

Self-Ensembling with GAN-based Data Augmentation for Domain Adaptation in Semantic Segmentation

Jaehoon Choi
KAIST

Taekyung Kim
KAIST

Changick Kim
KAIST

{whdns44, tkkim93, changick}@kaist.ac.kr

Abstract

Deep learning-based semantic segmentation methods have an intrinsic limitation that training a model requires a large amount of data with pixel-level annotations. To address this challenging issue, many researchers give attention to unsupervised domain adaptation for semantic segmentation. Unsupervised domain adaptation seeks to adapt the model trained on the source domain to the target domain. In this paper, we introduce a self-ensembling technique, one of the successful methods for domain adaptation in classification. However, applying self-ensembling to semantic segmentation is very difficult because heavily-tuned manual data augmentation used in self-ensembling is not useful to reduce the large domain gap in the semantic segmentation. To overcome this limitation, we propose a novel framework consisting of two components, which are complementary to each other. First, we present a data augmentation method based on Generative Adversarial Networks (GANs), which is computationally efficient and effective to facilitate domain alignment. Given those augmented images, we apply self-ensembling to enhance the performance of the segmentation network on the target domain. The proposed method outperforms state-of-the-art semantic segmentation methods on unsupervised domain adaptation benchmarks.

1. Introduction

Semantic segmentation has been widely studied in the computer vision field. Its goal is to assign image category labels to each pixel in the image. A wide variety of algorithms based on deep neural networks have achieved high performance with sufficient amounts of annotated datasets. However, creating large labeled datasets for semantic segmentation is cost-expensive and time-consuming [7]. To overcome the annotation burden, researchers utilize modern computer graphics to easily generate synthetic images with ground truth labels [36]. Unfortunately, in practice,

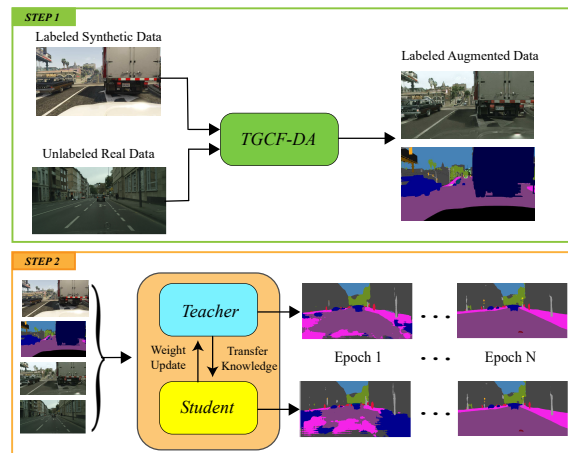


Figure 1. The overall framework of our method. Given labeled synthetic data and unlabeled real data, we propose a Target-Guided and Cycle-Free Data Augmentation (TGCF-DA) method to generate labeled augmented data (green). We introduce two segmentation networks as the teacher and the student in order to implement the self-ensembling algorithm (orange). Both segmentation networks are trained by augmented data as well as synthetic and real data. During the learning process, the teacher network transfers its knowledge to the student network.

models trained with synthetic data do not perform well on a realistic domain because there exists a distribution difference called domain shift. Unsupervised domain adaptation handles the domain shift by transferring knowledge from the labeled dataset in the source domain to the unlabeled dataset in the target domain [3].

Recent approaches for domain adaptation focus on aligning features extracted from the source and target data. In particular, most of the domain adaptation methods in semantic segmentation depend on adversarial training aiming to minimize the domain discrepancy through domain confusion [15, 14, 44, 41, 16, 52]. However, adversarial approaches suffer from a significant drawback. Since these methods seek to align the global distributions of two dif-

ferent domains, the adversarial loss may trigger a negative transfer, which aligns the target feature with the source feature in an incorrect semantic category. The negative transfer can have adverse effect on features that are already well aligned. Thus, this adaptation often performs even worse than a network trained solely on the source domain. Instead of adversarial training, we take an alternative way to perform feature-level domain alignment. We adopt self-ensembling [9], one of the effective methods for domain adaptation in classification.

Self-ensembling is composed of a teacher and a student network, where the student is compelled to produce consistent predictions provided by the teacher on target data. As the teacher is an ensembled model that averages the student’s weights, predictions from the teacher on target data can be thought of as the pseudo labels for the student. While recent self-ensembling proves its effectiveness in classification, these approaches require heavily-tuned manual data augmentation [9] for successful domain alignment. Furthermore, although such data augmentation consisting of various geometric transformations is effective in classification, it is not suited to minimize the domain shift in semantic segmentation. Two different geometric transformations on each input can cause spatial misalignment between the student and teacher predictions. Thus, we propose a novel data augmentation method to deal with this issue.

Our augmented image synthesis method is based on generative adversarial networks (GANs) [12]. We aim to generate augmented images, in which semantic contents are preserved, because these images with the inconsistent semantic content impair the segmentation performance due to the pixel-level misalignment between augmented images and source labels. Hence, we add a semantic constraint for the generator to preserve global and local structures, *i.e.* the semantic consistency. Furthermore, we propose a target-guided generator, which produces images conditioned on style information extracted from the target domain. In other words, our generator synthesizes augmented images maintaining semantic information, while only transferring styles from target images.

