This CVPR 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Semi-supervised Domain Adaptation based on Dual-level Domain Mixing for Semantic Segmentation

Shuaijun Chen[1†], Xu Jia[2†‡], Jianzhong He[1,3], Yongjie Shi[1,4], Jianzhuang Liu[1]
[1]Noah's Ark Lab, Huawei Technologies. [2]Dalian University of Technology.
[3]Data Storage and Intelligent Vision Technical Research Dept, Huawei Cloud.
[4]Key Lab of Machine Perception, Peking University
{chenshuaijun,jianzhong.he,shiyongjie2,liu.jianzhuang}@huawei.com,xjia@dlut.edu.cn

Abstract

Data-driven based approaches, in spite of great success in many tasks, have poor generalization when applied to unseen image domains, and require expensive cost of annotation especially for dense pixel prediction tasks such as semantic segmentation. Recently, both unsupervised domain adaptation (UDA) from large amounts of synthetic data and semi-supervised learning (SSL) with small set of labeled data have been studied to alleviate this issue. However, there is still a large gap on performance compared to their supervised counterparts. We focus on a more practical setting of semi-supervised domain adaptation (SSDA) where both a small set of labeled target data and large amounts of labeled source data are available. To address the task of SSDA, a novel framework based on dual-level domain mixing is proposed. The proposed framework consists of three stages. First, two kinds of data mixing methods are proposed to reduce domain gap in both region-level and sample-level respectively. We can obtain two complementary domain-mixed teachers based on dual-level mixed data from holistic and partial views respectively. Then, a student model is learned by distilling knowledge from these two teachers. Finally, pseudo labels of unlabeled data are generated in a self-training manner for another few rounds of teachers training. Extensive experimental results have demonstrated the effectiveness of our proposed framework on synthetic-to-real semantic segmentation benchmarks.

1. Introduction

Semantic segmentation with the goal of assigning semantic-level labels to every pixel in an image is one of the fundamental topics in computer vision due to its widely critical real-world applications, such as autonomous driving [11] and robotic navigation [28, 39]. Over the past few years, deep convolutional neural networks(CNNs) have achieved dramatic improvements in semantic segmentation [1, 25, 17, 26, 2, 48]. The success of CNN-based methods benefits from large volume of manually labeled data [24, 8], and the assumption of independent and identical data distribution between training and testing samples. However, performance drops significantly when the model trained on training set (source domain) is directly applied to unseen test scenarios (target domain). In addition, densely annotating pixel-wise labels of many samples in target domain is time-consuming and uneconomical.

To reduce the heavy demand for pixel-wise annotation, one way is to employ large amounts of easy-to-get simulation data which can be collected from game engines such as GTA5 [33] and SYNTHIA [34]. In addition, unsupervised domain adaptation (UDA) strategy, which aims at transferring knowledge from a synthetic label-rich source domain to a real-world label-scarce target domain, is required to bridge domain gap between synthetic and real-world domains. Impressive results have been achieved by UDA methods that extract domain-invariant representations via entropy minimization [31, 43], generative modelling [16, 12] and adversarial learning [42, 41]. However, domain shift cannot be completely eliminated by these methods due to weak supervision on target examples. There is still a big gap in performance compared with supervised methods. Another way in addressing the issue of heavy annotation is to annotate only a small set of images from target domain and make full use of plenty of unlabeled data with semi-supervised learning (SSL) techniques [10, 30, 9, 29]. Due to the shortage of labeled data in SSL setting, the obtained model has the risk of overfitting to the small amount of labeled data. How to effectively utilize available unlabeled and limited labeled data from different domains is the key in improving model's accuracy and generalization for pixel-wise prediction tasks.

Hence, a more practical task of semi-supervised domain

[†]Equal contribution
[‡]Part of this work was done while he was in Noah's Ark Lab

adaptation (SSDA) is recently introduced by combining the small set of labeled target data images in SSL with the large amounts of labeled source domain data and unlabeled target domain data. In order to address the SSDA problem, one naive way is to equip UDA methods with additional supervision on the extra labeled target images (see Baseline in Table 1.). For example, Alleviating Semantic-level Shift (ASS) model [44] is proposed for better promoting the distribution consistency of features by using adversarial learning on outputs from two labeled domains. However, these methods cannot fully explore rich information within available labeled and unlabeled data in two domains.

Semantic segmentation is a dense pixel-wise prediction task, and classification of one pixel depends not only on its own value but also on its neighbourhood's context. We focus on how to effectively utilize labeled data in different domains to extract domain-invariant representations in both region-level and sample-level. The proposed framework consists of three steps. First, two kinds of data mixing methods are proposed to reduce domain gap in both region-level and sample-level. The region-level data mixing is achieved by applying two binary masks to labeled images from two domains and combining the two masked regions, which encourages a model to extract domain-invariant features about semantic structure from partial view. On the other hand, the image-level data mixing directly mixes labeled images from two domains from holistic view. Such two mixing ways help train two complementary teacher models that work on both two kinds of data distribution. In the second step, we employ knowledge distillation technique to extract "dark knowledge" from these two complementary teachers, which works as guidance in the learning process of a student model for target domain. By integrating knowledge from two views and making full use of unlabeled data, the student model of the same network architecture can give even better performance than any of its teachers. Once a good student model for target domain is obtained, pseudo labels could be generated with self-training strategy to expand the set of labeled target domain data for iterative update. Compared with traditional self-training methods, which directly use pseudo labels to train a final model, we instead leverage these pseudo labels to obtain two stronger domain-mixed teachers, which also leads to stronger student by another round of knowledge distillation. Overall, in our framework, teachers and student are progressively growing, and we can obtain a final well-trained student model.

