Multi-Source Domain Adaptation with Collaborative Learning for Semantic Segmentation

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

Abstract

Multi-source unsupervised domain adaptation (MSDA) aims at adapting models trained on multiple labeled source domains to an unlabeled target domain. In this paper, we propose a novel multi-source domain adaptation framework based on collaborative learning for semantic segmentation. Firstly, a simple image translation method is introduced to align the pixel value distribution to reduce the gap between source domains and target domain to some extent. Then, to fully exploit the essential semantic information across source domains, we propose a collaborative learning method for domain adaptation without seeing any data from target domain. In addition, similar to the setting of unsupervised domain adaptation, unlabeled target domain data is leveraged to further improve the performance of domain adaptation. This is achieved by additionally constraining the outputs of multiple adaptation models with pseudo labels online generated by an ensembled model. Extensive experiments and ablation studies are conducted on the widely-used domain adaptation benchmark datasets in semantic segmentation. Our proposed method achieves 59.0% mIoU on the validation set of Cityscapes by training on the labeled Synscapes and GTA5 datasets and unlabeled training set of Cityscapes. It significantly outperforms all previous state-of-the-arts single-source and multi-source unsupervised domain adaptation methods.

1. Introduction

Semantic segmentation as one of the core tasks in computer vision community, aims to assign semantic label to each pixel of images, *e.g.*, person, car, road and *etc.*. With the development of convolutional neural networks (CNNs), semantic segmentation has made great progress recently. For example, recent deep methods [2, 5, 7, 11, 41], have achieved superior performance on almost all public benchmarks. However, their success is based on the large numbers of densely annotated images which used to train the networks. Dense pixel-level annotation for semantic segmentation is very laborious and expensive, *e.g.*, annotating one image in the Cityscapes dataset [4] takes about 90 minutes, which makes it difficult and sometimes even impossible to collect large amounts of densely annotated images for semantic segmentation. Thanks to the recent progress in graphics and simulation infrastructure, simulators can generate lots of images with dense annotation for semantic segmentation, such as recent proposed large-scale dense labeled datasets SYNTHIA [28], GTA5 [27] and Synscapes [36]. Although the huge amounts of annotated synthesized images are very close to the real scene, there is still great domain gap between synthetic datasets and real scene datasets. The domain gap causes another problem that networks trained on synthetic datasets often perform poorly on real target scenes. To handle this issue, many un-/semi-supervised domain adaptation (UDA) approaches are proposed, like [13, 21, 29, 31, 32, 35, 37] and *etc.*, with the purpose of mitigating the gap between synthetic source and

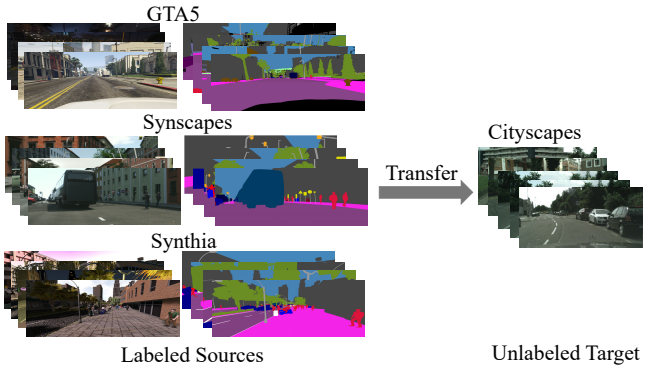

Figure 1. Multi-source domain adaptation for semantic segmentation. The left shows synthetic images and corresponding labels generated from different simulators, which suffer domain shift between each other but share similar semantic contexts. The right part shows unlabeled target images sampled from real scenes.

*Corresponding author

real target domain. Over the past years, UDA has made a great progress.

Although existing works have greatly boosted the performance of UDA for semantic segmentation, most of them focus on single source. Seldom works consider a more practical setting where labeled datasets from multiple sources with different distributions are available, *e.g.*, SYNTHIA and GTA5. Training with multiple sources can further alleviate the problem on lack of annotated data. Moreover, multiple sources sampled from different distribution can also encourage networks to learn more essential knowledge for semantic segmentation. A straightforward approach is to simply combine all source domains into a single one, and then trains a UDA model on the combined sources and target domain dataset. This simple method can indeed boost the performance, but it does not fully exploit the abundant information across multiple source domains. Domain shift across multiple sources restricts the power within them in learning a more powerful domain adaptation model.

There are several multi-source deep UDA methods are proposed recently to exploit multiple source domains for better adaptation. They align different domains by translating images from source domains to the target style via generative adversarial networks (GAN). However, most of them [23, 30, 15] work on image classification task except for MADAN [43] which works on semantic segmentation, a pixel-wise prediction task. In this paper, we propose an approach based on collaborative learning and image translation to address multi-source domain adaptation for semantic segmentation.

