Dual Path Learning for Domain Adaptation of Semantic Segmentation

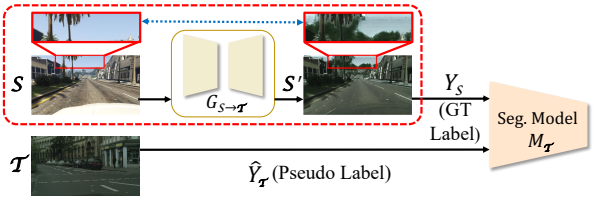
Yiting Cheng[1] Fangyun Wei*[2] Jianmin Bao[2] Dong Chen[2] Fang Wen[2] Wenqiang Zhang*[1]
[1]School of Computer Science, Fudan University [2]Microsoft Research Asia
{ytcheng18, wqzhang}@fudan.edu.cn {fawe, jianbao, doch, fangwen}@microsoft.com

Abstract

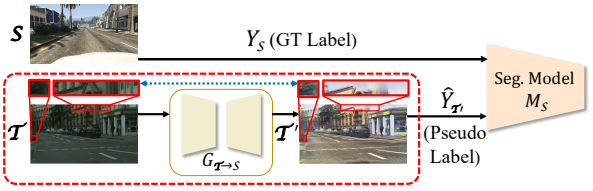
Domain adaptation for semantic segmentation enables to alleviate the need for large-scale pixel-wise annotations. Recently, self-supervised learning (SSL) with a combination of image-to-image translation shows great effectiveness in adaptive segmentation. The most common practice is to perform SSL along with image translation to well align a single domain (the source or target). However, in this single-domain paradigm, unavoidable visual inconsistency raised by image translation may affect subsequent learning. In this paper, based on the observation that domain adaptation frameworks performed in the source and target domain are almost complementary in terms of image translation and SSL, we propose a novel dual path learning (DPL) framework to alleviate visual inconsistency. Concretely, DPL contains two complementary and interactive single-domain adaptation pipelines aligned in source and target domain respectively. The inference of DPL is extremely simple, only one segmentation model in the target domain is employed. Novel technologies such as dual path image translation and dual path adaptive segmentation are proposed to make two paths promote each other in an interactive manner. Experiments on GTA5→Cityscapes and SYNTHIA→Cityscapes scenarios demonstrate the superiority of our DPL model over the state-of-the-art methods. The code and models are available at: https://github.com/royee182/DPL.

1. Introduction

In the past decades, significant progress [45, 4, 38, 35, 43, 31, 20] in semantic segmentation has been achieved with Deep Convolutional Neural Networks. The empirical observation [25, 39] demonstrates that the leading performance is partially attributed to a large volume of training data, thus dense pixel-level annotations are required in supervised learning, which is laborious and time-consuming. To avoid this painstaking task, researchers resort to train segmentation models on synthetic but photo-realistic large-scale datasets such as GTA5 [26] and SYNTHIA [27] with computer-generated annotations. However, due to the cross-domain differences, these well-trained models usually undergo significant performance drops when tested on realistic datasets (e.g., Cityscapes [6]). Therefore, unsupervised domain adaptation (UDA) methods have been widely adopted to align the domain shift between the rich-labeled source data (synthetic images) and the unlabeled target data (real images).



(a) Illustration of adaptation in domain-$\mathcal{T}$.



(b) Illustration of adaptation in domain-$\mathcal{S}$.

Figure 1: Illustration of single-domain adaptation pipelines. $\mathcal{S}$ is source image with ground-truth label $Y_{\mathcal{S}}$, and $\mathcal{T}$ is target image. $G_{\mathcal{S}\to\mathcal{T}}$ represents image translation from domain-$\mathcal{S}$ to domain-$\mathcal{T}$ and vice versa. $\mathcal{S}' = G_{\mathcal{S}\to\mathcal{T}}(\mathcal{S})$ and $\mathcal{T}' = G_{\mathcal{T}\to\mathcal{S}}(\mathcal{T})$ are translated images in the corresponding domain. $M_{\mathcal{S}}$ and $M_{\mathcal{T}}$ are semantic segmentation models in domain-$\mathcal{S}$ and domain-$\mathcal{T}$, respectively. $\hat{Y}_{\mathcal{T}}$ and $\hat{Y}_{\mathcal{T}'}$ represent the corresponding pseudo labels of $\mathcal{T}$ and $\mathcal{T}'$. Red dash rectangles denote that visual inconsistency raised by image translations disturbs domain adaptation learning in either supervised part or SSL part.

