn useful information. Chang et al. [3] follow a similar idea to first disentangle images into the domain-invariant structure and domain-specific texture representations, and then reconstruct input images. LSD-seg [43] first reconstructs images from the feature space, and then apply a discriminator to the reconstructed images. Rather than performing reconstruction from feature representations, we reconstruct both source and target images from their predicted labels.

3 Algorithm

3.1 Overview

The overall design of our framework is illustrated in Figure 2, mainly containing three complementary modules: a translation network $\mathcal{F}$, a segmentation network G, and a reconstruction network $\mathcal{M}$. Given a set of source domain images $\mathcal{X}_s$ with labels Y_s and a set of target domain images $\mathcal{X}_t$ without any annotations.

Our goal is to train G to predict accurate pixel-level labels for $\mathcal{X}_t$. To achieve this, we first use $\mathcal{F}$ to adapt pixel-level knowledge between $\mathcal{X}_t$ and $\mathcal{X}_s$ by translating $\mathcal{X}_t$ to source-like images $\mathcal{X}_{t\to s}$. This is different from existing prevalent methods that translate images from the source domain to the target domain. $\mathcal{X}_s$ and $\mathcal{X}_{t\to s}$ are then fed into G to predict their segmentation outputs $G(\mathcal{X}_s)$ and $G(\mathcal{X}_{t\to s})$, respectively. To further enforce semantic consistency of both source and target domains, $\mathcal{M}$ is then applied to reconstruct $\mathcal{X}_s$ and $\mathcal{X}_{t\to s}$ from their predicted labels. Specifically, a cycle-reconstruction loss is proposed to measure the reconstruction error, which enforces the semantic consistency and further guides segmentation network to predict more accurate segmentation outputs.

3.2 Target-to-source Translation

We first perform the image-to-image translation to reduce the pixel-level discrepancy between source and target domains. As opposed to the source-to-target translation reported in previous domain adaptation methods, we conduct the target-to-source translation through an unsupervised image translation network (Figure 3). Our goal is to learn a mapping $\mathcal{F}: \mathcal{X}_t \to \mathcal{X}_s$ such that the distribution of images from $\mathcal{F}(\mathcal{X}_t)$ is indistinguishable from the distribution of $\mathcal{X}_s$. As a counterpart, the inverse mapping $\mathcal{F}^{-1}: \mathcal{X}_s \to \mathcal{X}_t$, which maps images from $\mathcal{X}_s$ to $\mathcal{X}_t$, is introduced to prevent the mode collapse issue [18]. Two adversarial discriminators $\mathcal{D}_t$ and $\mathcal{D}_s$ are employed for distribution match, where $\mathcal{D}_t$ enforces indistinguishable distribution between $\mathcal{F}(\mathcal{X}_t)$ and $\mathcal{X}_s$, and $\mathcal{D}_s$ encourages indistinguishable distribution between $\mathcal{F}^{-1}(\mathcal{X}_s)$ and $\mathcal{X}_t$ (More details can be found in the **Supplementary**).

Based on the trained model $\mathcal{F}$, we first translate images from $\mathcal{X}_t$ to source-like images $\mathcal{X}_{t\to s} = \mathcal{F}(\mathcal{X}_t)$. Specifically, each image in $\mathcal{X}_{t\to s}$ preserves the same content as the corresponding image in $\mathcal{X}_t$ while demonstrating the common style (e.g., texture and lighting) as $\mathcal{X}_s$. $\mathcal{X}_s$ and $\mathcal{X}_{t\to s}$ are then fed into a segmentation network for semantic label prediction.

Compared to translating images from the source domain to the target domain, the target-to-source translation has three benefits. First, it allows full supervision on the source domain by training the segmentation network with original source images and their corresponding labels. Second, it enables to reduce the bias in translated images. Third, it is computationally efficient lying in the fact that $|\mathcal{X}_t| \ll |\mathcal{X}_s|$.

