The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

Addressing Domain Gap via Content Invariant Representation for Semantic Segmentation

Li Gao, [1] **Lefei Zhang,** [1*] **Qian Zhang** [2]

[1] National Engineering Research Center for Multimedia Software,
Institute of Artificial Intelligence, School of Computer Science and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan, China. [2] Horizon Robotics, Inc., Beijing, China.
gaoli1218@whu.edu.cn, zhanglefei@whu.edu.cn, qian01.zhang@horizon.ai

Abstract

The problem of unsupervised domain adaptation in semantic segmentation is a major challenge for numerous computer vision tasks because acquiring pixel-level labels is time-consuming with expensive human labor. A large gap exists among data distributions in different domains, which will cause severe performance loss when a model trained with synthetic data is generalized to real data. Hence, we propose a novel domain adaptation approach, called Content Invariant Representation Network, to narrow the domain gap between the source (S) and target (T) domains. The previous works developed a network to directly transfer the knowledge from the S to T. On the contrary, the proposed method aims to progressively reduce the gap between S and T on the basis of a Content Invariant Representation (CIR). CIR is an intermediate domain (I) sharing invariant content with S and having similar data distribution to T. Then, an Ancillary Classifier Module (ACM) is designed to focus on pixel-level details and generate attention-aware results. ACM adaptively assigns different weights to pixels according to their domain offsets, thereby reducing local domain gaps. The global domain gap between CIR and T is also narrowed by enforcing local alignments. Last, we perform self-supervised training in the pseudo-labeled target domain to further fit the distribution of the real data. Comprehensive experiments on two domain adaptation tasks, that is, GTAV → Cityscapes and SYNTHIA → Cityscapes, clearly demonstrate the superiority of our method compared with state-of-the-art methods.

Introduction

The past few years have witnessed the rapid development of deep learning, which made a vital effect on many computer vision tasks, such as semantic segmentation. Semantic segmentation aims to assign each pixel of a photograph to a semantic class label to distinguish different things on the image. That is, semantic segmentation can be understood as a pixel-level classification task.

For computer-generated images, obtaining their semantic labels is very simple, but acquiring the labels of real pictures requires expensive human labor (Luo et al. 2018). Moreover, in some scenarios, such as autonomous cars and industry robots, huge amount of real images with accurate labels are indispensable for training. In the case that the number of labeled real images is not enough, utilizing synthetically generated data is one promising approach that addresses the above issues (Richter et al. 2016; Ros et al. 2016). Nevertheless, models trained with the synthetic images perform unsatisfactorily when applied to a realistic domain because of the existence of cross-domain discrepancy (Shimodaira 2000).

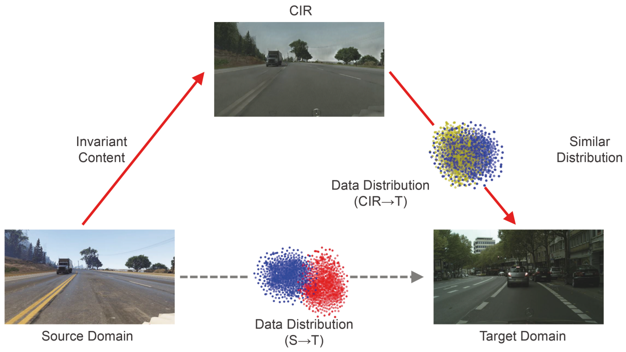

Figure 1: Connection of CIR with source and target domains. S and T denote source and target domain, respectively. The image from S has the same content with its corresponding CIR image. The blue, red, and yellow dots represent the distribution of T, S, and CIR, respectively. CIR has a similar distribution to T.

To solve the aforementioned problems, unsupervised domain adaptation (UDA) approaches (Saenko et al. 2010) are proposed to reduce distribution offsets between source and target domains. Numerous domain adaptation methods (Hoffman et al. 2016; Sankaranarayanan, Balaji, and Jain 2018; Song et al. 2020; Zhang et al. 2019) are designed to learn the domain-invariant features of the source and target domains. Thus, when we have no access to the target domain labels, the model learned from the source domain can be applied to the target domain.

Some prior works (Lee et al. 2019; Chen et al. 2017; Zou et al. 2018; Chen et al. 2019; Zhao et al. 2019; Murez et al. 2018) tend to directly adapt the information of the source domain to the target domain. The distance between these two feature spaces is often very far, and many categories have

[*] Corresponding author.

major differences across different domains. Hence, the adaptation results generated by those works are not satisfactory. To overcome this limitation, we propose a Content Invariant Representation Network (CIRN) to narrow the domain gap between the source (S) and target (T) domains. CIRN constructs a Content Invariant Representation (CIR) by introducing an additional intermediate domain (I) for semantic segmentation. Specifically, CIR has the invariant content with S but different appearance, while sharing similar data distribution to T but diverges in image details. Therefore, CIR could serve as a good intermediator to connect S and T, as illustrated in Figure 1.

Our adaptation happens in three stages. First, we construct a CIR to make the adaptation task easier because aligning data distributions between two distant domains is challenging. That is, we reduce the domain gap from S to T bridged by CIR instead of directly solving the domain shift between S and T. In practice, the ways to construct the CIR could be numerous. In this work, we adopt a widely-used algorithm to transfer the style of images from S to T without changing their original content. Image translation algorithms make the overall style, lighting, tone, and other stylization factors of the source domain images look like the target domain images. However, they still cannot perfectly translate due to the absence of paired training data. Thus, we define the translated images as images of CIR. In addition, they form a new intermediate domain I, which has a smaller distance with T.