Two commonly used paradigms in unsupervised domain adaptive segmentation are image-to-image translation based methods [21, 9] and self-supervised learning (SSL) based methods [50, 49, 44, 12]. The most common practice for image-to-image translation based methods is to translate synthetic data from source domain (denote as domain-$\mathcal{S}$) to target domain (denote as domain-$\mathcal{T}$) [9, 2] to reduce the visual gap between different domains. Then adaptive seg-

*Corresponding author

mentation is trained on translated synthetic data. However, by only applying the image-to-image translation to domain adaptation task, the results are always unsatisfying. One of the leading factors is that image-to-image translation may change the image content involuntarily and introduce *visual inconsistency* between raw images and translated images. Training on translated images with uncorrected ground-truth labels of source images introduces noise which disturbs the domain adaptation learning.

A combination of SSL and image-to-image translation [16, 41, 15] has been demonstrated great effectiveness in the UDA field. SSL utilizes a well-trained segmentation model to generate a set of pseudo labels with high confidence for unlabeled target data, then the adaptive segmentation training can be divided into two parallel parts, namely supervised part (training is performed on source data with ground-truth labels) and SSL part (training is performed on target data with pseudo labels). In this paradigm, the most prevalent practice is to perform adaptation to well align a single domain, i.e., either source domain (named domain-$\mathcal{S}$ adaptation) [16, 15] or target domain (named domain-$\mathcal{T}$ adaptation) [41]. However, both domain-$\mathcal{S}$ and domain-$\mathcal{T}$ adaptation heavily rely on the quality of image-to-image translation models, where visual inconsistency is always unavoidable. For domain-$\mathcal{T}$ adaptation (as shown in Figure 1.(a)), visual inconsistency brings in misalignment between translated source images and uncorrected ground-truth labels, which disturbs the supervised part. In contrast, domain-$\mathcal{S}$ adaptation (as shown in Figure 1.(b)) avoids image translation on source images, but simultaneously introduces visual inconsistency between target images and the corresponding translated images. Defective pseudo labels generated by unaligned images disturb the SSL part.

Notice the above single-domain adaptation pipelines are almost complementary in terms of the two training parts, i.e., visual inconsistency caused by image translation disturbs the training of supervised part in domain-$\mathcal{T}$ adaptation and SSL part in domain-$\mathcal{S}$ adaptation. In contrast, SSL part in domain-$\mathcal{T}$ adaptation and supervised part in domain-$\mathcal{S}$ adaptation are unaffected. It is natural to raise a question: *could we combine these two complementary adaptation pipelines into a single framework to make good use of each strength and make them promote each other?* Based on this idea, we propose the *dual path learning* framework which considers two pipelines from opposite domains to alleviate unavoidable visual inconsistency raised by image translations. We name two paths used in our framework as path-$\mathcal{T}$ (adaption is performed in domain-$\mathcal{T}$) and path-$\mathcal{S}$ (adaption is performed in domain-$\mathcal{S}$), respectively. Path-$\mathcal{S}$ assists path-$\mathcal{T}$ to learn precise supervision from source data. Meanwhile, path-$\mathcal{T}$ guides path-$\mathcal{S}$ to generate high-quality pseudo labels which are important for SSL in return. It is worth noting that path-$\mathcal{S}$ and path-$\mathcal{T}$ are not two separated pipelines in our framework, interactions between two paths are performed throughout the training, which is demonstrated to be effective in our experiments. The whole system forms a closed-loop learning. Once the training has finished, we only retain a single segmentation model well aligned in target domain for testing, no extra computation is required. The main contributions of this work are summarized as:

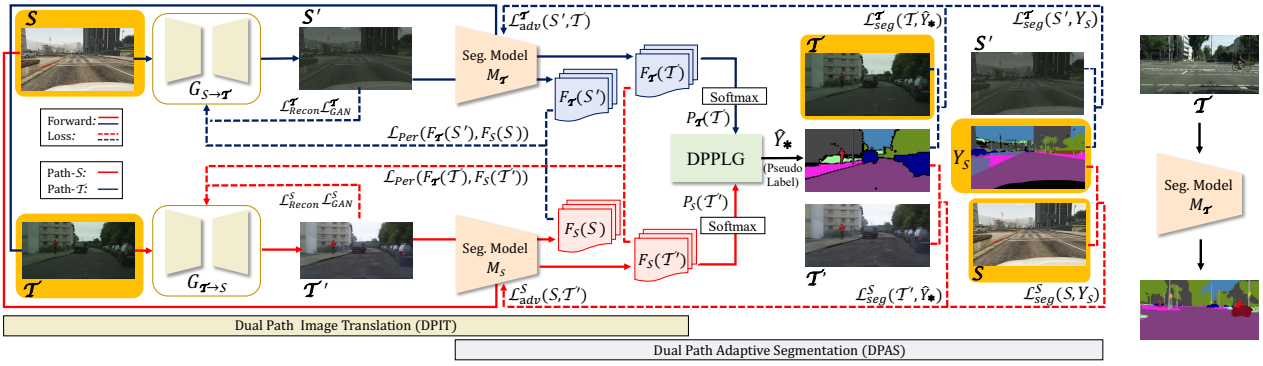
- We present a novel dual path learning (DPL) framework for domain adaptation of semantic segmentation. DPL employs two complementary and interactive single-domain pipelines (namely path-$\mathcal{T}$ and path-$\mathcal{S}$) in the training phase. In the testing, only a single segmentation model well aligned in target domain is used. The proposed DPL framework surpasses state-of-the-art methods on representative scenarios.
- We present two interactive modules to make two paths promote each other, namely dual path image translation and dual path adaptive segmentation.
- We introduce a novel warm-up strategy for the segmentation models which helps adaptive segmentation in the early training stage.