Our contributions of this paper are three-fold:

- Two kinds of data mixing methods are proposed to train domain-mixed teachers across domains in both region-level and sample-level to alleviate data distribution mismatch between different domains.
- A stronger student model on target domain can be obtained by distilling knowledge from complementary domain-mixed teachers. It can be further strengthened by employing pseudo labels which are generated for unlabeled target data in a self-training manner.
- Extensive experiments demonstrate that the proposed method can achieve superior performance on two common synthetic-to-real semantic segmentation benchmarks with small amounts of labeled data.

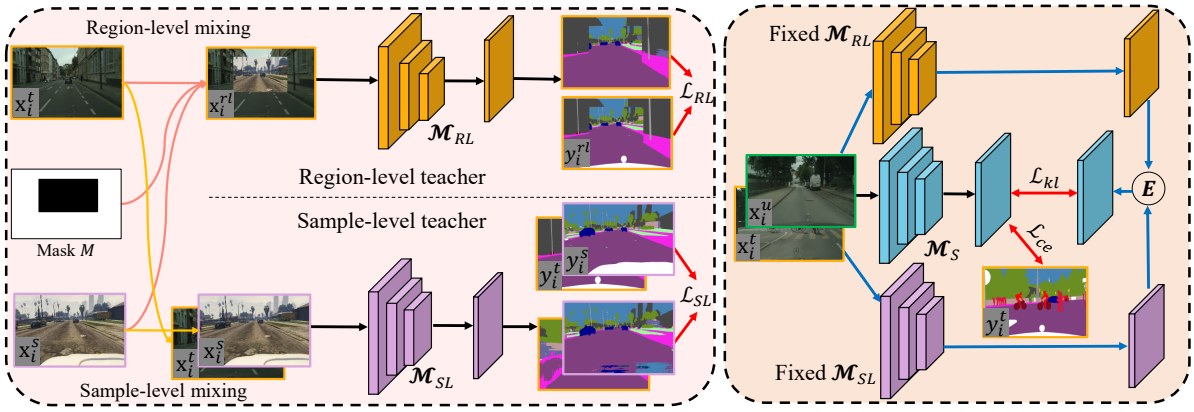
2. Related Works

Unsupervised domain adaptation for semantic segmentation. Unsupervised domain adaptation (UDA) methods for semantic segmentation have been extensively studied to address domain discrepancy between photo-realistic synthetic dataset and unlabeled real dataset. One mainstream approach is by adversarial learning [42, 41, 6, 5, 17, 37, 19], which aims to employ a discriminator to measure the divergence across two domains. Another approach to solve UDA problem is to utilize generative networks [38, 16, 49] to generate target-style images by applying style transfer technique on annotated source image. Some methods based on self-training [21, 50, 23, 14] have been employed to generate pseudo labels of unlabeled data and use them to fine-tune the model. [21] firstly generates different stylized annotated images to learn texture-invariant representation and then uses self-training to generate pseudo labels of unlabeled data to fine-tune the model on target domain.

Although impressive results have been achieved in UDA for semantic segmentation, the domain gap cannot be fully alleviated due to the lack of strong supervision in the target domain, and there is still an observed performance gap compared with their supervised counterparts.

Semi-supervised learning for semantic segmentation. One way to reduce the heavy demand for manual pixel-wise labeling is to only label a small amount of data from target distribution and adopt semi-supervised learning (SSL) strategy to learn a great generalized model among ample unlabeled and limited labeled data. Numerous methods have since been developed to improve model generalization [30, 20, 9, 29, 18, 4, 13]. Consistency regularization is one of the most popular methods and the key idea is to encourage the network to give consistent predictions for perturbed unlabeled inputs. One most related work is [10], which enforces a consistency between mixed output of teacher network and the prediction of student over the mixed inputs by a region-level data augmentation CutMix [46] with a teacher-student architecture [40]. Our method also shares similar philosophy as theirs, however, we propose to train two domain-mixed teachers with two kinds of domain-mixing methods to fully exploit two sets of data from two different domains.

Semi-supervised domain adaptation. Also aims to reduce the data distribution mismatch, compared with UDA, semi-supervised domain adaptation (SSDA) bridges domain dis-



(a) Domain-mixed Teachers based on Dual-level Domain Mixing (b) Multi-teacher Knowledge Distillation

Figure 1. The first two stages of the proposed SSDA framework, training of domain-mixing teachers and multi-teacher knowledge distillation. The domain-mixed teachers are trained based on the dual-level mixed data. Then these two domain-mixed teachers are used to train a good student. The student will generate pseudo labels for next round of teachers training. E means the ensemble operation of two domain-mixed teachers. The black arrows represent the training data flow, blue arrows represent the data flow of inference, which do not require gradient backward. The red arrows represent the computation of losses.

crepancy via introducing partially labeled target samples. Recently, a few methods have been proposed based on deep learning [45, 32, 22, 35] for image classification. [45] decomposes SSDA into two sub-problems: UDA and SSL, and employs co-training [3] to exchange the expertise between two classifiers, which are trained on MixUp-ed [47] data between labeled and unlabeled data of each view.