Second, we perform domain adaptation from CIR to T by introducing an ACM and a pixel-level adversarial loss. When stylized source images often fail to preserve details completely, these failure cases may lead CIR to have a large distance with the target distribution. Consequently, after the global domain adaptation in stage 1, we focus on the local pixel-level domain shift in stage 2 by assigning each pixel different weights. In other words, ACM aims to conduct global domain adaptation by guiding local domain adaptation.

Finally, generating the pseudo labels with the model trained in stage 2, we can finetune the network using a self-supervised learning strategy, which narrows the large distribution gap among the target data itself.

The main contributions of this work are summarized as follows:

- We propose a CIRN to address the problem of UDA-based semantic segmentation gradually, from global to local adaptation. Specifically, we construct a CIR lying between source and target domains to align the distribution shift based on an intermediate feature space. CIR shares invariant content and the same labels with the source domain and has similar data distribution to the target domain.
- We propose an ACM to help domain adaptation network focus on pixel-level details, thereby producing additional convinced results by introducing attention modules. ACM performs local domain alignment to adapt information from the intermediate domain to the target domain.
- Experiments conducted on two challenging benchmark datasets can validate the effectiveness of our proposed method against existing state-of-the-art approaches.

Related Work

This section focuses on adversarial training, image-level transferring, and curriculum style based approaches for UDA, which form the main motivations of our proposed method.

Adversarial Training Based Approaches

Adversarial training based approaches have been widely studied since FCNs-Wild (Hoffman et al. 2016). This kind of methods usually consists of a generator to predict the segmentation results and a discriminator to minimize the divergences between source and target domains. FCNs-Wild (Hoffman et al. 2016) is the first to adopt adversarial training for UDA semantic segmentation and not only aligns the global domain shift but also performs local alignment. Tsai et al. found multiple modes of patch-wise output distribution and thus proposed a method to learn discriminative feature representations of patches in the source domain (Tsai et al. 2019). ADVENT (Vu et al. 2019) enforced high prediction certainty (low-entropy) on target predictions by introducing an entropy adversarial loss to achieve domain adaptation. CLAN (Luo et al. 2019) took advantage of co-training by utilizing the discrepancy between two classifiers' outputs and proposed an adaptive adversarial loss to enforce domain alignment.

Image-Level Transferring Based Approaches

Image-level transferring based approaches are proposed to transfer the appearance of source or target images to make them visually similar. CyCADA (Hoffman et al. 2018) transferred the style of source images while enforcing cycle-consistency and leveraging segmentation loss. DCAN (Wu et al. 2018) explored statistics in each channel of feature maps by performing channel-wise feature alignment in an image translator and a segmentation classifier. Choi et al. raised a self-ensembling data augmentation method by transferring image style to facilitate domain alignment (Choi, Kim, and Kim 2019). Song et al. actively transferred the style from the target to the source domain to reduce the visual gap between them. They also introduced a perceptual loss to ensure image similarity (Song et al. 2020). Another related work considered the texture difference between the two domains and bridged the domain gap by diversifying the texture of synthetic images using two style transfer algorithms (Kim and Byun 2020). However, this method only considered global adaptation by transferring the style of source images twice, and bad stylized pixels may result in worse local adaptation. DISE (Chang et al. 2019) found an image representation, which comprises domain-invariant structure and domain-specific texture component and further realized image-translation across domains. CPN (Yang et al. 2020) and FDA (Yang and Soatto 2020) translated the style of source images through a simple Fourier Transform and its inverse. Dong et al. developed a transfer model to alternatively determine where and how to explore transferable domain-invariant knowledge between two domains