3.3 Semantic Segmentation

Given that source-like images $\mathcal{X}_{t\to s}$ preserves all semantic information from $\mathcal{X}_t$, we apply a shared segmentation network G to $\mathcal{X}_s$ and $\mathcal{X}_{t\to s}$ to predict their segmentation outputs with the loss function given by,

$$\begin{aligned}\mathcal{L}_G = \mathcal{L}_{seg}(G(\mathcal{X}_s), Y_s) + \mathcal{L}_{seg}(G(\mathcal{X}_{t\to s}), Y_t^{ssl}) + \\ \lambda \mathcal{L}_{adv}(G(\mathcal{X}_s), G(\mathcal{X}_{t\to s})),\end{aligned} \tag{1}$$

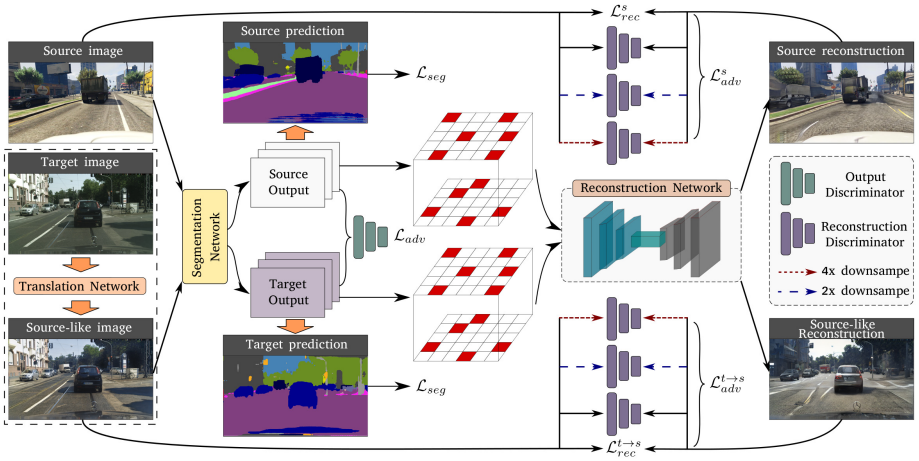

Fig. 3. Schematic overview of our framework which has three modules: (i) a translation network for pixel-level discrepancy reduction by translating target images to source-like images, where source-like images are indistinguishable from source images, (ii) a segmentation network that predicts segmentation outputs for source images and source-like images, and (iii) a reconstruction network for reconstructing source and source-like images from their corresponding label space.

where $\mathcal{L}_{seg}$ indicates the typical segmentation objective, Y_t^{ssl} is pseudo labels of $\mathcal{X}_{t\to s}$ which is derived from [28], $\mathcal{L}_{adv}(G(\mathcal{X}_s), G(\mathcal{X}_{t\to s}))$ is an adversarial loss, and λ leverages the importance of these losses. Specifically, $\mathcal{L}_{adv}(G(\mathcal{X}_s), G(\mathcal{X}_{t\to s}))$ is defined as,

$$\begin{aligned}\mathcal{L}_{adv}(G(\mathcal{X}_s), G(\mathcal{X}_{t\to s})) = \ & \mathbb{E}[logD(G(\mathcal{X}_s))]+ \\ & \mathbb{E}[log(1 - D(G(\mathcal{X}_{t\to s})))],\end{aligned} \tag{2}$$

which enforces G to learn domain-invariant features by confusing the discriminator D. It is noteworthy that we regard the segmentation outputs $G(\mathcal{X}_s)$ and $G(\mathcal{X}_{t\to s})$ as features in our study. This is based on the observation that $\mathcal{X}_s$ and $\mathcal{X}_{t\to s}$ share significant similarities in terms of spatial layouts and structures [45].

3.4 Image Reconstruction from the Label Space

To encourage G to generate segmentation outputs that are semantic consistent, we introduce a reconstruction network $\mathcal{M}$ to reconstruct $\mathcal{X}_\phi$ from $G(\mathcal{X}_\phi) \in \mathbb{R}^{H_\phi \times W_\phi \times C}$, where (H_ϕ, W_ϕ) indicates image size, C represents the number of label classes, and the subscript ϕ can be either s or $t{\to}s$ to denote the source or the target domain. However, directly reconstructing images from the feature space fails to provide semantic consistency constraint to G. On the one hand, $G(\mathcal{X}_\phi)$ encodes rich information which makes the image reconstruction