Most previous studies for GAN-based Image-to-Image translation methods [53, 49, 27, 25, 21, 19, 31] rely on various forms of cycle-consistency. However, incorporating cycle-consistency into unsupervised domain adaptation has two limitations. First, it needs redundant modules such as a target-to-source generator and corresponding computational burden. Second, cycle-consistency may be too strong when target data are scarce compared to source data [17], which is the general setting of unsupervised domain adaptation. Our proposed model does not consider all kinds of cycle-consistency. We refer to our method as Target-Guided and Cycle-Free Data Augmentation (TGCF-DA).

Our universal framework is illustrated in Fig. 1. We em-

ploy TGCF-DA to produce augmented images. Then, the segmentation network learns from the source, target and augmented data through self-ensembling. The main contributions of this paper are summarized as follows:

- We propose a novel data augmentation method with a target-guided generator and a cycle-free loss which is more efficient and suitable for semantic segmentation in unsupervised domain adaptation.
- We build a unified framework that collaborates the self-ensembling with TGCF-DA.
- Our approach achieves the state-of-the-art performances on challenging benchmark datasets. Also, we conduct extensive experiments and provide comprehensive analyses for the proposed method.

2. Related work

Unsupervised Domain Adaptation for Semantic Segmentation: Recently unsupervised domain adaptation for semantic segmentation has received much attention. The first attempt to this task is FCNs in the wild [15], which simultaneously performs the global and local alignment with adversarial training. Adversarial training is the predominant approach focusing on a feature-level adaptation to generate domain-invariant features through domain confusion, *e.g.*, [6, 5, 44, 41, 16, 39, 18]. This idea is extended to jointly adapt representations at both pixel and feature level through various techniques such as cycle-consistency loss [14, 34] or style transfer [8, 47]. Except for adversarial training methods, there is a different approach based on self-training. CBST [57] introduces self-training to produce pseudo labels and retrain the network with these labels.

Self-Ensembling: Self-Ensembling [56, 38] is proposed in the field of semi-supervised learning. A popular method for semi-supervised learning is the consistency regularization, which employs unlabeled data to produce consistent predictions under perturbations [40, 2]. Laine and Aila [24] propose Temporal Ensembling using a per-sample moving average of predictions for the consistent output. Tarvainen and Valpola [43] suggest an exponential moving average of the model weights instead of average of predictions. The self-ensembling method [9] applies a Mean Teacher framework to unsupervised domain adaptation with some modifications. In [35], Perone *et al.* address medical imaging segmentation tasks by applying the self-ensembling method akin to the previous method. Yonghao *et al.* [48] utilize the self-ensembling attention network to extract attention-aware features for domain adaptation.

Image-to-Image Translation: Recent approaches for Image-to-Image (I2I) Translation are based on Generative Adversarial Networks (GANs) [12]. In the case of unpaired training images, one popular constraint is cycle-consistency

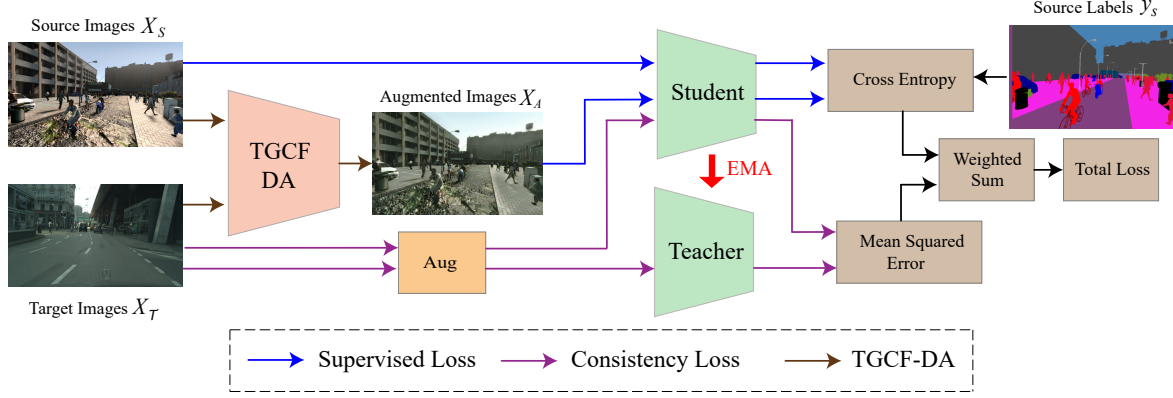


Figure 2. An overview of the proposed framework. 1) The source and target images are fed into Target-Guided generator to produce augmented images. 2) The supervised loss is a multi-class cross-entropy loss with source images and augmented images. 3) The consistency loss is a mean squared error between both prediction maps extracted from the student and teacher network. 4) A total training loss is the weighted sum of the supervised loss and the consistency loss. 5) We perform data augmentation only for target samples to complement the consistency loss. 6) The teacher network’s weights are the exponential moving average (EMA) of those of the student network.

that maps a given image to the target domain and reconstructs the original image [23, 53, 49]. UNIT [27] introduces a constraint for learning a shared latent space. However, all the aforementioned methods suffer from a lack of diversity in translated images. To produce multi-modal outputs, one possible approach injects noise vectors as additional inputs to the generator [54, 1, 11], but it could lead to the mode collapse problem. Also, Since cycle-consistency is too restrictive, variants of cycle-consistency [54, 21, 25] are developed for multi-modal I2I translation. A different approach is to apply neural style transfer [10, 45, 22, 20]. In particular, concurrent works [21, 31] employ an adaptive instance normalization [20] to transfer style from the exemplar to the original image. In addition, the authors of AugGAN [19] exploit the segmentation information for improving I2I translation network. Our task is entirely different from AugGAN because domain adaptation cannot use segmentation labels of the target data.