Our observation shows that appearance discrepancy especially color discrepancy between source domains and target domain has a great impact on the performance of adaptation. Existing works [34, 12, 38, 10] demonstrate that style transfer could reduce this discrepancy in some extent. However, most of them are complicated to plug in networks during training process. Therefore, we propose a simple image translation method to first mitigate domain gap between sources and target. Unlike MADAN [43], FDA [37] and GAN-based translation methods, we propose to translate source domain images to the target style by aligning different distributions to the target domain in LAB color space. In addition, we observe that apart from discrepancy in appearance, images from different domains do still share much similarity in semantic contexts as shown in Fig 1. The shape of instances (person, car, bike and *etc.*) and spatial layout of different instances (cars always on the road, sidewalk adjacent to the road, sky on the top and *etc.*) are almost the same in all domains. Two collaborative learning strategies are proposed to explore essential and domain-invariant semantic contexts across different domains. First we propose a collaborative learning between source domains to investigate the case of domain adaptation without seeing any data from target domain, which is also called domain generalization in previous works [6, 22, 40]. For each source domain, we have a semantic segmentation network supervised by annotation maps, and an additional soft supervision coming from other models trained on a different source domain. In addition, similar to previous UDA methods [43, 37], we also consider making full use of the unlabeled data of the target domain to further boost the performance. A collaborative learning based on target domain is proposed, in which an ensemble of models trained on source domains is used to produce pseudo labels for data from target domain in an online fashion. In turn, each model can be additionally supervised by the generated pseudo labels. Such two collaborations help constantly improve each model's adaptation capability to target domain during the training process.

The performance of our method significantly outperforms other state-of-the-art single-source and multi-source UDA methods. This success of proposed method is mainly attributed to the effective image translation and domain-invariant feature learning. Note that, our method can be trained in both end-to-end and stage-wise.

2. Related works

In this section, we briefly review some related works in the literature, *i.e.*, semantic segmentation, domain generalization and unsupervised domain adaptation.

2.1. Semantic Segmentation

Semantic segmentation plays a vital role in computer vision community and is beneficial to many practical applications, such as autonomous driving, virtual reality and medical imaging and *etc.*. It has developed several decade years and is well researched. Since Long *et al.* [16] propose to transform the classification CNNs to fully convolutional network for semantic segmentation, large numbers of deep learning based methods have been proposed and greatly boost the advances of this task. For example, Chen *et al.*propose the DeepLab series [1, 2] approaches which utilize átrous spatial pyramid pooling (ASPP) to capture different scale of context information. Fu *et al.* [5] involve non-local attention block into the CNNs architecture to exploit the global context of image and relation of objects. Hou *et al.* [9] propose a new effective strip pooling to model long-range dependencies. However, the advanced performance of these semantic segmentation methods often build on the large amounts of densely annotated images which are usually hard and sometimes impossible to collect.

2.2. Domain Generalization

Domain generalization is a particular case of transfer learning. It's purpose is to enhance the generalization ability of models on new domains that have not been

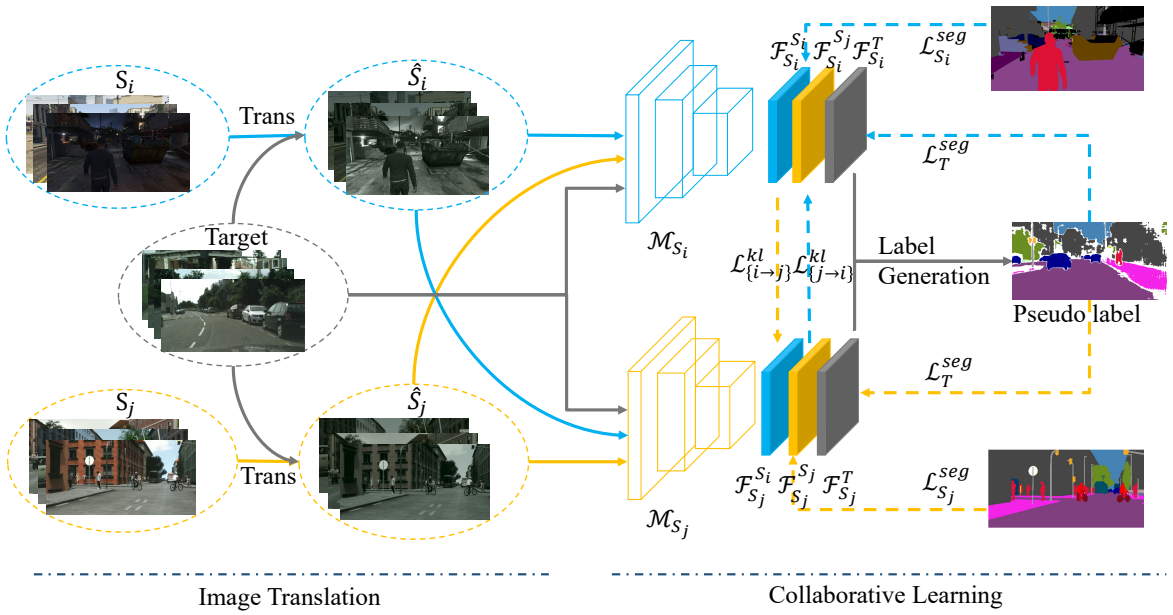

Figure 2. The overall framework of proposed approach consists of three components, including that image-to-image translation based on LAB color space, collaborative learning between source domains and collaborative learning on target domain. The solid arrows represent the forward data flow and different colors indicate different source domains or target domain data flow. The dash arrows represent the supervision to the network outputs. For illustration, we just show the case of two source domains as an example to explain our method.

seen during the training process. Currently, most of the domain generalization methods can be categorized to three parts: data-based, feature-based and meta-learning based. Most data-based methods employ GANs, Variational Autoencoder (VAE) or other image edit methods to generate new data for network training to enhance robustness and generalization of the models, *e.g.*, Dlow [6] and STRG [39]. Then, the feature-based methods mainly aim to learn representations invariant to different domains by adversarial learning to align features or employing normalization to eliminate the style information, like CADAG [25] and IBN-Net [22]. Meta-learning based methods aim to enhance the generalization ability by using meta learning, such as Zhang *et al.* [40] deal with the domain generalization from the training scheme perspective and develop a target-specific normalization method to further boost the generalization ability in unseen target domain. Domain generalization do not access the data in target domain which may not learn the optimal feature for target domain.