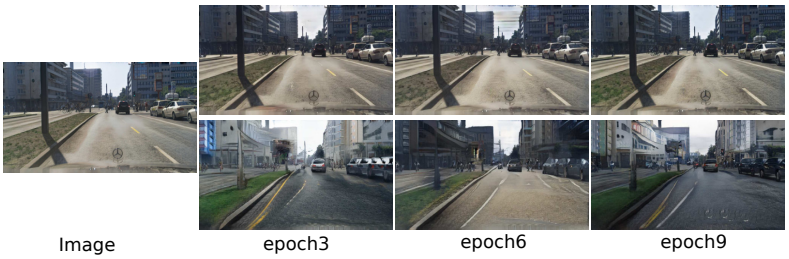

Fig. 4. A comparison between the image reconstruction from feature space and label space (ours). For each input image (first column), the first and second row indicate the images reconstructed from the feature and label space, respectively.

quite straightforward. As illustrated in Figure 4, in just a few epochs, the reconstructed images derived from $\mathcal{M}$ are almost identical to the input images. On the other hand, to enforce cross-domain features with the same category close to each other, it is essential to perform the reconstruction based on the label space. Unfortunately, $G(\mathcal{X}_\phi)$ lies in the feature space instead. To overcome these limitations, the most clear-cut way is to convert $G(\mathcal{X}_\phi)$ to have zeros everywhere except where the index of each maximum value in the last dimension. Doing so formulates the categorical representation of the predicted label that corresponds to $G(\mathcal{X}_\phi)$. Nevertheless, such conversion is non-differentiable and cannot be trained using standard backpropagation.

Driven by the softmax action selection which is commonly used in the reinforcement learning, we apply Boltzmann distributed probabilities to approximate the semantic label map of $G(\mathcal{X}_\phi)$, which is defined as,

$$\Omega_\phi{}^{(h,w,i)} = \frac{exp(G(\mathcal{X}_\phi)^{(h,w,i)}/\tau)}{\sum_{j=1}^{c} exp(G(\mathcal{X}_\phi)^{(h,w,j)}/\tau)}, \tag{3}$$

where τ is a temperature parameter. This conversion is continuous and differentiable, therefore, we use $\mathcal{M}$ to reconstruct input images $\mathcal{X}_\phi$ from Ω_ϕ (Figure 4).

To synthesize high-resolution images from the semantic label map, we use conditional GANs [23] to model the conditional distribution of $\mathcal{X}_\phi$ given Ω_ϕ. To this end, we introduce $\mathcal{M}$ and multi-scale domain discriminators D_k for $k = 1, 2, 3$. $\mathcal{M}$ is designed to reconstruct $\mathcal{X}_\phi$ from Ω_ϕ, and D_k aims to distinguish $\mathcal{X}_\phi$ from $\mathcal{M}(\Omega_\phi)$. Specifically, $\mathcal{M}$ follows the architecture proposed in [24], while D_k is based on PatchGAN [23] that penalizes structure at the scale of image patches. All D_k follow the same network architecture. Besides $\mathcal{X}_\phi$ and $\mathcal{M}(\Omega_\phi)$ themselves, they are downsampled by a factor of 2 and 4 to obtain pyramid of 3 scales for D_1, D_2, and D_3, respectively. It is worth mentioning that D_k is essential to differentiate real and reconstructed images with high resolution [50], owing to its ability in providing large receptive field. The objective function is

given by,

$$\begin{aligned}\mathcal{L}_{adv}^{\phi} = \sum\nolimits_{k=1}^{3}[\mathbb{E}[logD_k(\Omega_\phi, \mathcal{X}_\phi)]+ \\ \mathbb{E}[log(1 - D_k(\Omega_\phi, \mathcal{M}(\Omega_\phi)))]]\end{aligned} \tag{4}$$

To further enforce semantic consistency between $\mathcal{X}_\phi$ and $\mathcal{M}(\Omega_\phi)$, we introduce a cycle-reconstruction loss $\mathcal{L}_{rec}^{\phi}$ to match their feature representations. $\mathcal{L}_{rec}^{\phi}$ contains a VGG perceptual loss and a discriminator feature matching loss, which is defined as,

$$\begin{aligned}\mathcal{L}_{rec}^{\phi} = \mathbb{E}\sum_{m=1}^{M}[||V^{(m)}(\mathcal{M}(\Omega_\phi)) - V^{(m)}(\mathcal{X}_\phi)||_1]+ \\ \mathbb{E}\sum_{k=1}^{3}\sum_{n=1}^{N}[||D_k^{(n)}(\Omega_\phi, \mathcal{X}_\phi)) - D_k^{(n)}(\Omega_\phi, \mathcal{M}(\Omega_\phi))||_1]\end{aligned} \tag{5}$$

where V is a VGG19-based model for extracting high-level perceptual information [24], M and N represent the total number of layers in V and D_k for matching intermediate representations. Note that $\mathcal{L}_{rec}^{\phi}$ penalizes Ω_ϕ when it deviates from the corresponding image $\mathcal{X}_\phi$ in terms of semantic consistency. In this way, $\mathcal{M}$ enables to map features from $\mathcal{X}_{t\to s}$ closer to the features from $\mathcal{X}_s$ with the same label.