3. Proposed Method

In this work, we introduce the unified framework, which is built upon the self-ensembling for semantic segmentation. The key to improve the capacity of the self-ensembling for semantic segmentation is the GAN-based data augmentation to align representations of source and target rather than geometric transformations mostly used in existing self-ensembling for classification. To achieve this goal, we present a novel Target-Guided and Cycle-Free Data Augmentation (TGCF-DA) with a target-guided generator and a semantic constraint. The target-guided generator translates source images to different styles in the target domain. Our student network learns the source images and augmented images from TGCF-DA with a supervised loss by comput-

ing cross-entropy loss. Also, we only use target samples to compute the consistency loss, which is defined as the mean squared error between prediction maps generated from the student and teacher networks.

More formally, let X_S and X_T denote the source domain and target domain. We have access to N_s labeled source samples $\{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ with $x_s^i \in X_S$ and the corresponding label maps y_s^i . The target domain has N_t unlabeled target samples $\{x_t^i\}_{i=1}^{N_t}$, where $x_t^i \in X_T$. P_S and P_T denote the source and target data distributions, respectively. The source and target data share C categories. Let f_S and f_T be a student segmentation network and a teacher segmentation network.

3.1. Target-guided generator

Based on the assumption that image can be decomposed into two disentangled representations [27, 21], a content and a style, we adopt a source encoder for generating content representation and a target encoder for extracting style representation. To combine these two representations properly, we apply Adaptive Instance Normalization (AdaIN) [20] to feature maps of source images. As in [21], the target encoder with multiple fully connected layers provide the learnable affine transformation parameters (γ_t, β_t) to normalize the feature maps of a source image for each channel. The AdaIN operation is defined as:

$$\tilde{F}_s^i = \gamma_t^i \left(\frac{F_s^i - \mu(F_s^i)}{\sigma(F_s^i)} \right) + \beta_t^i, \quad (1)$$

where F_s^i denotes the source feature map for the i -th channel. $\mu(\cdot)$ and $\sigma(\cdot)$ respectively denote mean and variance across spatial dimensions. Our generator is guided by the style information of target samples through AdaINs at

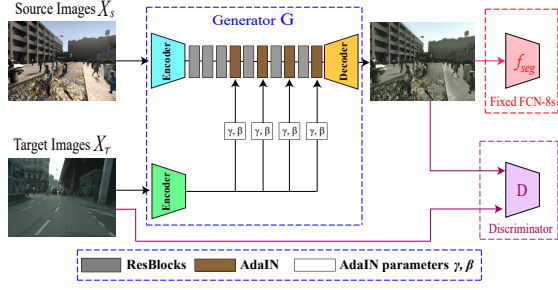


Figure 3. The overview of TGCF-DA based on GAN [12]. The blue box describes the target-guided generator G . The red box is the pretrained segmentation model f_{seg} with fixed weights. The purple box is the discriminator D .

intermediate residual blocks while preserving the spatial structure of source images, *i.e.* the semantic consistency of source images is retained.

3.2. Semantic constraint

We utilize a semantic constraint to preserve semantic content at pixel level. Given the labeled source data, we can pretrain the segmentation model such as FCN-8s [29] for constraining the generator. The pretrained segmentation model f_{seg} with fixed weights encourages the semantic consistency between the images before and after the translation. Thanks to this semantic constraint, our network can preserve the objects in images without distortion. Furthermore, this constraint is crucial to stabilizing the adversarial training without the cycle-consistency. Since the cycle-consistency enforces the strict constraint for matching two distributions, it is effective to prevent the mode collapse and to stabilize the adversarial training [26]. Without the cycle-consistency, our adversarial training is vulnerable to instability of GAN training. However, this semantic constraint guarantees stable adversarial training by strongly enforcing semantic consistency. We define the semantic constraint loss as the cross-entropy loss:

$$L_{sem}(f_{seg}, G) = \frac{-1}{HW} \sum_{k=1}^{H \times W} \sum_{c=1}^C y_s^{(k,c)} \log(f_{seg}(G(x_s, x_t))^{(k,c)}), \quad (2)$$

where $G(x_s, x_t)$ is the generated image of size $H \times W$ produced by the target-guided generator G .

3.3. Target-guided and cycle-free data augmentation

We introduce a GAN designed for Target-Guided and Cycle-Free Data Augmentation (TGCF-DA). As in Fig. 3, G is the target-guided generator and D is the discriminator proposed in [46]. We use the adversarial objective from

LSGAN [32] and apply the spectral normalization [33] to stabilize the GAN training. The GAN loss is defined as:

$$L_{GAN}(G, D) = E_{(x_s, x_t) \sim (P_S, P_T)} [D(G(x_s, x_t))^2] + E_{x_t \sim P_T} [(D(x_t) - 1)^2]. \quad (3)$$

This loss ensures that G produces new images visually similar to target images without losing semantic content in source images. Since the segmentation model f_{seg} is fixed, we jointly train the target-guided generator and discriminator to optimize the overall loss:

$$L_{TGCF-DA} = L_{GAN} + \lambda_{sem} L_{sem}, \quad (4)$$

where λ_{sem} is a weight to balance the contribution of the GAN loss and semantic constraint. The pretrained target-guided generator is employed to synthesize augmented images with the purpose of data augmentation in the self-ensembling.