Due to the complex densely pixel-wise prediction and no explicit decision boundaries between examples in semantic segmentation, SSDA methods based on discriminative class boundaries for image classification cannot be directly applied to semantic segmentation. Just one previous work have been developed to study SSDA for semantic segmentation. Wang *et al.* [44] propose Alleviating Semantic-level Shift (ASS) framework to realize feature alignment across domain from global and semantic level. ASS introduces an extra semantic-level adaptation module through adversarial training on the corresponding outputs of source and target labeled inputs besides the additional supervision on extra small amount of labeled target data upon the classical AdaptSeg framework [41]. However, the naive supervision of labeled target samples cannot fully take advantage of labeled two domains, and the adversarial loss makes training unstable due to the weak supervision. To solve this issue, we propose a novel iterative framework based on dual-level domain mixing methods without any adversarial training.

3. Method

3.1. Problem Statement

In the setting of semi-supervised domain adaptation (SSDA), we are provided with a small set of labeled target domain images upon the large amounts of labeled source and unlabeled target domain images. Let $D_S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{N_S}$ represents the N_S labeled source domain samples, and $D_T = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{N_T}$ represents the N_T labeled target domain samples, and $D_U = \{\mathbf{x}_i^u\}_{i=1}^{N_U}$ represents the N_U unlabeled target domain samples. With the SSDA setting, we aim at developing a way to efficiently utilize the available D_S, D_T and D_U and obtain a segmentation model which has great performance on unseen test data sampled from target data distribution.

3.2. Domain-mixed Teachers

Performance degradation comes from inconsistent data distribution in different domains. We propose two data mixing methods for domain adaptation, one is region-level data mixing and another is sample-level data mixing, to reduce the data distribution gap from two views. As we all know, data with labeled ground truth provides much information for training one model in deep learning-based methods. In SSDA, two types of labeled data, *i.e.*, D_S, D_T, are provided. Our region-level and sample-level data mixing methods are implemented on these two kinds of labeled data, and two domain-mixed teacher models can be trained on the mixed data. Because of different views of data mixing, these two domain-mixed teachers are complementary.

Region-level data mixing. Semantic segmentation is a dense pixel-wise prediction task, and the classification of one pixel depends not only on its own value but also on its regional neighbourhood's context. Thus, if one image contains both source domain and target domain content, the model can learn domain-invariant representation because different regions with different feature distribution can be seen at the same time during model training.

Inspired by CutMix [46] where patches from an image are cut and pasted to another one to augment data for improving model's generalization ability, here we pro-

pose to conduct region-level data mixing on set D_S and D_T to reduce domain gap. Given two labeled images $\{\mathbf{x}^t, y^t\}$, $\{\mathbf{x}^s, y^s\}$, the region-level mixing operation can be described as below.

$$\begin{aligned} \mathbf{x}^{rl} &= M \odot \mathbf{x}^t + (1 - M) \odot \mathbf{x}^s, \\ y^{rl} &= M \odot y^t + (1 - M) \odot y^s, \end{aligned} \tag{1}$$

where M denotes a binary mask indicating where the region needs to fusion, and $\odot$ is element-wise multiplication. As shown in Fig. 1, the mixed image $\mathbf{x}^{rl}$ contains both contents of $\mathbf{x}^s$ and $\mathbf{x}^t$, and the corresponding mixed labels y^{rl} are obtained for each pixel according to which domain the region containing that pixel comes from. In detail, a rectangular region is cropped from $\mathbf{x}^s$ according to randomly chosen coordinates, and then pasted on the same location of $\mathbf{x}^t$. The region-level data mixing is able to produce intermediate samples between different domains, which works as a bridge, filling in the gap between domains. This helps explore essential semantic contexts across different domains from partial view. Additionally, this operation can destroy the inherent structure of the original target picture, and regularize the training process of region-level teacher. Once the mixed images and their labels are ready, we can train a semantic segmentation model through supervised training on the mixed data. The training objective function can be written as follows.

$$\mathcal{L}_{RL} = \mathcal{L}_{ce}(\mathcal{M}_{RL}(\mathbf{x}^{rl}), y^{rl}), \tag{2}$$

where $\mathcal{M}_{RL}$ represents teacher model trained on region-level mixed data, $\mathcal{L}_{ce}$ denotes the cross entropy loss.

Sample-level data mixing. Sample-level data mixing aims to mix the data from different domains from holistic view. The source and target examples are sampled from inconsistent distribution with a big gap. We find that direct mixing of these data can already help reduce the gap between different domains to some extent. There are two advantages with sample-level data mixing method. On the one hand, the introduction of large amounts of source images alleviates the model overfitting to the small amount of target images. On the other hand, the sample-level mixing helps explore intermediate decision boundary between different domains from holistic view. In our experiments, we randomly sample two examples from source set D_S and target set D_T, then directly feed both of them into model during one iteration. Given two images from D_S and D_T, the training objective function of sample-level teacher is defined as follows.

$$\mathcal{L}_{SL} = \mathcal{L}_{ce}(\mathcal{M}_{SL}(\mathbf{x}^s), y^s) + \mathcal{L}_{ce}(\mathcal{M}_{SL}(\mathbf{x}^t), y^t), \tag{3}$$

where $\mathcal{M}_{SL}$ represents teacher model trained on sample-level mixed data.