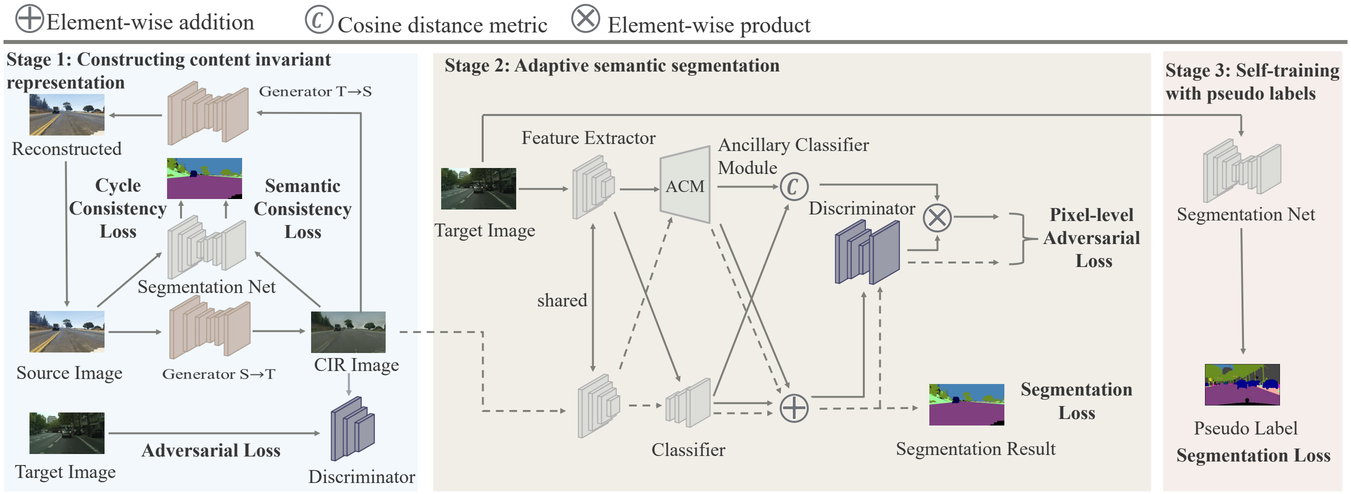

Figure 2: Framework of our CIRN. Our adaptation happens in three stages. First, we map the source domain to the target domain by transferring style with a CIR introducing. Second, we perform domain adaptation from CIR to the target domain by ACM and pixel-level adversarial loss. Finally, we can improve the framework with a self-supervised learning strategy.

(Dong et al. 2020). Yue et al. proposed to randomize the style of synthetic images with auxiliary datasets and introduced a pyramid consistency to learn domain-invariant representations (Yue et al. 2019). DLOW (Gong et al. 2019) bridged domain gap by generating a continuous sequence (style transferred images) flowing from one domain to the other.

Curriculum Style Based Approaches

Our work is also relevant to curriculum style based approaches (Zhang et al. 2020; Lian et al. 2019; Pan et al. 2020), which deal with easy tasks, such as inferring necessary properties about the target domain first. Zhang et al. initially gained some high-level properties about the unknown pixel-level labels for target images and then trained the semantic segmentation network (Zhang et al. 2020). PyCDA (Lian et al. 2019) constructed a pyramid curriculum, which contains various properties about the target domain. Those properties can improve the segmentation network's generalization capability to the target domain. Pan et al. conducted the inter-domain adaptation of the model first and then separated the target domain into an easy and hard split to reduce the intra-domain gap (Pan et al. 2020).

Content Invariant Representation Network

Overview of the Proposed Model

S denotes the source domain, containing source data X_S with pixel-level labels Y_S, and T denotes the target domain, where we only have access to the data X_T. We aim to train a semantic segmentation network to predict accurate labels for X_T. In this section, we present a network with three stages, which can progressively bridge the domain gap step by step. Figure 2 shows the framework of our proposed method. The first stage is constructing CIR by introducing an intermediate domain I. We apply CycleGAN (Zhu et al. 2017) to map the source domain images to the target distributions in the image-level. Then in the second stage, the domain gap between I and T is narrowed by training the adaptive semantic segmentation model. Finally, we can adopt the self-training strategy with direct supervision to obtain our final model based on the pseudo labels generated in stage 2.

Constructing Content Invariant Representation

We construct CIR by introducing an intermediate domain I (Hsu et al. 2020) with applying the image-to-image translation network CycleGAN (Zhu et al. 2017). Since the major contribution of our work is to propose a CIRN framework rather than develop a fixed approach to generate the CIR, we employ the CycleGAN to validate the effectiveness of CIR due to its simplicity and generality. In fact, CIR can still work well if another appropriate approach is employed to generate the CIR.

The objective is to map the source domain images to mimic the ones in the target domain, for ground truth labels are only available in the source domain. S and I have similar content but different appearance, whereas I and T diverge in image details, but distributions between them resemble. Thus, this generated domain serves as an intermediary to assist in reducing the adaptation difficulty when a large domain gap exists between S and T.

The goal of the first stage is to learn mapping functions between S and T. We have two mappings $G : S \to T$ and $F : T \to S$, two adversarial discriminators D_S and D_T. D_S aims to distinguish between source images $x_s \in X_S$ and translated images $F(x_t)$ $(x_t \in X_T)$, whereas D_T aims to discriminate between x_t and $G(x_s)$. We express the *adversarial loss* as:

$$\begin{aligned}\mathcal{L}_{GAN}(G, D_T, S, T) = {} & E_{x_t \sim X_T}[\log D_T(x_t)] \\ & + E_{x_s \sim X_S}[\log(1 - D_T(G(x_s))] \end{aligned} \quad (1)$$

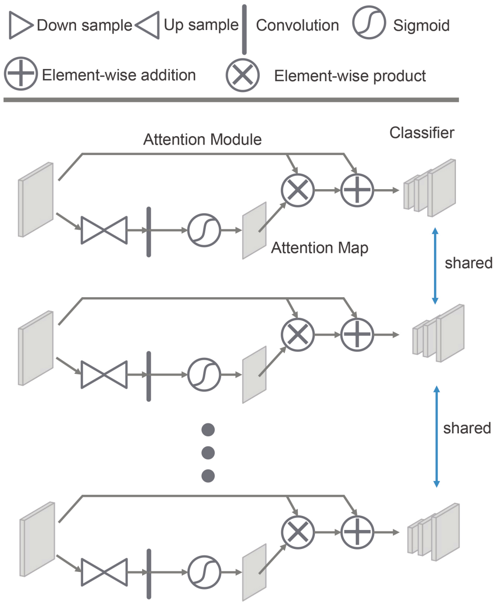

Figure 3: Framework of ACM, comprising K attention modules embedded in a classifier.

$$\begin{aligned}\mathcal{L}_{GAN}(F, D_S, T, S) = & E_{x_s \sim X_S}[\log D_S(x_s)] \\ & + E_{x_t \sim X_T}[\log(1 - D_S(F(x_t))] \end{aligned} \tag{2}$$

We impose a cycle-consistency constraint (Hoffman et al. 2018; Zhu et al. 2017) to prevent the learned mappings G and F from contradicting with each other to encourage the preservation of the content of the source images. The *cycle consistency loss* is:

$$\begin{aligned}\mathcal{L}_{cyc}(G, F, S, T) = & E_{x_s \sim X_S}[\|F(G(x_s)) - x_s\|_1] \\ & + E_{x_t \sim X_T}[\|F(G(x_t)) - x_t\|_1] \end{aligned} \tag{3}$$

where $\|\cdot\|_1$ denotes L1 norm.