2.3. Unsupervised Domain Adaptation

In contrast to domain generalization, unsupervised domain adaptation involves in unlabeled data of target domain during training process to learn knowledge of it. UDA has developed many years and large numbers of methods have been proposed. From the perspective of the number of source domains, domain adaptation methods can be split into two categories: single-source and multi-source.

Single-source domain adaptation focuses on single-source-single-target setting. Most of existing UDA works are for classification, like MMD [17, 18], ADDA [33] and *et al.*. With the synthetic dataset GTA5, SYNTHIA and Synscapes proposed , UDA for semantic segmentation has also achieved great advances in recent. For example, Tsai *et al.* [32] propose AdaptSeg that based on adversarial learning to aligns scene layout and local context of images between source and target domain. Yang *et al.* [37] draw on the image-to-image translation and propose to using Fourier Transform to translate images in source domain to the style of target domain. While Li *et al.* [13] propose a bidirectional framework for domain adaptation and learn a image translation model with perceptual loss to translate images to target style. These methods have significantly boost the performance of UDA in semantic segmentation. However, they do not take the existing multiple different source domains into account which is a great waste of labeled resources.

Multi-source domain adaptation aims to make full use of existing labeled source for adaptation. Compared to single-source UDA, it is relatively more challenging because of the shift between source domains. There are also some multi-source UDA works but most of them focus on image classification. Directly extending these methods for semantic segmentation may not work. Zhao *et al.* [43] propose MADAN for semantic segmentation. They first learns a cycle GAN [44] for each source domain to translate images to target, and then combine different adapted sources with specific weights. Finally, training a adaptation network on

Source Image Trans. on RGB Trans. on LAB Target Image

Figure 3. The qualitative comparison of image translation on different color space.

the combined data in a way similar to single-source UDA. In contrast to [43], we not only draw on the style-transfer for image-to-image translation, also aim to learn the shared semantic information of source domains and explore the knowledge of unlabeled target data.

3. Proposed Approaches

3.1. LAB-based Image Translation

We observe that the domain discrepancy between domains mainly lies in appearance of images, *i.e.*, color and texture. For example, the appearance discrepancy between source domains (GTA5 and SYNTHIA), and between source domains and target domain (GTA5 and Cityscapes) as demonstrated in Fig. 1. This discrepancy would further increase the difficulty of multi-source domain adaptation. An effective style transfer could reduce this discrepancy to some extent. For simplicity and efficiency, we proposed an image translation method that translates the style of images in source domains to the style of target domain by aligning the distribution of pixel values, like [26]. Specifically, due to the gamut of LAB color space is larger than RGB color space, and the style of images translated on LAB color space are closer to the style of images in target domain than directly operated on RGB color space as shown in Fig. 3, we involve the image translation based on LAB color space to achieve the target of reducing domain discrepancy.

Specifically, for a RGB image $\mathcal{X}_S^{RGB}$ in source domains, it is firstly converted to LAB color space to generate LAB image $\mathcal{X}_S^{LAB} = rgb2lab(\mathcal{X}_S^{RGB})$. Then, we calculate the mean μ_S and standard deviation values σ_S of each channel of the generated LAB image $\mathcal{X}_S^{LAB}$. At the same time, an image from target domain is randomly selected and converted to LAB color space in same way as image from source. After that, we also calculate the mean μ_T and standard deviation values σ_T of this converted target image. Finally, we translate the converted LAB image from source domains to the style of target by shifting the distribution of pixel values to the image of target domain, *ie.*,

$$\hat{\mathcal{X}}_S^{LAB} = \frac{(\mathcal{X}_S^{LAB} - \mu_S)}{\sigma_S} * \sigma_T + \mu_T. \tag{1}$$

After aligning distribution in LAB color space, we then convert the translated LAB image $\hat{\mathcal{X}}_S^{LAB}$ from source domains back to the RGB color space $\hat{\mathcal{X}}_S^{RGB} = lab2rgb(\hat{\mathcal{X}}_S^{LAB})$, for the subsequent training. The $rgb2lab$ and $lab2rgb$ are the library conversion function in python, and also easy to implement it with CUDA to accelerate computation. Fig 3 shows the qualitative comparison between RGB- and LAB-based translation. We can see that the images are close to the style of target domain after translation, and the gap between domains is reduced to some extent. Moreover, the LAB-based translated images are closer to the style of target images.

3.2. Collaborative Learning between Labeled Source Domains

As shown in Fig. 1, although the images of different source domains are sampled from different simulators, *i.e.*, from different *i.i.d* distributions, they are highly structured and share many similarities, *e.g.*, the spatial layout of different categories and local context of objects. Based on this observation and to encourage models learn these essential similar properties for semantic segmentation, we propose a collaborative learning (Co-Learning) method which fully takes advantage of the capacity of different models learned on different source domains. This approach utilizes a model learn from one source domain to teach the model trained on another one, which allows that model to learn the knowledge from current source domain.