3.3. Multi-teacher Knowledge Distillation

After obtaining two pre-trained domain-mixed teachers, we employ knowledge distillation (KD), a technique to distilling knowledge by minimizing the KL-divergence between outputs of these two models. Here we adapt it to extract "dark knowledge" from these two complementary teachers. The pipeline of multi-teacher KD is shown in Fig. 1 (b), including two pre-trained domain-mixed teachers and one student with the same network architecture as teacher. The outputs of two teachers are ensembled as a stronger guidance to supervise the training of the student model on unlabeled target data. Besides, the student model is also supervised by the labels on the small amount of labeled target data. The objective function of learning student $\mathcal{M}_S$ is defined as below.

$$\begin{aligned} \mathcal{L}_S =& \lambda_{kl}\mathcal{L}_{kl}(E(\mathcal{M}_{RL}(\mathbf{x}^u), \mathcal{M}_{SL}(\mathbf{x}^u)), \mathcal{M}_S(\mathbf{x}^u)) \\ &+ \lambda_{ce}\mathcal{L}_{ce}(\mathcal{M}_S(\mathbf{x}^t), y^t), \end{aligned} \tag{4}$$

where λ_{kl} and λ_{ce} are the weights of KL-divergence loss and cross entropy loss respectively, E denotes the ensemble operation of two models. In experiments, the ensemble operation is implemented by averaging the outputs of two complementary teachers.

By integrating knowledge from two views and making full use of unlabeled data, we can obtain one student with even superior performance than any one of its teachers.

Algorithm 1 Training process of our proposed framework.

Require: labeled source dataset $D_S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{N_S}$, unlabeled target dataset $D_U = \{\mathbf{x}_i^u\}_{i=1}^{N_U}$, labeled target dataset $D_T = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{N_T}$, initialized weights of teachers model $\mathcal{M}_{RL}^0$, $\mathcal{M}_{SL}^0$ and student model $\mathcal{M}_S^0$, iterative rounds R.

Procedure:

```
for r ← 1 to R do
    Dual-level domain mixing
    Optimize M_RL^r and M_SL^r by Eq. (2) and (3)   ▷ Training two teachers
    Optimize M_S^r by Eq. (4) ▷ Training student model
    Generate pseudo labels ŷ_i^u following [23] by M_S^r
    Update D_U = {x_i^u, ŷ_i^u}_{i=1}^{N_U} to labeled target dataset D_T
end for
return student model M_S^R
```

3.4. Progressive Improving Scheme

Normally, a teacher network usually has stronger ability than student network. However, here a good student model is obtained by distilling knowledge from the ensembled outputs of two complementary domain-mixed models on large amount of unlabeled data. We focus on how to use a student to further improve teachers' performance for next step.

Recently, self-training as a simple but effective technique to address the scarceness of labeled training data, and are widely applied in SSL and UDA for image classification

task. In our task, the teachers we obtain are trained based just on labeled source data and a small amount of labeled target data. Motivated by the success of self-training, we believe the teachers can be further improved with this strategy. In detail, following [23], pseudo labels of D_U are generated via the learned student model to update the labeled set of images in dataset of D_T for next round training of domain-mixed teachers. Once stronger domain-mixed teachers are obtained, a stronger student can be obtained by another round of multi-teacher KD.

Overall, the whole training process of our framework goes iteratively. Both the domain-mixed teachers and student are progressively growing, *i.e.*, they can help the learning of each other through knowledge distillation and self-training strategies. We summarize our proposed algorithm in Algorithm 1.

4. Experiments

4.1. Experimental Setup

Following the setting of unsupervised domain adaptation methods in semantic segmentation, we also conduct extensive experiments and report the mean intersection-over-union (mIoU) score on two commonly used synthetic-to-real benchmarks, which are GTA5 [33] and SYNTHIA [34] to Cityscapes [7] respectively.

Cityscapes is an autonomous driving dataset captured from 50 cities in real world. It contains densely annotated 2,975 and 500 images with a fixed resolution of 2048×1024 for training and validation respectively. All images are manually labeled by 19 semantic categories. For SSDA setting, we randomly select different numbers of images, such as (100, 200, 500, 1000), from the whole training set to demonstrate the effectiveness of our method across different settings. The validation set is used to evaluate the performance of our method.

GTA5 is a synthetic dataset in which the images are collected from game video and the corresponding semantic labels are automatically generated by computer graphics techniques. It includes 24,966 synthesized images with pixel-wise labels of 33 classes. In experiments, we consider the 19 common classes with Cityscapes dataset to train our models.

SYNTHIA is also a synthetic dataset and we use SYNTHIA-RAND-CITYSCAPES as another labeled source domain, which contains 9,400 fully annotated synthetic images with resolution of 1280×960. It has 16 common categories with Cityscapes dataset. We train our models with the common classes and report the 13-class mIoU on validation set.

Table 1. Semantic segmentation performance comparison with the state-of-the-art UDA, SSL and SSDA methods on GTA5→Cityscapes. 19-class mIoU (%) score are reported on Cityscapes validation set across 0, 100, 200, 500, 1000, 2975 numbers of labeled target images. "*" denotes our reimplementation on corresponding numbers of labeled Cityscapes images. GTA5 images are not introduced for implementing SSL methods. Best results are **highlighted**.

Type	Methods	Labeled target images					
		0	100	200	500	1000	2975
UDA	AdaptSeg [41]	42.4	-	-	-	-	-
	Advent [42]	44.8	-	-	-	-	-
	LTI [21]	50.2	-	-	-	-	-
	PIT [27]	**50.6**	-	-	-	-	-
Supervised	DeeplabV2	-	41.9	47.7	55.5	58.6	65.3
SSL	CutMix* [10]	-	50.8	54.8	61.7	64.8	-
	DST-CBC* [9]	-	48.7	54.1	60.6	63.2	-
SSDA	Baseline	-	52.6	53.6	58.4	61.6	66.6
	MME* [36]	-	52.6	54.4	57.6	61.0	64.2
	ASS [44]	-	54.2	56.0	60.2	64.5	69.1
	Ours	-	**61.2**	**60.5**	**64.3**	**66.6**	**69.8**

4.2. Implementation Details

For all the following experiments, similar to [42], a DeeplabV2 [1] model, which contains Atrous Spatial Pyramid Pooling (ASPP) module to extract multi-scale representations and utilizes a pre-trained ResNet-101 [15] on ImageNet as backbone, is employed as our semantic segmentation architecture. To train our proposed framework, we implement it using Pytorch deep learning toolbox. All the experiments are conducted on a single Tesla V100 GPU with 32GB memory to accelerate computing. We also implement our algorithm on MindSpore*, which is a new deep learning computing platform.