In addition, we explicitly encourage high semantic consistency before and after image translation to ensure the accuracy of the semantic labels when trained in the second stage. Thus, we pretrain a source segmentation model and fix its weights to compute the *semantic consistency loss* to stimulate an image to be classified in the same way after translation as it was before translation according to this classifier (Hoffman et al. 2018; Li, Yuan, and Vasconcelos 2019). We express the objective as:

$$\mathcal{L}_{sem}(G, S) = -\sum_{i=1}^{H\times W}\sum_{c=1}^{C} y_s^{H\times W\times C}\log P_S^{H\times W\times C} \tag{4}$$

where H, W, and C represent the height, width and number of categories, respectively. $y_s \in Y_S$ is the label of the source domain and P_s is the predicted probability of $G(x_s)$.

With the above loss terms, the overall loss function of stage 1 can be written as:

$$\begin{aligned}\mathcal{L}_{stage1}(G, F, S, T, D_S, D_T) = & \mathcal{L}_{GAN}(G, D_T, S, T) \\ & + \mathcal{L}_{GAN}(F, D_S, T, S) \\ & + \mathcal{L}_{cyc}(G, F, S, T) \\ & + \mathcal{L}_{sem}(G, S) \end{aligned} \tag{5}$$

Then, the intermediate domain I is constructed, X_I consists of $G(x_s)$ ($x_s \in X_S$), and Y_I consists of y_s ($y_s \in Y_S$).

Adaptive Semantic Segmentation

The second stage aims to address the domain shift between I and T. Our network is composed of a feature extractor F, a discriminator D, a classifier C and an ACM. For an intermediate domain image $x_i \in X_I$ with its ground-truth label $y_i \in Y_I$. F transforms x_i to high-level semantic features. Then, C and ACM generate a prediction P_I, with which we can compute supervised cross-entropy loss. In the same time, P_I is fed to D to generate an adversarial loss. For a target image x_t, its prediction map is also generated by C and ACM. Different from the intermediate data process, we additionally compute their distance out of p^1 and p^2 (Luo et al. 2019), represented as $\mathcal{R}(p^1, p^2)$, where p^1 denotes the output of C, p^2 denotes the output of ACM, and $\mathcal{R}(\cdot)$ denotes cosine distance between them. Then, we perform an element-wise multiplication with $\mathcal{R}(p^1, p^2)$ and the output adversarial loss of D. Also, in order to make the classifier in ACM and C have different parameters, we impose a *weight discrepancy loss* following the work(Luo et al. 2019).

In stage 1, the global domain gap becomes smaller, so we have to focus more on local pixel-level divergence. Different regions in the images usually correspond to different levels of domain shift. Thus, those noteworthy regions deserve additional attention. Following the works (Chen et al. 2016; Wang et al. 2017; Xu et al. 2019), we introduce the attention module into ACM to learn attention-aware features. The framework of ACM is shown in Figure 3, comprising K attention modules embedding in a classifier, where the output of ACM is the sum of K modules. ACM will focus on the pixels with a large domain gap to help C supplement details because C focuses on feature-level information. The attention map ranges from $[0, 1]$, thereby playing the role of controlling features for input. If some pixels in attention maps have large values, then these pixels are of a vital level of domain gap. Under this circumstance, the network will be pushed to accelerate the alignment process of those pixels.

Our second stage has two loss functions, namely, *segmentation loss* and *pixel-level adversarial loss*. The *segmentation loss* is:

$$\mathcal{L}_{seg}(X_I) = -\sum_{i=1}^{H\times W}\sum_{c=1}^{C} y_i^{H\times W\times C}\log P_I^{H\times W\times C} \tag{6}$$

where P_I denotes the predicted probability map generated by the sum of the outputs of ACM and C. Following the work (Luo et al. 2019), in order to ensure the classifier in ACM and C to have as different views as possible and in the

GTAV → Cityscapes																				
Method	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	trunk	bus	train	motorbike	bike	mIoU
Source Only	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
UIDA	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	**85.7**	40.5	79.7	58.7	**31.1**	**86.3**	31.5	**48.3**	0.0	30.2	35.8	46.3
CLAN	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
ADVENT	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
DISE	91.5	47.5	82.5	31.3	25.6	33.0	33.7	25.8	82.7	28.8	82.7	**62.4**	30.8	85.2	27.7	34.5	6.4	25.2	24.4	45.4
CRST	91.0	**55.4**	80.0	**33.7**	21.4	**37.3**	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	**26.9**	26.0	42.3	47.1
WeekSegDA	**91.6**	47.4	84.0	30.4	**28.3**	31.4	**37.4**	**35.4**	83.9	38.3	83.9	61.2	28.2	83.7	28.8	41.3	8.8	24.7	**46.4**	48.2
Ours	91.5	48.7	**85.2**	33.1	26.0	32.3	33.8	34.6	85.1	**43.6**	**86.9**	62.2	28.5	84.6	37.9	47.6	0.0	**35.0**	36.0	**49.1**