Assume that there are N different labeled source domains $S = \{S_1, S_2, \cdots, S_N\}$ which are sampled from N different *i.i.d* distributions, and N deep neural networks $\mathcal{M} = \{\mathcal{M}_{S_1}, \mathcal{M}_{S_2}, \cdots, \mathcal{M}_{S_N}\}$ of the same architecture but different weights learned on these source domains. Then, for an model $\mathcal{M}_{S_i}$, the learning process of model $\mathcal{M}_{S_i}$ is supervised by segmentation loss on labeled data from source S_i and collaborative loss on output from source $S_{k,k\neq i}$. That is, for model $\mathcal{M}_{S_i}$, the object function is

$$\mathcal{L}_i = \mathcal{L}_{S_i}^{seg}(\mathcal{F}_{S_i}^{S_i}, \mathcal{Y}_{S_i}) + \lambda_S^{col}\mathcal{L}_S^{col}(\{(\mathcal{F}_{S_i}^{S_k}, \mathcal{F}_{S_k}^{S_k})_{k\neq i}\}), \tag{2}$$

where the loss $\mathcal{L}^{seg}$ is the cross entropy loss, *i.e.*,

$$\mathcal{L}_S^{seg}(\mathcal{F}_S, \mathcal{Y}_S) = -\frac{1}{|\mathcal{X}_S|}\sum_{h,w}\sum_{c\in C}\mathcal{Y}_S^{(h,w,c)} log(\sigma(\mathcal{F}_S^{(h,w,c)})), \tag{3}$$

and the loss $\mathcal{L}^{col}$ is the average of Kullback-Leibler (KL) divergence loss, *i.e.*,

$$\mathcal{L}_S^{col}(\{(\mathcal{F}_{S_i}^{S_k}, \mathcal{F}_{S_k}^{S_k})_{k\neq i}\}) = \frac{1}{N-1}\sum_{k,k\neq i}\mathcal{L}_{k\to i}^{kl}(\mathcal{F}_{S_k}^{S_k} \| \mathcal{F}_{S_i}^{S_k}), \tag{4}$$

$$\mathcal{L}_{k\to i}^{kl}(\mathcal{F}_{S_k}^{S_k} \| \mathcal{F}_{S_i}^{S_k}) = -\frac{1}{|\mathcal{X}_{S_k}|}\sum \sigma(\mathcal{F}_{S_k}^{S_k}) \log(\frac{\sigma(\mathcal{F}_{S_i}^{S_k})}{\sigma(\mathcal{F}_{S_k}^{S_k})}). \tag{5}$$

$\mathcal{X}, \mathcal{F}, \mathcal{Y}$ are the input image, outputs of networks and the corresponding ground truth respectively. For $\mathcal{F}_{S_i}^{S_k}$, the

subscript indicates that $\mathcal{F}$ is generated by $\mathcal{M}_{S_i}$ while the superscript indicates $\mathcal{F}$ is the feature computed for images from domain S_k, *i.e.*, $\mathcal{F}_{S_i}^{S_k} = \mathcal{M}_{S_i}(\mathcal{X}_{S_k})$. C is the number of categories to be segmented, $\sigma(\cdot)$ indicates the softmax function, $|\mathcal{X}|$ represents the number of pixels in image $\mathcal{X}$. Collaborative learning can allow the model $\mathcal{M}_{S_i}$ learn the knowledge of other models $\mathcal{M}_{S_{k,k\neq i}}$ learned from corresponding source S_k. Thus, rather than stuck in the source domain S_i, model $\mathcal{M}_{S_i}$ also tries to learn essential properties between all the source domains for better generalization and segmentation on target domain.

Difference with existing works: In the co-training of CLAN [19] and CT [24], two diversified classifiers are produced to make predictions and their ensemble with either summation or multiplication is taken as the final prediction; however, in the collaborative learning framework, two segmentation models trained based on each source domain **teach each other** to extract essential semantic information across domains. CrDoCo [3] uses bi-directional KL divergence to encourage segmentation consistency between images in unlabeled target domain and their source-style transformed correspondences. However, in our work, images from both source domains are first transformed to the same style, i.e., target domain style. Then an arbitrary image is sent to two models that are trained for source 1 and source 2 respectively. The prediction of the model trained for the source domain that image comes from works as teacher to supervise the other model's prediction on the **same** image.

3.3. Collaborative Learning on Unlabeled Target Domain

In practice, unlabeled data is often relatively easy and cheap to collect. Moreover, models trained on the data of target domain can learn better features and perform well on target domain. Therefore, we propose a collaborative learning method to fully take advantage of the unlabeled images in target domain, and further boost the performance of models on target domain.

Denote target domain as T which has collected large amounts of unlabeled images $\{\mathcal{X}_T^j\}_{j=1}^{N_T}$. For an image $\mathcal{X}_T$ in the target domain T, we first feed it to all the N models $\mathcal{M} = \{\mathcal{M}_{S_i}, i = 1, 2, \cdots, N\}$ and compute the corresponding outputs of models $\mathcal{M}_{S_i}$ for the input image $\mathcal{X}_T$ as $\mathcal{F}_{S_i}^T$, *i.e.*, $\mathcal{F}_{S_i}^T = \mathcal{M}_{S_i}(\mathcal{X}_T)$. Outputs of these networks are ensembled and a softmax function $\sigma(\cdot)$ is used to compute the probability map, *i.e.*,

$$\hat{P} = \sigma(\frac{1}{N}\sum_i \mathcal{F}_{S_i}^T). \tag{6}$$

Finally, we generate one-hot pseudo labels $\hat{\mathcal{Y}}$ for the target image $\mathcal{X}_T$ by utilizing the probability map in a way similar to [13, 45], and the detail process is described in the Algorithm 1. After then, we can train models $\mathcal{M}_{S_i}$ with both source domains and target domain. Thus, the object function for model $\mathcal{M}_{S_i}$ becomes