Taken together, the training objective of our framework is formulated as,

$$\min_{G,\mathcal{M}} \max_{D,D_1,D_2,D_3} \mathcal{L}_G + \alpha(\mathcal{L}_{adv}^{s} + \mathcal{L}_{adv}^{t\to s}) + \beta(\mathcal{L}_{rec}^{s} + \mathcal{L}_{rec}^{t\to s}) \tag{6}$$

where α and β leverage the importance of losses above. Notably, our method is able to implicitly encourage G to generate semantic-consistent segmentation labels for the target domain.

4 Experiments

In this section, a comprehensive evaluation is performed on two domain adaption tasks to assess our framework for semantic segmentation. Specifically, we consider the large distribution shift of adapting from synthetic (i.e., GTA5 [41] and SYNTHIA [42]) to the real images in Cityscapes [11]. A thorough comparison with the state-of-the-art methods and extensive ablation studies are also carried out to verify the effectiveness of each component in our framework.

4.1 Datasets

Cityscapes is one of the benchmarks for urban scene understanding, which is collected from 50 cities with varying scene layouts and weather conditions. The 5,000 finely-annotated images from this dataset are used in our study, which contains 2,975 training images, 500 validation images, and 1,525 test images. Each image with a resolution of 2048 $\times$ 1024. The GTA5 dataset is synthesized from

Table 1. A performance comparison of our method with other state-of-the-art models on "GTA5 to Cityscapes". The performance is measured by the intersection-over-union (IoU) for each class and mean IoU (mIoU). Two base architectures, i.e., VGG16 (V) and ResNet101 (R) are used in our study.

GTA5 → Cityscapes																					
	Base	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bicycle	**mIoU**
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
SIBAN [34]	R	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
CLAN [35]	R	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 [3]	R	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
IntraDA [38]	R	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
BDL [28]	R	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
CrCDA [22]	R	92.4	**55.3**	82.3	31.2	29.1	32.5	33.2	35.6	83.5	34.8	84.2	58.9	32.2	84.7	40.6	46.1	2.1	31.1	32.7	48.6
SIM [51]	R	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
Kim et al. [25]	R	**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-MBT [54]	R	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**	**50.45**
Ours	R	90.8	41.4	84.7	**35.1**	27.5	31.2	38.0	32.8	85.6	42.1	84.9	59.6	**34.4**	85.0	**42.8**	52.7	3.4	30.9	38.1	49.5
Source only	V	26.0	14.9	65.1	5.5	12.9	8.9	6.0	2.5	70.0	2.9	47.0	24.5	0.0	40.0	12.1	1.5	0.0	0.0	0.0	17.9
SIBAN [34]	V	83.4	13.0	77.8	20.4	17.5	24.6	22.8	9.6	81.3	29.6	77.3	42.7	10.9	76.0	22.8	17.9	5.7	14.2	2.0	34.2
ASN [45]	V	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
CyCADA [1]	V	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
CLAN [35]	V	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
CrDoCo [9]	V	89.1	33.2	80.1	26.9	25.0	18.3	23.4	12.8	77.0	29.1	72.4	**55.1**	20.2	79.9	22.3	19.5	1.0	20.1	18.7	38.1
CrCDA [22]	V	86.8	37.5	80.4	30.7	18.1	26.8	25.3	15.1	81.5	30.9	72.1	52.8	19.0	82.1	25.4	29.2	10.1	15.8	3.7	39.1
BDL [28]	V	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 [54]	V	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
Kim et al. [25]	V	**92.5**	**54.5**	**83.9**	**34.5**	**25.5**	**31.0**	30.4	18.0	**84.1**	**39.6**	**83.9**	53.6	19.3	81.7	21.1	13.6	17.7	12.3	6.5	42.3
SIM [51]	V	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
Ours	V	90.1	41.2	82.2	30.3	21.3	18.3	**33.5**	23.0	**84.1**	37.5	81.4	54.2	**24.3**	**83.0**	**27.6**	32.0	8.1	**29.7**	**26.9**	**43.6**