3.4. Self-ensembling

We construct the teacher network f_T and the student network f_S . The teacher's weights t_i at training step i are updated by the student's weights s_i following the formula:

$$t_i = \alpha t_{i-1} + (1 - \alpha) s_i, \quad (5)$$

where α is an exponential moving average decay. During training, each mini-batch consists of source samples, augmented samples, and target samples. We use source samples and augmented samples to compute the supervised loss L_{sup} , which is cross-entropy function for semantic segmentation. This loss function enables the student network to produce the semantically accurate prediction for the source and augmented samples. The consistency loss L_{con} is formulated as the mean-squared error between the prediction maps generated from the student and teacher network:

$$L_{con}(f_S, f_T) = E_{x_t \sim P_T} [\|\sigma(f_S(x_t)) - \sigma(f_T(x_t))\|^2], \quad (6)$$

where σ is a softmax function to compute probability of prediction maps. The total loss L_{total} is the weighted sum of the supervised loss L_{sup} and the consistency loss L_{con} :

$$L_{total} = L_{sup} + \delta_{con} L_{con}, \quad (7)$$

where δ_{con} is the weight of consistency loss subject to the ramp-ups. Contrary to [9], we empirically observe that weight ramp-up is necessary for enhancing the effectiveness of the consistency loss.

3.5. Data augmentation for target samples

Here, data augmentation for target samples is not relevant to TGCF-DA. This data augmentation is only applied

to target samples in order to compute the consistency loss for the self-ensembling in Section 3.4. In classification [9], the goal of random data augmentations for target samples is forcing the student network to produce different predictions for the same target sample. Aforementioned above, image-level transformations such as geometric transformations are not helpful for the pixel-level prediction task like semantic segmentation [28]. Thus, we inject Gaussian noise to target samples, which are fed to student and target networks respectively. In addition, we apply Dropout [42] for weight perturbation. As a result, our student network is forced to produce consistent predictions with the teacher network under different perturbations for target samples and parameters of each network.

4. Experiments

This section describes experimental setups and details of synthetic-to-real domain adaptation. Then, we will report the experiment results compared with the previous researches. Furthermore, we will provide ablation studies to validate the effectiveness of our method.

4.1. Datasets

For a synthetic source domain, we used SYNTHIA [37] and GTA5 [36] datasets. Then, we evaluated our method on Cityscapes dataset [7] as a real-world target domain following similar settings in [15, 51, 44, 41]. We briefly introduce the details of datasets as following:

GTA5. GTA5 [36] contains 24966 urban scene images with pixel-level annotations. These high-resolution images are rendered from the gaming engine Grand Theft Auto V. Following [15], we used the 19 categories of the annotations compatible with those of the Cityscapes. We randomly picked 1000 images from GTA5 for validation purpose.

SYNTHIA. SYNTHIA [37] is a large-scale dataset of video sequences rendered from a virtual city. We used SYNTHIA-RAND-CITYSCAPES, consisting of 9400 images with pixel-level annotations. Inheriting from the previous work [51], we chose 16 categories common in both SYNTHIA and Cityscapes. We randomly selected 100 images for evaluation.

Cityscapes. Cityscapes [7] contains urban street scenes collected from 50 cities around Germany and neighboring countries. It has a training set with 2975 images and a validation set with 500 images.

We can utilize source images and labels from either SYNTHIA or GTA5, as well as target images without labels from the training set of Cityscapes. The validation set in Cityscapes is treated as the evaluation set for our domain adaptation experiment. We report IoU (Intersection-over-Union) for each class and mIoU (mean IoU) to measure the segmentation performance. In supplementary material, we

provide additional experimental results on the BDD100K dataset [50].

4.2. Experiment setup and implementation details

TGCF-DA. Our augmentation network for TGCF-DA is composed of the generator, the discriminator and the segmentation model. The generator is built upon the auto-encoder architecture used by MUNIT [21], but modified to act as the cycle-free generator. It consists of the source encoder, the target encoder and the decoder. The source encoder includes strided convolutional layers to downsample the source images and residual blocks [13] to compute the content representations. The decoder consists of residual blocks and transposed convolutional layers to upsample the combined representations. The target encoder is comprised of strided convolutional layers and fully connected layers to provide the style representations. Multi-scale discriminators described in [46] are employed as our discriminator. We set the weight λ_{sem} to 10 in all experiments.

Self-Ensembling. In all our experiments, we employed a VGG-16 backbone for our semantic segmentation network. Following Deeplab [4], we incorporated ASPP (Atrous Spatial Pyramid Pooling) as the decoder and then used an upsampling layer to get the final segmentation output. Before the upsampling layer, the output of the final classifier is used to compute the consistency loss in Section 3.4. Motivated by [43], we utilized the sigmoid ramp-up for the consistency loss weight δ_{con} . The details of the consistency loss weight is analyzed in Section 5.3. During training process, the images are resized and cropped to 480×960 resolution, and for evaluation we upsample our prediction maps to 1024×2048 resolution. The details of our architecture and experiments will be available in the supplementary material.