$$\mathcal{L}_i = \mathcal{L}_{S_i}^{seg}(\mathcal{F}_{S_i}^{S_i}, \mathcal{Y}_{S_i}) + \frac{cur\_it}{max\_its}\lambda_T^{seg}\mathcal{L}_T^{seg}(\mathcal{F}_{S_i}^T, \hat{\mathcal{Y}}_T), \tag{7}$$

where cross entropy loss is employed to compute the loss $\mathcal{L}_T^{seg}$ and is same as Equ. 3, the only difference is that the target labels are generated pseudo labels. Here the the weight term $\frac{cur\_it}{max\_its}$ is designed to prevent poor predictions of the target image in early training process collapsing the network training, cur_it and max_its represent current and maximum iterations of training process.

Algorithm 1: Pseudo Labels Generation

Data: The probability map $\hat{P} \in \mathcal{R}^{C\times H\times W}$, keep proportion α, maximum thresh τ, the ignore label l_{ig}

Result: one-hot hard pseudo labels $\hat{\mathcal{Y}}$

$\hat{\mathcal{Y}} \leftarrow argmax(\hat{P}, dim = 0)$, $\hat{\mathcal{Y}} \in \mathcal{R}^{H\times W}$

for $c \leftarrow 0$ *to* $C-1$ **do**

  $\hat{P}_c \leftarrow sort(\hat{P}_{\{c,\cdot,\cdot\}}, order = Descending)$;

  get the number of pixels n_c which are predicted to category c: $n_c \leftarrow sum(\hat{\mathcal{Y}} == c)$;

  get the threshold t that used to filter the prediction: $t \leftarrow min(\hat{P}_c[n_c \times \alpha], \tau)$;

  $mask1 \leftarrow \hat{\mathcal{Y}} == c$;

  $mask2 \leftarrow \hat{P}_{\{c,\cdot,\cdot\}} <= t$;

  $\hat{\mathcal{Y}}[mask1 \,\&\, mask2] \leftarrow l_{ig}$.

end

3.4. Network Architecture and Training

The overall architecture of proposed approach are shown in Fig. 2. In summary, it consists of three parts, *i.e.*, image-to-image translation based on LAB color space, collaborative learning between labeled source domains and on unlabeled target domain. Firstly, images of source domains are translated to the style of target domain as described in Sec. 3.1 during training process. Note that for simplicity, we just showcase the collaborative learning approach for multi-source domain adaptation with two source domains. As shown in Fig. 2, we train a segmentation network, *e.g.*, DeepLab v2 [1], for each source domain. These networks can be trained end-to-end which initialized from ImageNet pretrained model, or trained with stage-wise which initialized from source fine-tuned model. For each network, the final overall object function is as follows,

$$\mathcal{L} = \mathcal{L}_S^{seg} + \lambda_S^{col}\mathcal{L}_S^{col} + \frac{cur\_it}{max\_its}\lambda_T^{seg}\mathcal{L}_T^{seg}. \tag{8}$$

The loss $\mathcal{L}_S^{seg}$ ensures the correct functionality of networks for segmentation, while the loss $\mathcal{L}_S^{col}$ allows the network

Table 1. Validity of the proposed image translation method. The performance comparison with the recent single-source UDA methods trained on images that before and after translation.

GTA5→Cityscapes			
Methods	Before	+Trans	Diff.
Direct Transfer	39.53	43.36	↑ 3.83
AdaptSeg [32]	41.32	43.66	↑ 2.43
AdaptSeg-LS [32]	43.11	45.95	↑ 2.84
Advent [35]	44.30	45.96	↑ 1.66

to learn the similarities that shared by source domains. In addition, the loss $\mathcal{L}_T^{seg}$ allows the network learn the property of target data. When network training get converged, we have two strategies to obtain the final model. One is to choose the model which achieves the best performance on target domain. The other is to keep all models, and ensemble their outputs when inference. Unless otherwise specified, all refer to the last case. The following section tests the validity of proposed approach.

4. Experiments

4.1. Datasets

In this subsection, we briefly introduce the datasets used to validate our adaptation method, *i.e.*, the widely used dataset Cityscapes [4], and recent proposed synthetic datasets GTA5 [27], SYNTHIA [28], and Synscapes [36].

Cityscapes [4] consists of 5,000 real-world urban traffic scene images with 2048×1024 resolution and dense-pixel annotation. This dataset is split to 2,975 for training, 500 for validation and 1,525 for testing. Cityscapes annotates 33 categories of which 19 are used for training and evaluation. The train set without ground truth is used for training adaptation models and validation set for evaluation.

GTA5 [27] includes 24,966 dense annotated images synthesized from a game engine with the resolution of 1914×1052, and the annotation are consistent with Cityscapes [4]. In a way similar to previous works [13, 35], common categories between GTA5 and Cityscapes are used in all experiments.

SYNTHIA [28] is a large synthetic dataset that consists of photo-realistic frames rendered from a virtual city. In experiments, we use the SYNTHIA-RAND-CITYSCAPES [28] set for adaptation. It contains 9,400 images with a resolution of 1280×760 which are annotated into 16 categories. Similar to GTA5, its annotations are also automatically produced and compatible with Cityscapes. Following previous works [37, 32, 35], we evaluate performance on 16 and 13 common categories between Cityscapes and SYNTHIA when SYNTHIA is used.