the game Grand Theft Auto V (GTA5), including a total of 24,966 labeled images whose annotations are compatible with Cityscapes. The resolution of each image is 1914 × 1052. The SYNTHIA-RAND-CITYSCAPES (or SYNTHIA for short) contains 9,400 pixel-level annotated images (1280 × 760), which are synthesized from a virtual city. Following the same setting reported in the previous studies, we use the labeled SYNTHIA or GTA5 dataset as the source domain, while using the unlabeled training dataset in the CITYSCAPES as the target domain. Only the 500 labeled validation images from CITYSCAPES are used as test data in all of our experiments.

4.2 Network Architecture

We use two segmentation baseline models, i.e., FCN-VGG16 and DeepLab-ResNet101 to investigate the effectiveness and generalizability of our framework. Specifically, FCN-VGG16 is the combination of FCN-8s [32] and VGG16 [44], while DeepLab-ResNet101 is obtained by integrating DeepLab-V2 [6] into ResNet101 [20]. These two segmentation models share the same discriminator which has 5 convolution layers with channel number 64,128, 256, 512, 1. For each layer, a leaky ReLU parameterized by 0.2 is followed, except the last one. The kernel size and stride are set to 4×4 and 2, respectively. The reconstruction model follows the architecture in [24], containing 3 convolution layers (kernel

Table 2. A performance comparison of our method with other state-of-the-art models on "SYNTHIA to Cityscapes". The performance is measured by the IoU for each class and mIoU. Two base architectures, i.e., VGG16 (V) and ResNet101 (R) are used in our study.

SYNTHIA → Cityscapes																		
	Base	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	sky	person	rider	car	bus	motorbike	bicycle	**mIoU**
Source only	R	55.6	23.8	74.6	—	—	—	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	38.6
ASN [45]	R	84.3	42.7	77.5	—	—	—	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	46.7
DISE [3]	R	91.7	**53.5**	77.1	—	—	—	6.2	7.6	78.4	81.2	55.8	19.2	82.3	30.3	17.1	34.3	48.8
IntraDA [38]	R	84.3	37.7	79.5	—	—	—	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	48.9
Kim et al. [25]	R	**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
DADA [49]	R	89.2	44.8	**81.4**	—	—	—	8.6	11.1	**81.8**	84.0	54.7	19.3	79.7	40.7	14.0	38.8	49.8
CrCDA [22]	R	86.2	44.9	79.5	—	—	—	9.4	11.8	78.6	**86.5**	57.2	26.1	76.8	39.9	21.5	32.1	50.0
BDL [28]	R	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 [51]	R	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 [54]	R	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	R	85.1	44.5	81.0	—	—	—	16.4	15.2	80.1	84.8	59.4	31.9	73.2	41.0	32.6	44.7	**53.1**
CrCDA [22]	V	74.5	30.5	78.6	6.6	**0.7**	21.2	2.3	8.4	77.4	79.1	45.9	16.5	73.1	24.1	9.6	14.2	35.2
ROAD-Net [8]	V	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
SPIGAN [27]	V	71.1	29.8	71.4	3.7	0.3	**33.2**	6.4	15.6	81.2	78.9	52.7	13.1	75.9	25.5	10.0	20.5	36.8
GIO-Ada [7]	V	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
TGCF-DA [10]	V	**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
BDL [28]	V	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
FDA-MBT [54]	V	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
Ours	V	73.7	29.6	77.6	1.0	0.4	26.0	**14.7**	**26.6**	80.6	81.8	**57.2**	**24.5**	76.1	**27.6**	13.6	**46.6**	**41.1**

3×3 and stride 1), 9 ResNet blocks (kernel 3×3 and stride 2), and another 3 transposed convolution layers (kernel 3×3 and stride 2) for upsampling. The 3 multi-scale discriminators share the identical network, each of which follows the architecture of PatchGAN [23]. More details regarding the architecture of discriminators in both segmentation and reconstruction models can be found in the **Supplementary**.

4.3 Implementation Details

Our framework is implemented with PyTorch [39] on two TITAN Xp GPUs, each of which with 12GB memory. The batch size is set to one for training all the models discussed above. Limited by the GPU memory space, the translation network is first trained to perform target-to-source image translation by using Adam optimizer [26]. The initial learning rate is set to 0.0001, which is reduced by half after every 100,000 iterations. We use momentum {0.5, 0.999} with weight decay 0.0001. The maximum training iteration is $1000k$.