4.3. Experimental results

We report experimental results of the proposed method on two adaptation experiments in Table 1. We compare our proposed method with Curriculum DA [51], CyCADA [14], MCD [39], LSD-seg [41], AdaptSegNet [44], ROAD [5], Conservative Loss [55], DCAN [47], and CBST [57]. In Table 1, **Self-Ensembling (SE)** represents the segmentation performance of the network trained by source and target through the self-ensembling, without our data augmentation method. **TGCF-DA** indicates the segmentation network trained by the source data and augmented data generated from TGCF-DA with corresponding labels. **Ours (TGCF-DA + SE)** denotes our proposed framework comprised of TGCF-DA and the self-ensembling method. The proposed method significantly outperforms the baseline by 14.2% on GTA5→Cityscapes and 13.1% on SYNTHIA→Cityscapes. Our method makes further improvement compared to the source only baseline and also achieves the state-of-the-art mIoU scores on both experiments.

(a) GTA5 → Cityscapes																					
Method	Mech.	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bike	mIoU
Baseline (Source Only)	-	61.0	18.5	66.2	18.0	19.6	19.1	22.4	15.5	79.6	28.5	58.0	44.5	1.7	66.6	14.1	1.1	0.0	3.2	0.7	28.3
Curriculum DA [51]	ST	72.9	30.0	74.9	12.1	13.2	15.3	16.8	14.1	79.3	14.5	75.5	35.7	10.0	62.1	20.6	19.0	0.0	19.3	12.0	31.4
CyCADA [14]	AT	85.2	37.2	76.5	21.8	15.0	23.8	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
MCD [39]	AT	86.4	8.5	76.1	18.6	9.7	14.9	7.8	0.6	82.8	32.7	71.4	25.2	1.1	76.3	16.1	17.1	1.4	0.2	0.0	28.8
LSD-seg [41]	AT	88.0	30.5	78.6	25.2	23.5	16.7	23.5	11.6	78.7	27.2	71.9	51.3	19.5	80.4	19.8	18.3	0.9	20.8	18.4	37.1
AdaptSegNet [44]	AT	87.3	29.8	78.6	21.1	18.2	22.5	21.5	11.0	79.7	29.6	71.3	46.8	6.5	80.1	23.0	26.9	0.0	10.6	0.3	35.0
ROAD [5]	AT	85.4	31.2	78.6	27.9	22.2	21.9	23.7	11.4	80.7	29.3	68.9	48.5	14.1	78.0	19.1	23.8	9.4	8.3	0.0	35.9
Conservative Loss [55]	AT	85.6	38.3	78.6	27.2	18.4	25.3	25.0	17.1	81.5	31.3	70.6	50.5	22.3	81.3	25.5	21.0	0.1	18.9	4.3	38.1
DCAN [47]	SR	82.3	26.7	77.4	23.7	20.5	20.4	30.3	15.9	80.9	25.4	69.5	52.6	11.1	79.6	24.9	21.2	1.3	17.0	6.7	36.2
CBST [57]	ST	66.7	26.8	73.7	14.8	9.5	28.3	25.9	10.1	75.5	15.7	51.6	47.2	6.2	71.9	3.7	2.2	5.4	18.9	32.4	30.9
Self-Ensembling (SE)	ST	76.4	16.7	71.5	13.0	13.1	17.5	17.3	8.3	76.5	16.3	67.4	42.5	10.4	78.1	27.9	37.2	0.0	22.2	7.4	32.6
TGCF-DA	AT	73.9	19.8	74.8	19.7	21.8	20.7	26.7	12.4	78.0	22.3	72.0	53.4	12.9	73.3	24.5	28.5	0.0	24.3	14.1	35.4
Ours (TGCF-DA + SE)	AT+ST	90.2	51.5	81.1	15.0	10.7	37.5	35.2	28.9	84.1	32.7	75.9	62.7	19.9	82.6	22.9	28.3	0.0	23.0	25.4	42.5
Target Only	-	94.3	77.7	86.6	52.9	50.4	50.1	52.9	57.0	81.4	64.8	94.1	57.8	55.5	87.6	79.0	56.1	19.6	45.3	20.9	62.3

(b) SYNTHIA → Cityscapes																					
Method	Mech.	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	sky	person	rider	car	bus	motorbike	bike	mIoU	mIoU*		
Baseline (Source Only)	-	6.8	15.4	56.8	0.8	0.1	14.6	4.7	6.8	72.5	78.6	41.0	7.8	46.9	4.7	1.8	2.1	22.6	24.1		
Curriculum DA [51]	ST	65.2	26.1	74.9	0.1	0.5	10.7	3.7	3.0	76.1	70.6	47.1	8.2	43.2	20.7	0.7	13.1	29.0	34.8		
LSD-seg [41]	AT	80.1	29.1	77.5	2.8	0.4	26.8	11.1	18.0	78.1	76.7	48.2	15.2	70.5	17.4	8.7	16.7	36.1	-		
AdaptSegNet [44]	AT	78.9	29.2	75.5	-	-	-	0.1	4.8	72.6	76.7	43.4	8.8	71.1	16.0	3.6	8.4	-	37.6		
ROAD [5]	AT	77.7	30.0	77.5	9.6	0.3	25.8	10.3	15.6	77.6	79.8	44.5	16.6	67.8	14.5	7.0	23.8	36.2	-		
Conservative Loss [55]	AT	80.0	31.4	72.9	0.4	0.0	22.4	8.1	16.7	74.8	72.2	50.9	12.7	53.9	15.6	1.7	33.5	34.2	40.3		
DCAN [47]	SR	79.9	30.4	70.8	1.6	0.6	22.3	6.7	23.0	76.9	73.9	41.9	16.7	61.7	11.5	10.3	38.6	35.4	-		
CBST [57]	ST	69.6	28.7	69.5	12.1	0.1	25.4	11.9	13.6	82.0	81.9	49.1	14.5	66.0	6.6	3.7	32.4	35.4	36.1		
Self-Ensembling (SE)	ST	40.1	19.6	75.2	2.6	0.2	23.2	4.0	9.8	60.3	38.3	49.1	14.0	67.0	17.4	6.4	11.9	27.5	29.2		
TGCF-DA	AT	63.9	25.6	75.9	5.4	0.1	22.6	2.6	6.8	78.4	77.2	48.7	16.5	62.2	24.2	5.0	22.1	33.6	39.8		
Ours (TGCF-DA + SE)	AT+ST	90.1	48.6	80.7	2.2	0.2	27.2	3.2	14.3	82.1	78.4	54.4	16.4	82.5	12.3	1.7	21.8	38.5	46.6		
Target Only	-	89.2	85.3	90.7	65.5	60.7	21.5	2.1	7.2	74.2	93.2	61.8	40.1	78.4	81.4	36.7	24.8	57.1	64.1		