Synscapes [36] is a synthetic dataset created by photo-realistic rendering techniques. It consists of 25,000 images at 1440×720 resolution with 33 categories dense annotation and only 19 of them are used. Similarly, its annotations are compatible with Cityscapes. The style of Syncscapes is closer to Cityscapes than GTA5 and SYNTHIA.

Table 2. The validity of model selection and the proposed collaborative learning on the GTA5 + Synscapes to Cityscapes. (a) shows the performance of each single model and the final ensemble, (b) shows the comparison of proposed collaborative learning between source domains (Co-Learning-Src) with baseline and MLDG [40]. **E:** End-to-End, **S:** Stage-Wise.

(a)

Model	E	S
$\mathcal{M}_{S_{GTA5}}$	56.90	57.72
$\mathcal{M}_{S_{Syns}}$	56.65	57.81
$\mathcal{M}_{Ensemble}$	58.55	59.04

(b)

Methods	mIoU	Diff.
Data Combination	51.56	–
MLDG+TN [40]	52.73	↑ 1.17
Co-Learning-Src	55.79	↑ 4.23

4.2. Implementation Details

We both implement our proposed approach with PyTorch* and Mindspore†, and conduct experiments by adopting DeepLab-v2 [1] with ResNet-101 [8] as backbone. The performance of initialization from both ImageNet pretrained model and source pretrained model are reported, which denote as end-to-end and stage-wise training strategies respectively. Following prior works [35, 32], all the networks are trained with stochastic gradient descent (SGD) optimizer. The initial learning rate and momentum are set to $2.5 \times e^{-4}$ and 0.9, respectively, and the polynomial decay policy with power of 0.9 is adopted to adjust the learning rate. We set the batch size to 1 for all datasets during training because of memory limitation. The hyper-parameters in collaborative learning on target domain, α and τ are set to 50% and 0.9, respectively. It indicates that we keep pixels that prediction probability within the top 50% or higher than 0.9 as true labels, and reminder are ignored. The weights λ_S^{col} and λ_T^{seg} for collaborative learning losses are set to 0.5, 0.1 for end-to-end training and 9.5, 0.1 for stage-wise training. Following Advent [35], we set the number of maximum iterations to 250,000 but early stop at 120,000 iterations. Same as previous works [37, 35, 32, 21], the metric mean intersection-over-union (mIoU) is used to evaluate the performance of our adaptation method.

4.3. Ablation Study

In this subsection, extensive experiments on different adaptation settings are conducted to study the effectiveness of each component in the proposed approach.

Firstly, LAB-based image translation is applied to the state-of-the-art UDA methods in semantic segmentation, *i.e.*, AdaptSeg [32] and Advent [35], to investigate the effectiveness of proposed image translation method. Table 1 is the performance comparison of previous UDA methods training on original and translated images. From the results,

*https://www.pytorch.org/
†https://www.mindspore.cn/

Table 3. Ablation studies of proposed methods. Note that, the performances are achieved by end-to-end training strategy for comparison with simple combination of sources.

GTA5 + Synscapes → Cityscapes				
LAB-based Trans.	Data Comb.	Co-Learning between Src.	Co-Learning on Target	mIoU
	✓			51.59
✓	✓			54.38
✓			✓	54.03
✓		✓		56.03
		✓	✓	57.27
✓		✓	✓	58.55

we can see that LAB-based image translation can greatly boost the performance of UDA methods for semantic segmentation. For example, LAB-based translation advances the performance of direct transfer from 39.38% to 43.36%, which improves about 3.83%. The performance of AdaptSeg and Advent are also significantly boosted, *i.e.*, 2.43%, 2.84% and 1.66% for AdaptSeg, AdaptSeg-LS and Advent respectively. These results demonstrate that the proposed simple image translation method based on LAB color space is effective for UDA in semantic segmentation. It is easy to implement, with only a little additional computation and without any extra hyper parameter. Note that, all results are reproduced with the code provided by the authors.

Because our method can be trained in different strategies and there are N models in the framework, we then test the performance of each single model and the final ensemble one with different training strategies. All these results are based on the setting that adapting from GTA5+Synscapes to Cityscapes. As shown in Table 2 (a), the performance of all models are significantly improved and the ensembled model achieves the best performance. Therefore, we only report the ensembled model's performance later. The collaborative learning between source domains does not access to the unlabeled target data and therefore can be used to address the task of domain generalization. Table 2 (b) shows the performance comparison with the recent multi-source domain generalization method MLDG [40] and a baseline based on simple data combination. We can see that collaborative learning achieves better generalization performance. For example, MLDG only marginally outperforms the simple baseline about 1.17%, while collaborative learning boosts the performance about 4.23%. Note that, collaborative learning here is only applied to source domains without image translation for fair comparison.