DeepLab-ResNet101 is trained using Stochastic Gradient Descent optimizer with initial learning rate 2.5×10^{-4}. The polynomial decay with power 0.9 is applied to the learning rate. The momentum and weight decay are set to 0.9 and 5×10^{-4}, respectively. For FCN-VGG16, the Adam optimizer with momentum {0.9, 0.99} and initial learning rate 1×10^{-5} is used for training. The learning rate is decreased using step decay with step size 50000 and drop factor 0.1. In equation 1, λ is set to 1×10^{-3} for DeepLab-ResNet101 and 1×10^{-4} for FCN-VGG16.

Table 3. Ablation study of the target-to-source translation and the reconstruction network. S → T and T → S indicate source-to-target and target-to-source translation.

Base	S → T	T → S	Reconstruction	GTA5	SYNTHIA
R	✓			48.5	51.4
R		✓		49.1	52.0
R		✓	✓	49.5	53.1
V	✓			41.3	39.0
V		✓		42.3	40.1
V		✓	✓	43.6	41.1

The reconstruction network is first pre-trained by reconstructing source images $\mathcal{X}_s$ from the corresponding labels Y_s. We use the Adam optimizer with initial learning rate 2×10^{-4} and momentum $\{0.5, 0.999\}$, where the learning rate is linearly decreased to zero. In equation 6, we set $\beta = 10$. α is set to 0.01 and 0.001 for FCN-VGG16 and DeepLab-ResNet101 respectively.

4.4 GTA5 → Cityscapes

We carry out the adaptation from GTA5 to Cityscapes by following the same evaluation protocol as previously reported in [45, 28]. The overall quantitative performance is assessed on 19 common classes (e.g., road, wall, and car) between these two domains. As shown in Table 1, we demonstrate competitive performance against ResNet101-based methods, but are inferior to two newly published models [25, 54]. For the VGG16-based backbone, however, we are able to achieve the best results compared to existing state-of-the-art methods including [25, 54]. Specifically, our method surpasses the source-only model (without adaptation) by 12.9% and 25.7% on ResNet101 and VGG16, respectively. Compared with CyCADA [1] and BDL [28] that rely on source-to-target translation, we demonstrate significant improvements (i.e., 8.2% and 2.3% on VGG16) by reducing image translation bias. CLAN [35] aims to enforce local semantic consistency by a category-level adversarial network. However, such a strategy fails to account for the global semantic consistency. Our reconstruction network shares a similar spirit with CLAN in terms of joint distribution alignment but enables us to enforce semantic consistency from a global view. As a consequence, we get 6.3% and 7.0% improvement on ResNet101 and VGG16, respectively.

4.5 SYNTHIA → Cityscapes

We then evaluate our framework on the adaptation from SYNTHIA to Cityscapes based on 13 classes on ResNet101 and 16 classes on VGG16. The results exhibit that our method outperforms other competing methods on average as shown in Table 2. Both ASN [45] and BDL [28] adapt output space in their models. However, simply aligning segmentation outputs may lead to negative transfer

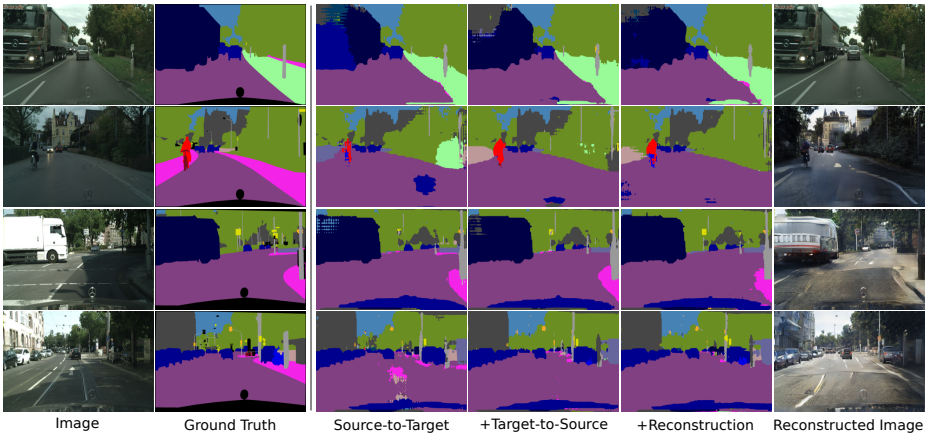

Fig. 5. Qualitative examples of semantic segmentation results in Cityscapes. For each target-domain image (first column), its ground truth and the corresponding segmentation prediction from the baseline model (source-to-target translation) are given. The following are predictions of our method by incorporating target-to-source translation and reconstruction, together with the reconstructed image.