Table 1. The semantic segmentation results on Cityscapes validation set when evaluating the model trained on (a) GTA5 and (b) SYNTHIA. All segmentation models in table use VGG-16 based models. The mIoU* denotes the segmentation results over the 13 common classes. “Source Only” denotes the evaluation result of models only trained on source data. “Target Only” denotes the segmentation results in supervised settings. The mechanism “AT”, “ST” and “SR” stand for adversarial training, self-training, and style transfer respectively.

4.4. Ablation studies

Ablation for Self-Ensembling (SE): Comparing the baseline and SE, SE shows small improvement in mIoUs by 4.3% in Table 1-(a) and by 4.9% in Table 1-(b). However, in details, we observe that SE does not perform well during the whole training process as shown in Fig. 4 (blue and orange lines). In contrast to our proposed method (TGCF-DA + SE), the teacher and student networks do not maintain complementary correlations.

Ablation for TGCF-DA: TGCF-DA is necessary to generate synthetic data, which help the network reduce the domain shift. Compared to the baseline, TGCF-DA improves the mIoUs by 7.1% in Table 1-(a) and by 11.0% in Table 1-(b). Such improvements validate that TGCF-DA serves as a useful way to reduce the domain shift. Except for TGCF-DA, SE shows the poor results in both experiments. On the contrary, our proposed method in Fig. 4 (grey and yellow lines) clearly demonstrates that the teacher updated by the student continues to improve segmentation capability, and successfully transfer its knowledge to the student. As a re-

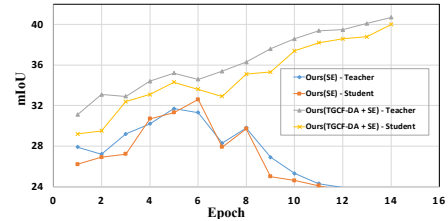


Figure 4. The testing mIoUs of SE (blue and orange) and our method (grey and yellow) *w.r.t* training epochs on the GTA5 → Cityscapes experiment.

sult, the teacher and student of our method enhance their performance simultaneously. These results substantiate our intuition that TGCF-DA enhances the capability of the self-ensembling algorithm for semantic segmentation.

5. Analysis

In this section, we provide visualization results and analysis on various components of our proposed framework.

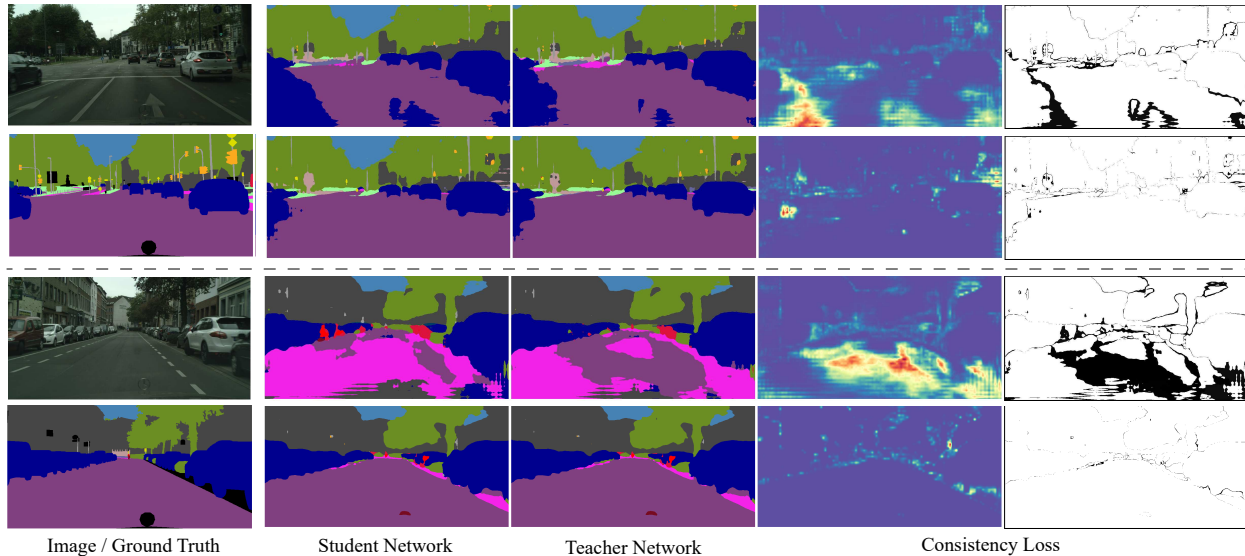


Figure 5. Visualization results of GTA5 → Cityscapes (first and second rows) and SYNTHIA → Cityscapes (third and fourth rows). Segmentation results at 10K training steps (first and third rows) and 56K training steps (second and fourth rows). The fourth and fifth columns illustrate the heatmap of the consistency loss and disagreement map between the student and teacher networks.