An optional operation before training the model is to apply a simple image translation method to source domain images to reduce the visual difference between source and target domain. Here, following [14], images are converted into LAB color space and are matched to the statistics of target domain. Image translation is applied at the beginning in most experiments except as otherwise noted. Then sample-level and region-level data mixing are conducted on labeled source data with target style and the target data. We then train domain-mixed teachers with cross entropy loss on supervised data. Student model is obtained on both labeled and unlabeled target data with cross entropy and KL loss. The weight λ_{kl}, λ_{ce} in Eq. 4 are set to 0.5 and 1 respectively. For self-training, the portion of selected pseudo labels and the confidence threshold are separately, similar to [23], set to 0.5 and 0.9. Iterative rounds R are set to 4 and 3 for GTA5→Cityscapes and SYNTHIA→Cityscapes respectively. All the models are trained by the Stochastic Gradient Descent (SGD) optimizer with initial learning rate

*https://www.mindspore.cn/

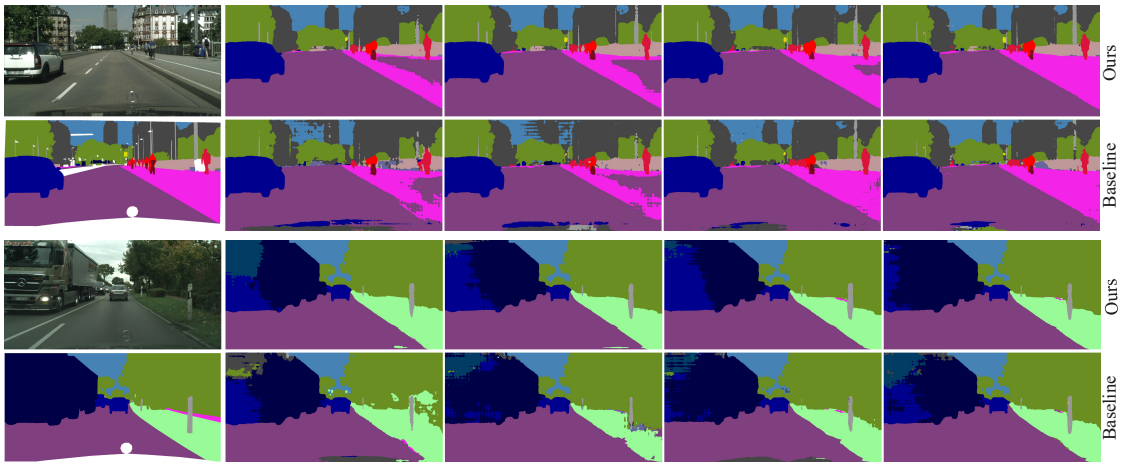


Figure 2. Qualitative results of our method and baseline method on different numbers of labeled target images on GTA5 to Cityscapes. (a) target images and corresponding ground truth (GT), (b)-(e) segmentation results of different numbers of labeled target images.

Table 2. Semantic segmentation performance comparison with the state-of-the-art UDA, SSL and SSDA methods on SYNTHIA→Cityscapes. Here we train the DeeplabV2 model with 16 classes and report 13-class mIoU (%) score following the previous works on UDA. Other settings are kept same as in Table 1. Best results are **highlighted**.

Type	Methods	Labeled target images					
		0	100	200	500	1000	2975
UDA	AdaptSeg [41]	46.7	-	-	-	-	-
	Advent [42]	48.0	-	-	-	-	-
	LTI [21]	49.3	-	-	-	-	-
	PIT [27]	**51.8**	-	-	-	-	-
Supervised	DeeplabV2	-	53.0	58.9	61.0	67.5	72.2
SSL	CutMix* [10]	-	61.3	66.7	71.1	73.0	-
	DST-CBC* [9]	-	59.7	64.3	68.9	70.5	-
SSDA	Baseline	-	58.5	61.9	64.4	67.6	73.1
	MME* [36]	-	59.6	63.2	66.7	68.9	72.7
	ASS [44]	-	62.1	64.8	69.8	73.0	77.1
	Ours	-	**68.4**	**69.8**	**71.7**	**74.2**	**77.2**

2.5×10^{-4}, the momentum 0.9 and weight decay 10^{-4} as mentioned in [42]. The learning rate is decreased with the polynomial annealing procedure with power of 0.9.

4.3. Performance Comparison

Our proposed method is conducted on two common synthetic-to-real GTA5 to Cityscapes and SYNTHIA to Cityscapes benchmarks to demonstrate the effectiveness of proposed framework. The performance is compared with the baseline method and existing state-of-the-art methods on UDA, SSL and SSDA settings. More extensive experiments can be seen in supplementary materials.

Baseline. SSDA aims to alleviate the domain shift problem by introducing extra a small amount of labeled target data compared with UDA setting. As mentioned in Section 1, one naive way to address SSDA problem is by adding additional supervision upon UDA methods. Therefore, here we employ the classical UDA method named AdaptSeg [41], one multi-level adaptation method by adversarial learning on multi-level outputs, with extra supervised cross entropy loss on limited labeled target images as our baseline model.