Table 1: Results of domain adaptation task GTAV → Cityscapes over 19 classes.

same time make as similar predictions as possible, we compute the cosine distance between the outputs/parameters of ACM and C to focus on the pixels suffering from major domain shift. Benefiting from that, when the cosine distance $\mathcal{R}(p^1, p^2)$ between ACM and C is large, large weights will be assigned as to encourage segmentation model to fool discriminator D. In addition, the computation of distance corresponds to our ACM. The *pixel-level adversarial loss* is:

$$\begin{aligned}\mathcal{L}_{adv}(X_I, X_T) = -E[\log P_I] - \\ E[(\lambda_{cos}\mathcal{R}(p^1, p^2) + \epsilon)\log(1 - P_T)]\end{aligned} \tag{7}$$

In Equation 7, we add a small number ϵ to stabilize the training process. Then, the overall loss function of stage 2 can be formulated as:

$$\mathcal{L}_{stage2}(X_I, X_T) = \mathcal{L}_{seg}(X_I) + \lambda_{adv}\mathcal{L}_{adv}(X_I, X_T) \tag{8}$$

Self-Training with Pseudo Labels

In the third stage, we generate pseudo labels for target domain images with the model trained in stage 2. Then we calculate the supervised segmentation loss to finetune our adaptive semantic segmentation model:

$$\mathcal{L}_{stage3}(X_T) = -\sum_{i=1}^{H\times W}\sum_{c=1}^{C} \hat{y}_t^{H\times W\times C} \log P_T^{H\times W\times C} \tag{9}$$

where $\hat{y}_t$ represents the pseudo label and P_T represents the predicted map for target image.

Experiments

Datasets

Following the experimental setup of previous works (Tsai et al. 2019; Chang et al. 2019), we conduct extensive experiments on two adaptation tasks, that is, GTAV (Richter et al. 2016) to Cityscapes (Cordts et al. 2016) and SYNTHIA (Ros et al. 2016) to Cityscapes.

GTAV is a dataset, which contains 24,966 synthetic urban scene images with a resolution of $1,914 \times 1,052$. For training, we consider 19 common categories semantic labels to be compatible with the Cityscapes dataset.

SYNTHIA: SYNTHIA-RAND-CITYSCAPES is another photorealistic synthetic image dataset, which consists of 9,400 images with a resolution of $1,280 \times 760$. We validate on 16 common classes with the Cityscapes dataset, and the evaluation of 13 classes is also reported.

Cityscapes is a real-world collected dataset which provides 5,000 densely annotated images with $2,048 \times 1,024$ resolution. We use 2,975 training images for training and 500 validation images for testing.

Implementation Details

We implement the proposed framework using the PyTorch toolbox on a single Tesla V100 GPU with 16 GB memory. For stage 1, we adopt CycleGAN (Zhu et al. 2017) as our image-to-image translation network. The model is trained using Adam (Kingma and Ba 2014) optimizer with the initial learning rate of 2×10^{-4} and $\beta_1 = 0.5$, $\beta_2 = 0.999$. Batch-size is set to 1 for all stages. For stage 2, we adopt the DeepLab-v2 (Chen et al. 2018) framework with the ResNet-101 (He et al. 2016) architecture pretrained on ImageNet (Deng et al. 2009) as our segmentation backbone network. C and the classifier in ACM are copies of the last classification module of the backbone network. For D, we adopt a similar structure with CLAN (Luo et al. 2019), which consists of 5 convolution layers with kernel 4×4 with channel numbers $\{64, 128, 256, 512, 1\}$ and stride of 2. Each convolution layer is followed by a Leaky-ReLU activation parameterized by 0.2 except the last layer. During training, we use SGD (Bottou 2010) as the optimizer for segmentation network with a momentum of 0.9 and initial learning rate of 2.5×10^{-4}, while utilizing Adam to optimize D with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and initializing the learning rate to 1×10^{-4}. We set optimizers a weight decay of 5×10^{-4} with a poly learning rate decay policy. We train the network for 100,000 iterations. We resize Cityscapes, GTAV, SYNTHIA to $1,024 \times 512$, $1,280 \times 720$, and $1,280 \times 760$ respectively. We also set K, λ_{adv}, λ_{cos} and ϵ to 2, 0.001, 40, and 0.4. For stage 3, the same learning rate and optimizer with stage 2 are used to finetune our network.