Table 3 shows different contribution of each component to performance of our proposed approach. As the results shown, training on combination of source domains can improve the performance on target domain to some extent, which achieves 51.59% and 54.38% respectively. Collaborative learning between different source domains

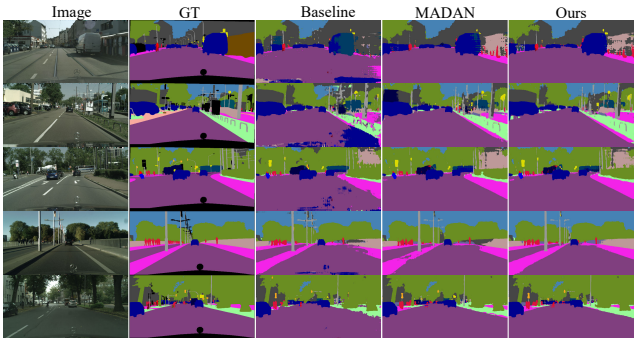

Figure 4. Visual Comparison with baseline and other methods. Left to right: input image from Cityscapes, corresponding ground-truth, segmentation results of baseline that simple combination of source domains, MADAN [43] and our proposed method. Note that, all these results are adapting from GTA5+Synscapes.

further boosts the performance to 56.03%. Collaborative learning on target domain also boosts the performance on target which achieves 54.03%. This result shows that collaborative learning between source domains brings more improvements than on target domain. Moreover, full version of our approach achieves the best performance, achieving 58.55% on target domain. Thus, we can conclude that the proposed approach is effective for unsupervised domain adaptation in semantic segmentation.

4.4. Comparison with SOTA

In this subsection, our proposed approach is compared to the recent state-of-the-art single-source and multi-source UDA methods on the GTA5+Synscapes to Cityscapes, including a baseline of source-only direct transfer (DT), single-source UDA methods [32, 21, 35, 14, 13, 37, 20], the multi-source baseline that simple combination of source domains and multi-source UDA methods [42, 43]. We further validate the effectiveness of our proposed method by conducting experiments based on adapting different number of source domains to target domain.

Table 4 shows the results comparison with other methods and Fig. 4 demonstrates the visual comparison with baseline and MADAN [43]. From the results of Table 4, we can see that simply combining the images of source domains for networks training can already greatly boost the generalization performance on target domain, which advances from the 44.1% of GTA5→Cityscapes and 45.3% of Synscapes→Cityscapes to 51.6%. Simply adapting UDA methods to train on combination of source domains does not bring much improvements, such as AdaptSeg has only 1% improvement and Advent only 0.5%. By employing adaptation strategy, multi-source UDA further boost the performance, *i.e.*, MDAN [42] achieves 55.2% and MADAN [43] achieves 55.7%. By further integrating collaborative learning on the source domains and target domain, we achieve 58.6% mIoU with the end-to-end

Table 4. The quantitative comparison with the state-of-the-art methods. DT is the abbreviation of direct transfer. G, S and A indicate GTA5, Synscapes and All respectively. Adv, CL, ST and RL indicate Adversarial learning, Curriculum Learning, Self Training and Reconstruction Learning respectively. Ours-E and Ours-S represent end-to-end training and stage-wise training of our proposed method respectively.

Methods	Appr.	Source	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
DT [32]	–		81.8	40.6	76.1	23.3	16.8	36.9	36.8	40.1	83.0	34.8	84.9	59.9	37.7	78.5	20.4	20.5	7.8	27.3	52.5	45.3
AdaptSeg [32]	Adv		94.2	60.9	85.1	29.1	25.2	38.6	43.9	40.8	85.2	29.7	88.2	64.4	40.6	85.8	31.5	43.0	28.3	30.5	56.7	52.7
FDA [37]	ST	S	93.6	58.1	84.0	30.4	29.2	39.0	43.1	**51.7**	85.9	28.8	86.9	64.0	**45.7**	84.7	30.4	36.5	28.5	34.4	**62.4**	53.5
Advent [35]	Adv		92.2	51.3	85.0	40.8	31.2	39.0	42.5	42.5	86.5	46.1	84.8	65.2	39.0	87.0	32.6	49.0	29.5	28.6	50.0	53.8
UIA [21]	Adv		94.0	60.0	84.9	29.5	26.2	38.5	41.6	43.7	85.3	31.7	88.2	66.3	44.7	85.7	30.7	53.0	29.5	36.5	60.2	54.2
DT [32]	–		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
AdaptSeg [32]	Adv		86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
Advent [35]	Adv		89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
UIA [21]	Adv	G	90.6	36.1	82.6	29.5	21.3	27.6	31.4	23.1	85.2	39.3	80.2	59.3	29.4	86.4	33.6	53.9	0.0	32.7	37.6	46.3
PyCDA [14]	CL		90.5	36.3	84.4	32.4	28.7	34.6	36.4	31.5	86.8	37.9	78.5	62.3	21.5	85.6	27.9	34.8	18.0	22.9	49.3	47.4
BDL [13]	ST		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
FDA [37]	ST		92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
PIT [20]	RL		87.5	43.4	78.8	31.2	30.2	36.3	39.9	42.0	79.2	37.1	79.3	65.4	37.5	83.2	46.0	45.6	25.7	23.5	49.9	50.6
Data Comb.	–		85.1	36.9	84.1	39.0	33.3	38.7	43.1	40.2	84.8	37.1	82.4	65.2	37.8	69.4	43.4	38.8	34.6	33.2	53.1	51.6
AdaptSeg [32]	Adv		89.3	47.3	83.6	40.3	27.8	39.0	44.2	42.5	86.7	45.5	84.5	63.1	38.0	79.4	34.9	48.3	42.1	30.7	52.3	53.7
Advent [35]	Adv		91.8	49.0	84.6	39.4	31.5	39.9	42.9	43.5	86.3	45.1	84.6	65.3	41.0	87.1	37.9	49.2	31.0	30.3	48.8	54.2
MDAN [42]	Adv	A	92.4	56.1	86.8	42.7	32.9	39.3	48.0	40.3	87.2	47.2	**90.5**	64.1	35.9	87.8	33.8	48.6	39.0	27.6	49.2	55.2
MADAN [43]	Adv		94.1	61.0	86.4	43.3	32.1	40.6	49.0	44.4	87.3	47.7	89.4	61.7	36.3	87.5	35.5	45.8	31.0	33.5	52.1	55.7
Ours-E	–		**94.2**	**61.8**	86.7	**47.7**	34.1	39.3	44.6	34.2	87.2	**49.6**	89.7	**65.6**	38.1	**88.2**	**48.1**	**63.0**	41.9	**39.2**	59.2	58.6
Ours-S	–		93.6	59.6	**87.1**	44.9	**36.7**	**42.1**	**49.9**	42.5	**87.7**	47.6	89.9	63.5	40.3	88.2	41.0	58.3	**53.1**	37.9	57.7	**59.0**