GTA5 to Cityscapes. The performance comparisons with several state-of-the-art methods on GTA5 to Cityscapes are shown in Table 1. In experiment, iterative round R is set to 4. After the iterative training, our method achieves the best performance on different ratios of labeled target domain images compared with existing methods in UDA, SSL and SSDA settings. Compared with UDA methods such as AdaptSeg, Advent [42], LTI [21], and PIT [27], our method can obtain above 10% performance improvement by labeling just 100 target images and significantly reduce the performance gap compared with the oracle model. Particularly, our method outperforms the SSL methods CutMix [10] and DST-CBC [9], which use related CutMix and self-training techniques respectively, by a large performance gain. ASS [44], to be our known, which is the first work on SSDA for semantic segmentation, employs additional semantic-level adaptation on the outputs of both labeled source and target images to alleviate semantic-level shift except the additional supervision. We modify MME [36], which is used to address image classification in SSDA setting, for semantic segmentation task, and obtain inferior results. We think the reason is that SSDA methods for classification without taking into account the semantic contexts in an image and cannot be directly applied to segmentation task. The proposed approach obtains superior results on all ratios of labeled data. The reason is that the supervision of adversarial learning is weak and we can fully take advantage of available labeled data to reduce domain gap by dual-level data mixing. In addition, our method also performs well on

Table 3. Performance comparison of ensembled models from single view and inconsistent views in terms of per-class IoUs and mIoU (%). The $\mathcal{M}_{RL}^{1'}$ means the repeat run of the first round of $\mathcal{M}_{RL}^{1}$. Best results are **highlighted**.

GTA5→Cityscapes																				
Model	road	sidewalk	building	wall	fence	pole	light	sign	veg.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
$\mathcal{M}_{RL}^{1}$	95.0	66.6	85.4	19.7	20.1	38.7	37.3	50.6	87.5	46.6	89.5	65.2	33.9	89.1	46.8	37.6	13.1	35.1	59.6	53.5
$\mathcal{M}_{RL}^{1'}$	95.5	67.6	85.6	30.5	19.6	36.9	35.0	49.6	87.6	46.4	**90.0**	64.8	29.5	88.3	37.7	39.1	**15.0**	37.7	58.8	53.4
$\mathcal{M}_{SL}^{1}$	93.8	59.2	85.4	33.9	29.0	37.5	**42.3**	45.3	86.8	44.4	86.0	63.8	37.6	87.7	45.9	49.5	0.1	39.0	56.2	53.9
$\mathcal{M}_{SL}^{1'}$	94.1	61.2	85.6	32.5	31.2	38.2	38.7	44.6	86.4	46.3	86.6	64.1	38.3	88.3	42.9	48.5	1.5	35.9	54.7	53.7
$E(\mathcal{M}_{RL}^{1}, \mathcal{M}_{RL}^{1'})$	**95.6**	**68.4**	86.0	24.6	20.8	38.8	37.6	**50.9**	87.8	47.7	89.7	65.4	33.1	89.2	44.7	41.6	10.9	35.3	60.2	54.1
$E(\mathcal{M}_{SL}^{1}, \mathcal{M}_{SL}^{1'})$	94.4	62.1	86.0	33.4	**31.7**	38.9	42.0	45.6	87.2	47.6	87.0	64.9	39.1	88.8	48.4	50.8	0.5	39.9	57.0	55.0
$E(\mathcal{M}_{RL}^{1}, \mathcal{M}_{SL}^{1})$	95.5	67.4	86.0	30.1	26.3	39.6	41.7	50.0	**88.0**	49.3	89.1	**66.7**	**40.0**	**90.0**	**53.2**	49.6	0.7	**43.3**	**61.3**	**56.2**
$E(\mathcal{M}_{RL}^{1}, \mathcal{M}_{SL}^{1'})$	95.5	68.0	86.1	30.8	28.5	**40.2**	39.2	49.4	87.9	50.0	89.1	**66.7**	**40.0**	89.8	50.0	48.0	2.8	42.6	59.9	56.0
$E(\mathcal{M}_{RL}^{1'}, \mathcal{M}_{SL}^{1})$	95.5	67.8	**86.4**	33.0	26.9	39.3	42.1	49.5	**88.0**	50.2	89.1	66.3	39.1	89.8	50.6	**51.2**	0.2	42.0	60.4	**56.2**
$E(\mathcal{M}_{RL}^{1'}, \mathcal{M}_{SL}^{1'})$	95.5	68.2	**86.4**	**33.7**	29.0	39.9	39.6	48.9	**88.0**	**50.5**	89.1	66.4	39.7	89.7	46.7	49.3	1.2	41.4	59.4	55.9

Table 4. The detailed results of domain-mixed teachers and student model during different rounds in the whole training process on GTA5→Cityscapes. For 2975 labeled images, our framework is justly trained one round.

Number	100				200				500				1000				2975			
Rounds R	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$\mathcal{M}_{RL}^{r}$	53.5	59.6	59.9	60.4	56.6	59.2	59.8	59.1	61.7	62.7	63.1	63.6	65.4	65.3	65.4	64.7	68.2	-	-	-
$\mathcal{M}_{SL}^{r}$	53.9	57.8	59.7	59.3	54.4	57.3	58.9	58.5	58.4	61.0	61.1	61.4	61.7	63.1	63.7	63.3	65.8	-	-	-
$\mathcal{M}_{S}^{r}$	57.1	59.8	61.0	61.2	58.3	60.2	60.3	60.5	62.5	63.7	64.1	64.3	65.5	66.0	66.6	66.0	69.8	-	-	-

fully 2975 images with the performance of 69.8%.