Comparison with State-of-the-Art Methods

In this section, we compare our method with several state-of-the-art methods with the same backbone network, the results of the comparison methods we cited here are reported in their original papers. The results of our proposed model

SYNTHIA → Cityscapes																	
Method	road	sidewalk	building	wall*	fence*	pole*	light	sign	vegetation	sky	person	rider	car	bus	motorbike	bike	mIoU	mIoU*
Source Only	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	38.6
CrCDA	86.2	44.9	79.5	8.3	0.7	27.8	9.4	11.8	78.6	**86.5**	57.2	26.1	76.8	39.9	21.5	32.1	42.9	50.0
LTIR	**92.6**	**53.2**	79.2	-	-	-	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	**39.4**	-	49.3
UIDA	84.3	37.7	79.5	5.3	0.4	24.9	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	41.7	48.9
MaxSquare	82.9	40.7	80.3	**10.2**	0.8	25.8	12.8	18.2	**82.5**	82.2	53.1	18.0	79.0	31.4	10.4	35.6	41.4	48.2
AdaptPatch	82.4	38.0	78.6	8.7	0.6	26.0	3.9	11.1	75.5	84.6	53.5	21.6	71.4	32.6	19.3	31.7	40.0	46.5
LSE	82.9	43.1	78.1	9.3	0.6	28.2	9.1	14.4	77.0	83.5	**58.1**	25.9	71.9	38.0	**29.4**	31.2	42.6	49.4
Ours	85.8	40.4	**80.4**	4.7	**1.8**	**30.8**	**16.4**	**18.6**	80.7	80.4	55.2	**26.3**	**83.9**	**43.8**	18.6	34.3	**43.9**	**51.1**

Table 2: Results of domain adaptation task SYNTHIA → Cityscapes. mIoU is calculated over 16 classes, and mIoU* denotes the mean IoU of 13 classes, excluding the classes with *.

we presented in this paper were the best in many times experiments.

GTAV to Cityscapes. Table 1 shows the results of the domain adaptation task GTAV → Cityscapes over 19 classes. Our method outperforms the source-only segmentation method by +12.5% in mIoU. Moreover, our method can achieve significantly better results than that of the compared state-of-the-art works UIDA (Pan et al. 2020), CLAN (Luo et al. 2019), ADVENT (Vu et al. 2019), DISE (Chang et al. 2019), CRST (Zou et al. 2019), and WeekSegDA (Paul et al. 2020). In particular, our method is good at predicting background objects that occupy a large area, such as “road”, “building”, “terrain”, and “sky”.

SYNTHIA to Cityscapes. We also conduct extensive experiments on the challenging SYNTHIA dataset. Table 2 presents experimental results, and our method obtains approximately 10.4% higher in 16 classes and 12.5% higher in 13 classes than non-adapted baseline in terms of mIoU. Similarly, CIRN achieves the best performance when compared with CrCDA (Huang et al. 2020), LTIR (Kim and Byun 2020), UIDA (Pan et al. 2020), MaxSquare (Chen, Xue, and Cai 2019), AdaptPatch (Tsai et al. 2019), and LSE (Subhani and Ali 2020).

K	0	1	2	3	4
GTAV	44.5	45.0	**46.1**	45.1	44.6
SYNTHIA	39.9	41.1	41.5	**41.7**	40.8

Table 3: Comparisons about different K in terms of mIoU on stage 2 for domain adaptation tasks.

Parameter Analysis and Ablation Study

In this section, we aim to study the proper values for hyperparameter K, λ_{cos} and ϵ. Following the work (Luo et al. 2019), for λ_{cos}, and ϵ, we utilize the loss of D to indicate convergence performance. That is to say, if the loss of D converges approximately 0.5, D has the same probability to distinguish two domains, then a stable adversarial training is achieved. First, we set K to 3 and test our model using λ_{cos} = 10, with varying ϵ over a range $\{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$. In this experiment, D suffers from poor convergence when we set ϵ to relatively small or big values, for example, 0.1, 0.2, 0.8, or 1. As a result, we then test our model using ϵ = 0.4 with varying λ_{cos} over a range $\{1, 10, 20, 40, 80\}$. The experimental results show that the convergence performance of D is not very sensitive to λ_{cos} because the loss converges to proper values despite the value of λ_{cos}. Thus, we choose λ_{cos} = 40 and ϵ = 0.4, for this combination achieves the best performance in terms of mIoU. Then, using λ_{cos} = 40 and ϵ = 0.4, we conduct a study to choose the proper value for hyperparameter K. As shown in Table 3, under different settings, when $K = 2$, the adaptation result is the best for GTAV, and when $K = 3$, the adaptation result is the best for SYNTHIA. Considering that the performance on SYNTHINA is slightly worse when $K = 2$ than $K = 3$, we finally set $K = 2$. Notebly, when $K = 0$, ACM will degenerate to a normal classifier. Moreover, this finding can prove that the introducing of ACM boosts the performance of domain adaptation.

	SO	CIR	ASS	ST	mIoU
GTAV	✓				36.6
		✓			42.7
			✓		43.6
		✓	✓		46.1
			✓	✓	44.2
		✓	✓	✓	**49.1**
SYNTHIA	✓				33.5
		✓			39.7
			✓		39.5
		✓	✓		41.5
			✓	✓	40.8
		✓	✓	✓	**43.9**

Table 4: Ablation study of our proposed framework in terms of mIoU.