Table 5. The performance of our proposed method that uses different source domains for adaptation. **G: GTA5, S: Synscapes, Y: SYNTHIA.** mIoU19, mIoU16 and mIoU13 indicate performance on different number of categories.

	sources	mIoU19	mIoU16	mIoU13
Source-Only	G	39.53	43.28	48.25
	S	44.43	48.74	54.09
	Y	–	32.31	37.41
Multi-Sources	G+S	59.04	61.25	65.87
	G+Y	–	54.03	59.42
	S+Y	–	58.19	63.18
	G+S+Y	–	62.24	67.15

training strategy and 59.0% mIoU with stage-wise training strategy, which greatly outperforms all the previous methods. When compared to single-source UDA, we observed that our approach achieves more significant improvement on categories such as train, truck, bus and *et al.*. These objects have rigid body and share much similarity in shape among different source domains. These results also validate the effectiveness of our proposed method. It is noteworthy that our approach does not employ any adversarial learning or any other sophisticated tricks, such as curriculum learning or self-training. More experiments are shown in supplementary.

Table 5 shows the performance comparison that adapting different source domains to Cityscapes. As the results shown, our approach can make full use of the labeled source domains and significantly improve the performance on target domain. We can see that adapting from GTA5 and Synscapes achieves 59.04% mIoU on 19 categories and 61.25% mIoU on 16 categories, which brings about 15% and 13% improvements *w.r.t* to the best model training on single-source. When adapting from all the three labeled source domains, our method further improve the performance (mIoU on 16 categories) from 61.25% to 62.24%. These results further illustrates the effectiveness of our method.

5. Conclusion

In this paper, we present an effective multi-source domain adaptation framework for semantic segmentation based on collaborative learning. A simple image translation method is proposed to reduce the gap between domains. A collaborative learning method based on both labeled source domains and unlabeled target domain is proposed to fully explore essential semantic contexts across domains. Extensive experiments and ablation studies show that the proposed framework is able to significantly outperform all previous state-of-the-arts single-source and multi-source unsupervised domain adaptation methods, by effectively taking advantage of labeled data from multiple source domains and unlabeled data from target domain.

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. 2, 5, 6

[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, 2

[3] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1791–1800, 2019. 5

[4] 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. 1, 6

[5] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 1, 2

[6] 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. 2, 3

[7] Jianzhong He, Shiliang Zhang, Ming Yang, Yanhu Shan, and Tiejun Huang. Bdcn: Bi-directional cascade network for perceptual edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[8] 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. 6

[9] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2

[11] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[12] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468, 2018. 2

[13] 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. 1, 3, 5, 6, 7, 8

[14] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 7, 8

[15] Hongfu Liu, Ming Shao, and Yun Fu. Structure-preserved multi-source domain adaptation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1059–1064. IEEE, 2016. 2

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[17] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 3

[18] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in neural information processing systems*, pages 136–144, 2016. 3

[19] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 5

[20] 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. 7, 8

[21] 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, 6, 7, 8

[22] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 2, 3

[23] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[24] Can Qin, Lichen Wang, Yulun Zhang, and Yun Fu. Generatively inferential co-training for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 5

[25] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Correlation-aware

adversarial domain adaptation and generalization. *Pattern Recognition*, 100:107124, 2020. 3

[26] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 4

[27] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 1, 6

[28] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 6

[29] Chen Shuaijun, Jia Xu, He Jianzhong, and Liu Jianzhuang. Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 2021. 1

[30] Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A two-stage weighting framework for multi-source domain adaptation. In *Advances in neural information processing systems*, pages 505–513, 2011. 2

[31] Isobe Takashi, Jia Xu, Chen Shuaijun, He Jianzhong, Shi Yongjie, Liu Jianzhuang, Wang Shengjin, and Lu Huchuan. Multi-target domain adaptation with collaborative consistency learning. In *Proceedings of the IEEE international conference on computer vision*, 2021. 1

[32] 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, 3, 6, 7, 8

[33] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 3

[34] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017. 2

[35] 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, 6, 7, 8

[36] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018. 1, 6

[37] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3, 6, 7, 8

[38] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019. 2

[39] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2100–2110, 2019. 3

[40] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Generalizable semantic segmentation via model-agnostic learning and target-specific normalization. *arXiv preprint arXiv:2003.12296*, 2020. 2, 3, 6, 7

[41] 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

[42] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In *Advances in neural information processing systems*, pages 8559–8570, 2018. 7, 8

[43] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 7287–7300, 2019. 2, 3, 4, 7, 8

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3

[45] Yang Zou, Zhiding Yu, BVK 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. 5