Table 4. Ablation study of the temperature τ on GTA5 $\rightarrow$ Cityscapes.

τ	0.0001	0.001	0.01	0.1	1
mIoU	42.7	**43.6**	42.8	42.9	41.5

issue, owing to the dramatic differences of the layout and structure between SYNTHIA and Cityscapes. We achieve 6.4% and 1.7% improvement than ASN and BDL on ResNet101, respectively. It is noteworthy that we also outperform [54] on both ResNet101 and VGG16-based backbone.

4.6 Ablation Study

Target-to-source Translation and Reconstruction For GTA5 to Cityscapes, 0.6% improvement is achieved by considering target-to-source translation on ResNet101 compared to the source-to-target translation model (Table 3). By further enforce semantic consistency through a reconstruction network, our method achieves 49.5 mIoU. Similar improvements are also observed on VGG16, with 1.0% improvement by performing target-to-source translation. The prediction power of our method is further boosted by combining translation and reconstruction, giving rise to another 1.3% mIoU improvement. The qualitative study of each module in our method is showcased in Figure 5.

Table 5. Ablation study of the feature space vs. label space reconstruction.

	Feature space	Label space
GTA5 → Cityscapes	41.48	43.6
SYNTHIA → Cityscapes	40.13	41.1

For SYNTHIA to Cityscapes, we achieve a performance boost of 0.6% and 1.1% by considering target-to-source translation on ResNet101 and VGG16, respectively (Table 3). The performance gain is 1.1% and 1.0% by incorporating the reconstruction network. Our results prove the effectiveness of target-to-source translation and reconstruction in adapting domain knowledge for semantic segmentation.

Table 6. Ablation study of the reconstruction loss on GTA5 → Cityscapes with VGG16 backbone.

VGG	Discriminator	mIoU
		41.53
✓		42.82
	✓	41.95
✓	✓	43.6

Parameter Analysis We investigate the sensitivity of temperature parameter τ in this section and find that $\tau = 0.001$ achieves the best performance (Table 4). Therefore, τ is set to 0.001 in all of our experiments to approximate semantic label maps.

Feature Space VS. Label Space Reconstruction We also evaluate the feature space reconstruction based on the VGG16-based backbone. Table 5 highlights the benefits of our label-driven reconstruction that enforces semantic consistency of target images and their predicted labels.

Reconstruction loss Table 6 shows the complementary role of VGG perceptual loss and discriminator feature matching loss (equation 5) in maintaining semantic consistency.

5 Conclusion

We propose a novel framework that exploits cross-domain adaptation in the context of semantic segmentation. Specifically, we translate images from the target domain to the source domain to reduce image translation bias and the computational cost. To enforce cross-domain features with the same category close to each other, we reconstruct both source and target images directly from the label space. Experiments demonstrate that our method achieves significant improvement in adapting from GTA5 and SYNTHIA to Cityscapes.

Acknowledgments This work was partially supported by US National Science Foundation IIS-1718853, the CAREER grant IIS-1553687 and Cancer Prevention and Research Institute of Texas (CPRIT) award (RP190107).

References

1. Cycada: Cycle consistent adversarial domain adaptation. In: International Conference on Machine Learning (ICML) (2018)
2. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: Advances in neural information processing systems (NIPS). pp. 343–351 (2016)
3. Chang, W.L., Wang, H.P., Peng, W.H., Chiu, W.C.: All about structure: Adapting structural information across domains for boosting semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1900–1909 (2019)
4. Chen, L.C., Collins, M.D., Zhu, Y., Papandreou, G., Zoph, B., Schroff, F., Adam, H., Shlens, J.: Searching for efficient multi-scale architectures for dense image prediction. In: Advances in neural information processing systems (NIPS) (2018)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: International Conference on Learning Representations (ICLR) (2014)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence (TPAMI) **40**(4), 834–848 (2018)
7. Chen, Y., Li, W., Chen, X., Gool, L.V.: Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1841–1850 (2019)
8. Chen, Y., Li, W., Van Gool, L.: Road: Reality oriented adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7892–7901 (2018)
9. Chen, Y.C., Lin, Y.Y., Yang, M.H., Huang, J.B.: Crdoco: Pixel-level domain transfer with cross-domain consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1791–1800 (2019)
10. Choi, J., Kim, T., Kim, C.: Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In: Proceedings of the IEEE international conference on computer vision (ICCV) (2019)
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3213–3223 (2016)
12. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 1635–1643 (2015)
13. Du, L., Tan, J., Yang, H., Feng, J., Xue, X., Zheng, Q., Ye, X., Zhang, X.: Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision (ICCV) (2019)
14. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
15. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning (ICML) (2015)

16. Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 597–613 (2016)
17. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 1440–1448 (2015)
18. Goodfellow, I.: Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160 (2016)
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems (NIPS). pp. 2672–2680 (2014)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
21. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016)
22. Huang, J., Lu, S., Guan, D., Zhang, X.: Contextual-relation consistent domain adaptation for semantic segmentation. Proceedings of the European Conference on Computer Vision (ECCV) (2020)
23. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1125–1134 (2017)
24. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 694–711 (2016)
25. Kim, M., Byun, H.: Learning texture invariant representation for domain adaptation of semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12975–12984 (2020)
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2014)
27. Lee, K.H., Ros, G., Li, J., Gaidon, A.: Spigan: Privileged adversarial learning from simulation. In: International Conference on Learning Representations (ICLR) (2019)
28. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
29. Lian, Q., Lv, F., Duan, L., Gong, B.: Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In: Proceedings of the IEEE international conference on computer vision (ICCV) (2019)
30. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in neural information processing systems (NIPS). pp. 700–708 (2017)
31. Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X.: Semantic image segmentation via deep parsing network. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 1377–1385 (2015)
32. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440 (2015)
33. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning (ICML) (2015)

34. Luo, Y., Liu, P., Guan, T., Yu, J., Yang, Y.: Significance-aware information bottleneck for domain adaptive semantic segmentation. In: Proceedings of the IEEE international conference on computer vision (ICCV) (October 2019)
35. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: 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 (CVPR). pp. 2507–2516 (2019)
36. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4500–4509 (2018)
37. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 13 (2018)
38. Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S.: Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3764–3773 (2020)
39. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
40. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1713–1721 (2015)
41. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 102–118 (2016)
42. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: 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 (CVPR). pp. 3234–3243 (2016)
43. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R.: Learning from synthetic data: Addressing domain shift for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
44. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
45. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
46. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 4068–4076 (2015)
47. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
48. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
49. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Dada: Depth-aware domain adaptation in semantic segmentation. In: Proceedings of the IEEE international conference on computer vision (ICCV) (2019)

50. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
51. Wang, Z., Yu, M., Wei, Y., Feris, R., Xiong, J., Hwu, W.m., Huang, T.S., Shi, H.: 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 (CVPR). pp. 12635–12644 (2020)
52. Wu, Z., Han, X., Lin, Y.L., Gokhan Uzunbas, M., Goldstein, T., Nam Lim, S., Davis, L.S.: Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 518–534 (2018)
53. Yang, J., An, W., Yan, C., Zhao, P., Huang, J.: Context-aware domain adaptation in semantic segmentation. arXiv preprint arXiv:2003.04010 (2020)
54. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4085–4095 (2020)
55. Ying, W., Zhang, Y., Huang, J., Yang, Q.: Transfer learning via learning to transfer. In: International Conference on Machine Learning (ICML). pp. 5072–5081 (2018)
56. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE international conference on computer vision (ICCV) (2017)
57. Zhang, Y., Wei, Y., Wu, Q., Zhao, P., Niu, S., Huang, J., Tan, M.: Collaborative unsupervised domain adaptation for medical image diagnosis. IEEE Transaction on Image Processing (TIP) (2020)
58. Zhang, Y., Qiu, Z., Yao, T., Liu, D., Mei, T.: Fully convolutional adaptation networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6810–6818 (2018)
59. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2881–2890 (2017)
60. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision (ICCV) (2017)
61. Zhu, X., Zhou, H., Yang, C., Shi, J., Lin, D.: Penalizing top performers: Conservative loss for semantic segmentation adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 568–583 (2018)
62. Zou, Y., Yu, Z., Kumar, B.V., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 297–313. Springer (2018)