In Fig. 2, we further display some qualitative segmentation results of both our method and baseline method on 100, 200, 500 and 1000 labeled target images. Overall, our method achieves more complete segmentation results than baseline model in the same ratio of labeled images. As the number of labeled images increases, more refined segmentation results we can obtain by our proposed approach.

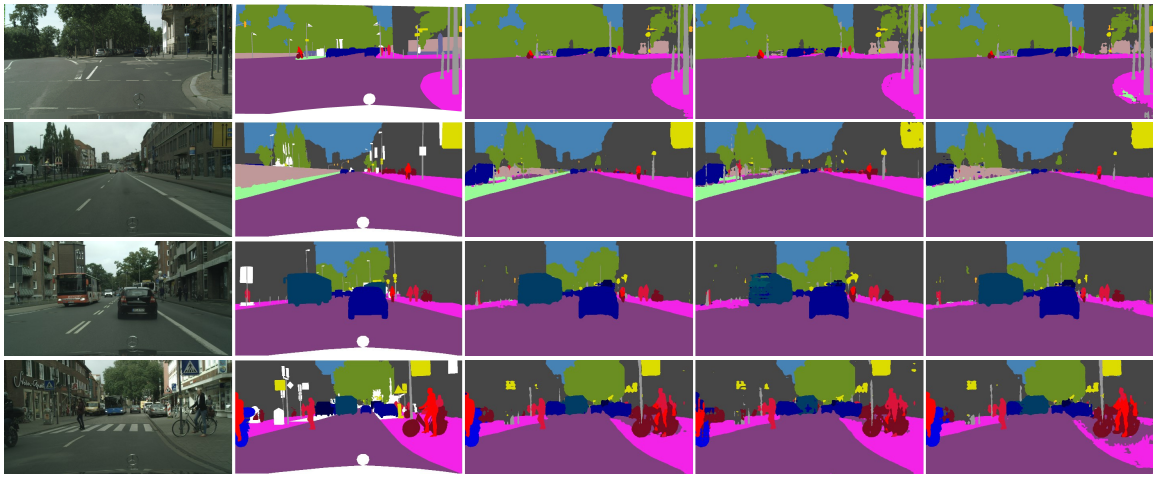
SYNTHIA to Cityscapes. In order to further measure the performance of our approach, we also compare the results with several state-of-the-art methods on the SYNTHIA to Cityscapes. Since there are only 16 common categories between the SYNTHIA and Cityscapes, we just train a segmentation model with the common categories. As shown in Table 2, following previous UDA works [42, 41], we also report 13-class mIoU score to compare with existing other methods. From the results, it is clear that our method outperforms the UDA, SSL and SSDA methods with a large performance gain. And the similar discussions we can draw as in "GTA5 to Cityscapes".

5. Ablation Study

5.1. Complementarity

To examine the complementarity of models trained from different views, we select 100 labeled target images and train the domain-mixed teachers twice separately from sample-level and region-level mixed data. Then ensemble of different models including two region-level teachers and two sample-level teachers, are conducted and results are shown in Table 3. Overall, from Table 3, we can draw a conclusion that model ensemble is effective for improving the performance, and the ensembled models from dual-level data mixing views can achieve better results than that from single-level data mixing view. In detail, region-level teachers perform better in categories which can be predicted without strictly relying on the structural information, such as road, sidewalk, vegetation and sky. However, they have poor predictions on the fence, light and bus classes whose shape is distinctive. We explain that the region-level data mixing operation could destroy the structure of these classes. Although model ensemble from one single view can realize impressive results on its own advantageous categories, the categories with poor performance are still poor. For example, the ensembled model from two region-level models achieves the best IoU score on road and sidewalk classes, and the worst results on rider and fence classes. Such best-worst phenomenon also occurs in ensembled model of sample-level teachers, but on different categories compared with ones in region-level. So we can fuse the models with different complementary levels and achieve a good result in all categories.

We also visualize some segmentation results of ensemble of different-level models in Fig. 3. From Fig. 3, the pixels what is wrongly classified in one view will be corrected in another view.



(a) Target Image (b) GT (c) Ensemble (d) Region-level (e) Sample-level

Figure 3. Qualitative results of ensembled models from region-level and sample-level views on GTA5 to Cityscapes. (a) target images, (b) ground truth, (c) segmentation results of model ensemble of different views, (d) results of model trained on region-level mixed data, (e) results of model trained on sample-level mixed data.

Table 5. Results of two domain-mixed teachers and student in the first round of our framework whether using style transfer or not across different number of labeled target images.

Model	Use-trans	100	200	500	1000	2975
$\mathcal{M}^1_{RL}$	T	53.5	56.6	61.7	65.4	68.2
	F	52.7	55.6	62.1	65.2	67.9
$\mathcal{M}^1_{SL}$	T	53.9	54.4	58.4	61.7	65.8
	F	51.5	53.5	56.4	59.8	64.0
$\mathcal{M}^1_{S}$	T	57.1	58.3	62.5	65.5	69.8
	F	55.8	57.0	62.1	65.0	68.9

5.2. Number of Iterative Rounds

We discuss our results reported in Table 1 during different rounds in the whole training process on GTA5 to Cityscapes, and the detailed results of three models, domain-mixed teachers and student, are shown in Table 4. All three models can be improved with obvious performance gain compared with first round in training process. During different rounds, the student model will outperform both of two teachers, and the stronger student will correct the learning of teachers through generating more accurate pseudo labels, thus the teachers and student are progressive growing. This demonstrates the effectiveness of our proposed iteratively framework. We notice that the best models are achieved in different rounds on different numbers of labeled images.