We report the segmentation results of every stage in Table 4 to show the effectiveness of our entire framework. SO means results trained with source only, ASS means our stage 2 adaptive semantic segmentation network, ST means stage 3 self-training. The domain gap is bridged progressively in our whole adaptation process, and each stage improves the overall performance. Also, results without CIR are much worse than results with CIR introducing, which can support

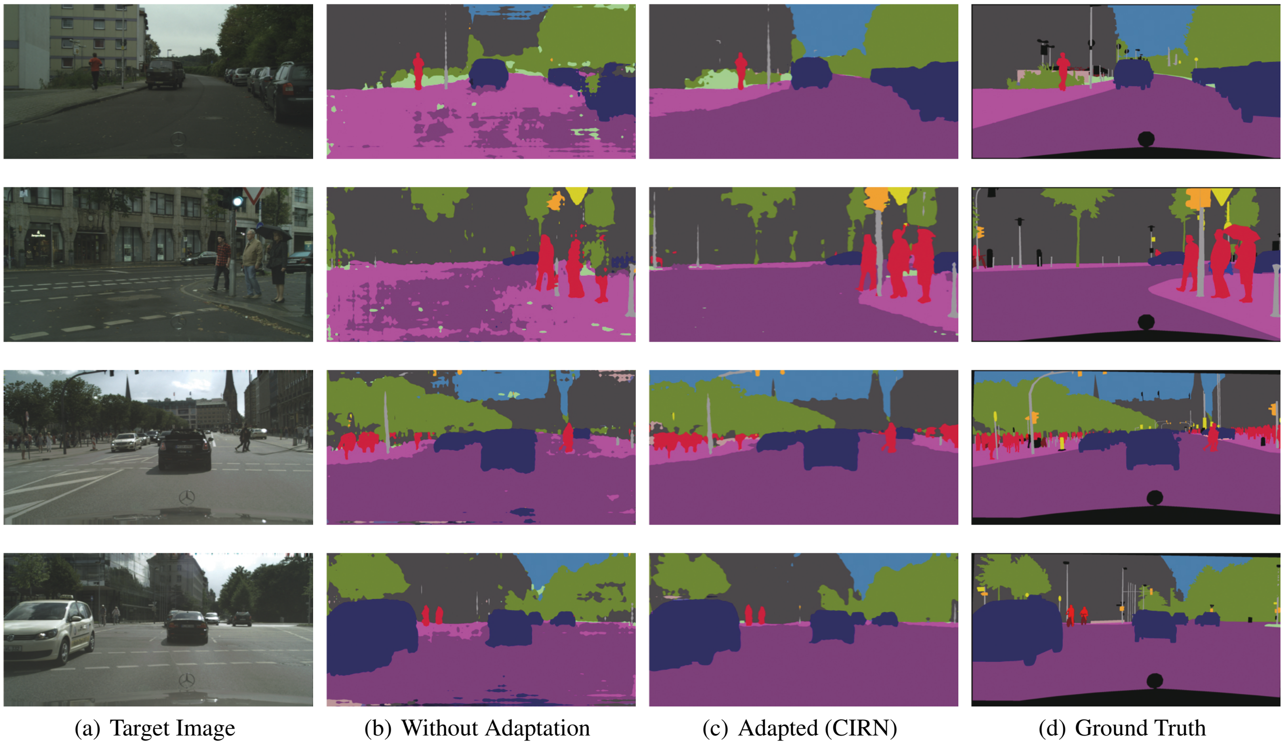

Figure 4: Qualitative semantic segmentation results on GTAV → Cityscapes.

our claim that CIR serves as a good intermediate domain.

In Figure 4, we show qualitative results after CIRN while comparing our method against ”Without Adaptation” and Ground Truth. The segmentation predictions made by our method are very similar to the ground truth labels.

Conclusion

In this study, we propose a novel framework, called CIRN, to bridge the domain gap among different domains. CIRN constructs CIR by introducing an intermediate domain to make the entire task easier. Instead of directly applying domain adaptation from source to target, we reduce the domain gap from source domain to target domain bridged by CIR. CIR shares the same content and label-distribution with the source domain and has similar data distribution to the target domain. Hence, CIR can serve as a good transitional domain to connect two distant domains. Through CIR, CIRN further focuses on pixel-level divergences to boost the performance of domain adaptation. From global to local alignment and image to pixel level, CIRN achieves good performances in reducing cross-domain discrepancy. Extensive experimental results on two challenging datasets validate the superiority of CIRN over several state-of-the-art methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62076188 and 61771349, and in part by the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170. The numerical calculations in this paper had been supported by the supercomputing system in the Supercomputing Center of Wuhan University.

References

Bottou, L. 2010. Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010*, 177–186.

Chang, W.-L.; Wang, H.-P.; Peng, W.-H.; and Chiu, W.-C. 2019. All About Structure: Adapting Structural Information Across Domains for Boosting Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1900–1909.

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. 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.

Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; and Yuille, A. L. 2016. Attention to Scale: Scale-Aware Semantic Image Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3640–3649.

Chen, M.; Xue, H.; and Cai, D. 2019. Domain Adaptation for Semantic Segmentation with Maximum Squares Loss. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2090–2099.

Chen, Y.-C.; Lin, Y.-Y.; Yang, M.-H.; and Huang, J.-B. 2019. CrDoCo: Pixel-Level Domain Transfer With Cross-Domain Consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1791–1800.

Chen, Y.-H.; Chen, W.-Y.; Chen, Y.-T.; Tsai, B.-C.; Frank Wang, Y.-C.; and Sun, M. 2017. No More Discrimination: Cross City Adaptation of Road Scene Segmenters. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1992–2001.

Choi, J.; Kim, T.; and Kim, C. 2019. Self-Ensembling With GAN-Based Data Augmentation for Domain Adaptation in Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 6830–6840.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Kai Li; and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Dong, J.; Cong, Y.; Sun, G.; Zhong, B.; and Xu, X. 2020. What Can Be Transferred: Unsupervised Domain Adaptation for Endoscopic Lesions Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4023–4032.