5.1. Visualization

The effectiveness of the self-ensembling is visualized in Fig. 5. We validate that the teacher network generates better predictions, and then different predictions between the teacher and student networks cause consistency loss to enforce the consistency of their predictions. In Fig. 5, the first and third rows show that predictions of the teacher can be a good proxy for training the student network early in the training. In addition, we point out that the consistency loss concentrates on the boundary of each object in the later training stage. Hence, the consistency loss can play a role in refining boundaries of semantic objects where the segmentation model are likely to output wrong predictions.

In Fig. 7, we show the example results of TGCF-DA compared with other Image-to-Image (I2I) translation methods: CycleGAN [53], UNIT [27], and MUNIT [21]. Both CycleGAN and UNIT often generate distorted images containing corrupted objects and artifacts. MUNIT is capable of preserving objects in images, but we observe that the style of the majority classes in the target image is often matched to elements of different classes in the source image, which is similar to “spills over” problem in [30]. For example, the translated image from MUNIT shows artifacts in the sky like road texture of the target domain. Compared to the methods mentioned above, our method is not only computationally cheap and memory efficient due to the cycle-free loss but also demonstrating compelling visual results with preserving semantic consistency.

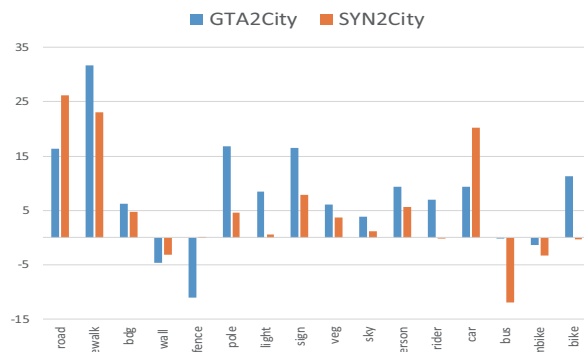


Figure 6. Per-class IoU gains through the self-ensembling. The blue bar represents per-class IoU gains in GTA5→Cityscapes experiment. The orange bar indicates the per-class IoU gains in SYNTHIA→Cityscapes experiment.

5.2. Analysis of self-ensembling with per-class IoUs

To better understand the self-ensembling, we compare the per-class IoUs of our method with and without the self-ensembling. In Fig. 6, we show the per-class IoU gains between TGCF-DA and Ours (TGCF-DA + SE). Although the IoU scores in the most categories are generally improved, there is a difference in performance gains among different categories. Figure 6 demonstrates that the IoU gains in majority classes (such as “road”) are generally better than those in minority classes (like “bus”). These experimental results are attributed to the self-ensembling and class imbalance issues. Due to the class imbalance, the segmentation network often produces incorrect predictions on minority

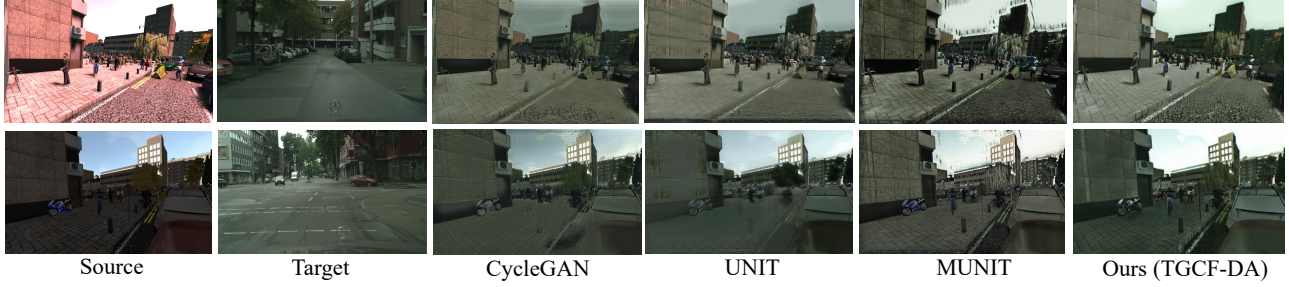


Figure 7. Example images of SYNTHIA synthesized in the style of Cityscapes with CycleGAN [53], UNIT [27], and MUNIT [21].

	Ramp-up coefficient δ_0				EMA decay α		
	1	3	30	50	0.9	0.99	0.999
GTA5	41.3	42.3	42.5	33.6	37.6	38.9	42.5
SYN	35.4	36.1	38.5	32.5	36.2	38.5	37.8

Table 2. Hyperparameter sensitivity. GTA5 denotes $\text{GTA5} \rightarrow \text{Cityscapes}$ experiment and SYN denotes $\text{SYNTHIA} \rightarrow \text{Cityscapes}$ experiment.

classes [57]. In the self-ensembling method, this effect can be strengthened because the student is iteratively learned from predictions of the teacher, which tends to make incorrect predictions on minority classes rather than majority classes. Thus, the self-ensembling gives rise to large improvements in per class IoUs of majority classes compared to minority classes. It is worth noting that this result accords with our intuition that predictions of the teacher network serve as pseudo labels for the student network.