5.3. Image Translation

In the above experiments, a simple image translation method in LAB color space is firstly taken to further reduce the visual difference between different domain images. Additionally, the experiments without style transfer are also conducted to demonstrate the effectiveness of our approach. We just compare the results of two-domain mixed teachers and student model in the first round of our framework. From Table 5, we can draw the following three observations. First, the student model using style transfer achieves better performance than ones without it. Therefore, dual-level data mixing with style transfer can further reduce distribution mismatch across domains. Secondly, the teacher model trained on region-level mixed data becomes insensitive as the number of images increases. Superior performance without style transfer is obtained on 500 labeled target images than using style transfer. We argue that the reason is that region-level data mixing is relatively robust to whether style transfer is conducted in the one patch cropped from source image. Because of significant improvement in sample-level data mixing, we can also obtain better results with style transfer. Finally, our proposed framework can obtain better results than ASS even if without style transfer on 100, 200, 500, 1000 labeled images.

6. Conclusion

In this paper, we propose a novel framework based on dual-level domain mixing to address semi-supervised domain adaptation problem. Two complementary domain-mixed teachers can be obtained based on proposed two kinds of data mixing methods in both region-level and sample-level. A stronger student model on target domain is obtained by distilling knowledge from these two domain-mixed teachers. Then pseudo labels generated the strong student are used for next round training of domain-mixed teachers, which is conducted in a self-training fashion. Extensive experiments demonstrate the proposed framework can fully take advantage of the available data from source and target domains, and achieve superior performance on two commonly used synthetic-to-real benchmarks.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1, 5

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[3] Minmin Chen, Kilian Q Weinberger, and John Blitzer. Co-training for domain adaptation. In *Advances in neural information processing systems*, pages 2456–2464, 2011. 3

[4] Shuai Chen, Gerda Bortsova, Antonio García-Uceda Juárez, Gijs van Tulder, and Marleen de Bruijne. Multi-task attention-based semi-supervised learning for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–465. Springer, 2019. 2

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

[6] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1992–2001, 2017. 2

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

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

[9] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum. *arXiv preprint arXiv:2004.08514*, 2020. 1, 2, 5, 6

[10] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 1, 2, 5, 6

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1

[12] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2477–2486, 2019. 1

[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005. 2

[14] Jianzhong He, Xu Jia, Shuaijun Chen, and Jianzhuang Liu. Multi-source domain adaptation with collaborative learning for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 2021. 2, 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 1, 2

[17] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 1, 2

[18] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. 2

[19] Takashi Isobe, Xu Jia, Shuaijun Chen, Jianzhong He, Yongjie Shi, Jianzhuang Liu, Shengjin Wang, and Huchuan Lu. Multi-target domain adaptation with collaborative consistency learning. In *Proceedings of the IEEE international conference on computer vision*, 2021. 2

[20] Hoel Kervadec, Jose Dolz, Éric Granger, and Ismail Ben Ayed. Curriculum semi-supervised segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 568–576. Springer, 2019. 2

[21] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12975–12984, 2020. 2, 5, 6

[22] Taekyung Kim and Changick Kim. Attract, perturb, and explore: Learning a feature alignment network for semi-supervised domain adaptation. *arXiv preprint arXiv:2007.09375*, 2020. 3

[23] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 2, 4, 5

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[25] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE international conference on computer vision*, pages 1377–1385, 2015. 1

[26] Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. Predicting deeper into the future

of semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 648–657, 2017. 1

[27] Fengmao Lv, Tao Liang, Xiang Chen, and Guosheng Lin. Cross-domain semantic segmentation via domain-invariant interactive relation transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4334–4343, 2020. 5, 6

[28] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 2229–2235. IEEE, 2018. 1

[29] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 2

[30] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 1, 2

[31] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020. 1

[32] Can Qin, Lichen Wang, Qianqian Ma, Yu Yin, Huan Wang, and Yun Fu. Opposite structure learning for semi-supervised domain adaptation. *arXiv preprint arXiv:2002.02545*, 2020. 3

[33] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 1, 5

[34] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 1, 5

[35] Danila Rukhovich and Danil Galeev. Mixmatch domain adaptaion: Prize-winning solution for both tracks of visda 2019 challenge. *arXiv preprint arXiv:1910.03903*, 2019. 3

[36] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8050–8058, 2019. 5, 6

[37] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017. 2

[38] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018. 2

[39] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 624–628. IEEE, 2018. 1

[40] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017. 2

[41] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 1, 2, 3, 5, 6, 7

[42] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 1, 2, 5, 6, 7

[43] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7364–7373, 2019. 1

[44] Zhonghao Wang, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S Huang, and Honghui Shi. Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 936–937, 2020. 2, 3, 5, 6

[45] Luyu Yang, Yan Wang, Mingfei Gao, Abhinav Shrivastava, Kilian Q Weinberger, Wei-Lun Chao, and Ser-Nam Lim. Mico: Mixup co-training for semi-supervised domain adaptation. *arXiv preprint arXiv:2007.12684*, 2020. 3

[46] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019. 2, 3

[47] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3

[48] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1

[49] Xinge Zhu, Hui Zhou, Ceyuan Yang, Jianping Shi, and Dahua Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 568–583, 2018. 2

[50] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2