Gong, R.; Li, W.; Chen, Y.; and Van Gool, L. 2019. DLOW: Domain Flow for Adaptation and Generalization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2472–2481.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J. Y.; Isola, P.; Saenko, K.; Efros, A. A.; and Darrell, T. 2018. CyCADA: Cycle-Consistent Adversarial Domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 3162–3174.

Hoffman, J.; Wang, D.; Yu, F.; and Darrell, T. 2016. FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation. *arXiv preprint arXiv:1612.02649* .

Hsu, H. K.; Yao, C. H.; Tsai, Y. H.; Hung, W. C.; Tseng, H. Y.; Singh, M.; and Yang, M. H. 2020. Progressive Domain Adaptation for Object Detection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 738–746.

Huang, J.; Lu, S.; Guan, D.; and Zhang, X. 2020. Contextual-Relation Consistent Domain Adaptation for Semantic Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Kim, M.; and Byun, H. 2020. Learning Texture Invariant Representation for Domain Adaptation of Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12975–12984.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* .

Lee, C.-Y.; Batra, T.; Baig, M. H.; and Ulbricht, D. 2019. Sliced Wasserstein Discrepancy for Unsupervised Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10285–10295.

Li, Y.; Yuan, L.; and Vasconcelos, N. 2019. Bidirectional Learning for Domain Adaptation of Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6929–6938.

Lian, Q.; Lv, F.; Duan, L.; and Gong, B. 2019. Constructing Self-motivated Pyramid Curriculums for Cross-Domain Semantic Segmentation: A Non-Adversarial Approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 6758–6767.

Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; and Yang, Y. 2019. Taking A Closer Look at Domain Shift: Category-level Adversaries for Semantics Consistent Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2502–2511.

Luo, Y.; Zheng, Z.; Zheng, L.; Guan, T.; Yu, J.; and Yang, Y. 2018. Macro-Micro Adversarial Network for Human Parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 424–440.

Murez, Z.; Kolouri, S.; Kriegman, D.; Ramamoorthi, R.; and Kim, K. 2018. Image to Image Translation for Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4500–4509.

Pan, F.; Shin, I.; Rameau, F.; Lee, S.; and Kweon, I. S. 2020. Unsupervised Intra-domain Adaptation for Semantic Segmentation through Self-Supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3764–3773.

Paul, S.; Tsai, Y.-H.; Schulter, S.; Roy-Chowdhury, A. K.; and Chandraker, M. 2020. Domain Adaptive Semantic Segmentation Using Weak Labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for Data: Ground Truth from Computer Games. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 102–118.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3234–3243.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting Visual Category Models to New Domains. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 213–226.

Sankaranarayanan, S.; Balaji, Y.; and Jain. 2018. Learning from Synthetic Data: Addressing Domain Shift for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3752–3761.

Shimodaira, H. 2000. Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function. *Journal of Statistical Planning and Inference* 90(2): 227–244.

Song, L.; Wang, C.; Zhang, L.; Du, B.; Zhang, Q.; Huang, C.; and Wang, X. 2020. Unsupervised Domain Adaptive Re-Identification: Theory and Practice. *Pattern Recognition* 102: 1–11.

Song, L.; Xu, Y.; Zhang, L.; Du, B.; Zhang, Q.; and Wang, X. 2020. Learning From Synthetic Images via Active Pseudo-Labeling. *IEEE Transactions on Image Processing* 29: 6452–6465.

Subhani, M. N.; and Ali, M. 2020. Learning from Scale-Invariant Examples for Domain Adaptation in Semantic Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Tsai, Y.-H.; Sohn, K.; Schulter, S.; and Chandraker, M. 2019. Domain Adaptation for Structured Output via Discriminative Patch Representations. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1456–1465.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2517–2526.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual Attention Network for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6450–6458.

Wu, Z.; Han, X.; Lin, Y. L.; Uzunbas, M. G.; Goldstein, T.; Lim, S. N.; and Davis, L. S. 2018. DCAN: Dual Channel-Wise Alignment Networks for Unsupervised Scene Adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 535–552.

Xu, Y.; Du, B.; Zhang, L.; Zhang, Q.; Wang, G.; and Zhang, L. 2019. Self-Ensembling Attention Networks: Addressing Domain Shift for Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 5581–5588.

Yang, Y.; Lao, D.; Sundaramoorthi, G.; and Soatto, S. 2020. Phase Consistent Ecological Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9011–9020.

Yang, Y.; and Soatto, S. 2020. FDA: Fourier Domain Adaptation for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4085–4095.

Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A.; Keutzer, K.; and Gong, B. 2019. Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization Without Accessing Target Domain Data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2100–2110.

Zhang, Q.; Zhang, J.; Liu, W.; and Tao, D. 2019. Category Anchor-Guided Unsupervised Domain Adaptation for Semantic Segmentation. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 435–445.

Zhang, Y.; David, P.; Foroosh, H.; and Gong, B. 2020. A Curriculum Domain Adaptation Approach to the Semantic Segmentation of Urban Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(8): 1823–1841.

Zhao, S.; Fu, H.; Gong, M.; and Tao, D. 2019. Geometry-Aware Symmetric Domain Adaptation for Monocular Depth Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9788–9798.

Zhu, J. Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2242–2251.

Zou, Y.; Yu, Z.; Kumar, B. V.; and Wang, J. 2018. Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 297–313.

Zou, Y.; Yu, Z.; Liu, X.; Kumar, B. V.; and Wang, J. 2019. Confidence Regularized Self-Training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5982–5991.