5.3. Hyperparameter sensitivity on self-ensembling

In the self-ensembling, the consistency loss weight δ and the exponential moving average (EMA) decay α are important hyperparameters. We conduct the experiments to explore the sensitivity of these hyperparameters. Table 2 shows that setting a proper value for the EMA decay is significant. In all our experiments, the EMA decay is 0.99 during the first 37K iterations, and 0.999 afterward. The teacher benefits from new and accurate student’s weight early in the training because the student improves its segmentation capacity rapidly. On the other hand, since the student improves slowly in the later training, the teacher can gain knowledge from the old ensembled model.

The consistency loss weight δ follows the formula $\delta = 1 + \delta_0 e^{-5(1-x)^2}$, where $x \in [0, 1]$ denotes the ratio between the current epoch and the whole epochs and δ_0 is a ramp-up coefficient. Different from the usual sigmoid ramp-up [43], we add one to the formula because it is essential to guarantee the contribution of the consistency loss at the beginning of training. We decide to use $\delta_0 = 30$ for all our experiments.



Figure 8. The change of augmented images *w.r.t* the value of weight λ_{seg} . From left to right: source input, output with $\lambda_{seg} = 1$, output with $\lambda_{seg} = 10$.

5.4. Hyperparameter sensitivity on TGCF-DA

The weight λ_{sem} for the semantic constraint is a hyperparameter for training our augmentation network. Figure 8 shows some example results on $\text{SYNTHIA} \rightarrow \text{Cityscapes}$. When we use a lower value ($\lambda_{sem} = 1$) for semantic constraint, the generator is prone to mix up objects and scenes in the augmented images. On the other hand, the proper value for semantic constraint ($\lambda_{sem} = 10$) helps the network preserve the local and global structures of images. These results confirm that the semantic constraint enforces our augmentation network to retain semantic consistency.

6. Conclusion

We have proposed a novel framework comprised of two complementary approaches for unsupervised domain adaptation for the semantic segmentation. We present the GAN-based data augmentation with the guidance of target samples. Without the use of cycle consistency, our augmentation network produces augmented images for domain alignment. Moreover, the self-ensembling with those augmented images can perform successful adaptation by transferring pseudo labels from the teacher network to the student network. Experimental results verify that our proposed model is superior to existing state-of-the-art approaches.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2018R1A5A7025409).

References

- [1] Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented CycleGAN: Learning many-to-many mappings from unpaired data. In *Proceedings of the 35th International Conference on Machine Learning*, pages 195–204, 2018.
- [2] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. 2018.
- [3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.
- [4] 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, 2018.
- [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.
- [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.
- [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.
- [8] Aysegül Dundar, Ming-Yu Liu, Ting-Chun Wang, John Zedlewski, and Jan Kautz. Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation. *arXiv preprint arXiv:1807.09384*, 2018.
- [9] Geoff French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018.
- [10] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.
- [11] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Advances in Neural Information Processing Systems*, pages 1294–1305, 2018.
- [12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [13] 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.
- [14] 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 *Proceedings of the 35th International Conference on Machine Learning*, pages 1989–1998, 2018.
- [15] 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.
- [16] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2018.
- [17] Ehsan Hosseini-Asl, Yingbo Zhou, Caiming Xiong, and Richard Socher. Augmented cyclic adversarial learning for low resource domain adaptation. In *International Conference on Learning Representations*, 2019.
- [18] Haoshuo Huang, Qixing Huang, and Philipp Krahenbuhl. Domain transfer through deep activation matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 590–605, 2018.
- [19] Sheng-Wei Huang, Che-Tsung Lin, Shu-Ping Chen, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai. Auggan: Cross domain adaptation with gan-based data augmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 718–731, 2018.
- [20] 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.
- [21] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.
- [22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
- [23] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org, 2017.
- [24] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.
- [25] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018.
- [26] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. Alice: Towards understanding adversarial learning for joint distribution matching. In *Advances in Neural Information Processing Systems*, pages 5495–5503, 2017.
- [27] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

- [28] Shuangting Liu, Jiaqi Zhang, Yuxin Chen, Yifan Liu, Zengchang Qin, and Tao Wan. Pixel level data augmentation for semantic image segmentation using generative adversarial networks. *arXiv preprint arXiv:1811.00174*, 2018.
- [29] 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.
- [30] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4990–4998, 2017.
- [31] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. In *International Conference on Learning Representations*, 2019.
- [32] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.
- [33] Takeru Miyato, Toshiaki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.
- [34] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.
- [35] Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *arXiv preprint arXiv:1811.06042*, 2018.
- [36] 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.
- [37] 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.
- [38] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *Proceedings of the Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION’05) - Volume 1 - Volume 01*, pages 29–36. IEEE Computer Society, 2005.
- [39] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.
- [40] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 1163–1171, 2016.
- [41] 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.
- [42] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [43] 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.
- [44] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schuster, 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.
- [45] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
- [46] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.
- [47] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–534, 2018.
- [48] Yonghao Xu, Bo Du, Lefei Zhang, Qian Zhang, Guoli Wang, and Liangpei Zhang. Self-ensembling attention networks: Addressing domain shift for semantic segmentation. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.
- [49] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2849–2857, 2017.
- [50] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.
- [51] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017.
- [52] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6810–6818, 2018.

- [53] 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.
- [54] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017.
- [55] 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.
- [56] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.
- [57] 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.