2. Related Work

Domain Adaptation. Domain adaptation is a broadly studied topic in computer vision. It aims to rectify the mismatch in cross-domains and tune the models toward better generalization at testing [23]. A variety of domain adaptation methods for image classification [28, 3, 33, 13] and object detection [5, 1] have been proposed. In this paper, we focus on the unsupervised domain adaptation of semantic segmentation.

Domain Adaptation for Semantic Segmentation. Semantic segmentation needs a large volume of pixel-level labeled training data, which is laborious and time-consuming in annotation. A promising solution to reduce the labeling cost is to train segmentation networks on synthetic dataset (e.g., GTA5 [26] and SYNTHIA [27]) with computer-generated annotations before testing on realistic dataset (e.g., Cityscapes [6]). Although synthetic images have similar appearance to real images, there still exist domain discrepancies in terms of layouts, colors and illumination conditions, which always cripples the models' performance. Domain adaptation is necessary to align the synthetic and the real dataset [37, 50, 46, 14].

Adversarial-based methods [10, 19, 32] are broadly explored in unsupervised domain adaptation, which align different domains at image-level [21, 9, 37] or feature-level [32, 11]. The image-level adaptation regards domain adaptation as an image synthesis problem, and aims to reduce visual discrepancy (e.g., lighting and object texture)



(a) Training pipeline of DPL. (b) Testing.

Figure 2: (a) Overview of DPL framework. Inputs are highlighted by orange rectangles. DPL consists of two complementary single-domain paths: path-$\mathcal{S}$ (learning is performed in *source* domain) and path-$\mathcal{T}$ (learning is performed in *target* domain). Dual path image translation (DPIT) and dual path adaptive segmentation (DPAS) are proposed to make two paths interactive and promote each other. In DPIT, unpaired image translation models ($G_{\mathcal{T}\to\mathcal{S}}$ and $G_{\mathcal{S}\to\mathcal{T}}$) are supervised by general GAN loss and cross-domain perceptual loss. DPAS employs the proposed dual path pseudo label generation (DPPLG) module to produce pseudo labels $\hat{Y}_*$ of target images, then segmentation models ($M_{\mathcal{S}}$ and $M_{\mathcal{T}}$) are trained on both source images (or translated source images) with ground-truth labels and target images (or translated target images) with pseudo labels. (b) Testing of DPL. Only $M_{\mathcal{T}}$ is used for inference.

in cross-domains with unpaired image-to-image translation models [47, 17, 22]. However, the performance is always unsatisfactory by simply applying image translation to domain adaptation task. One reason is that image-to-image translation may change the image content involuntarily and further disturb the following segmentation training [16].

In recent years, self-supervised learning (SSL) [7, 48] shows tremendous potential in adaptive segmentation [50, 49, 30, 12]. The key principle for these methods is to generate a set of pseudo labels for target images as the approximation to the ground-truth labels, then segmentation model is updated by leveraging target domain data with pseudo labels. CRST [50] is the first work to introduce self-training into adaptive segmentation, it also alleviates category imbalance issue by controlling the proportion of selected pseudo labels in each category. Recent TPLD [12] proposes a two-phase pseudo label densification strategy to obtain dense pseudo labels for SSL.

Two works [16, 41] which explore the combination of image translation and SSL are closely related to ours. Label-Driven [41] performs a target-to-source translation and a label-driven reconstruction module is used to reconstruct source and target images from the corresponding predicted labels. In contrast, BDL [16] represents a bidirectional learning framework which alternately trains the image translation and the adaptive segmentation in target domain. Meanwhile, BDL utilizes a single-domain perceptual loss to maintain visual consistency. We will demonstrate this kind of design is suboptimal compared with the proposed dual path image translation module in Section 3.2. These two works demonstrate the combination of image translation and SSL can promote adaptive learning. Different from these single-domain adaptation methods, the proposed dual path learning framework integrates two complementary single-domain pipelines in an interactive manner to address visual inconsistency problem by: 1) utilizing segmentation models aligned in different domains to provide cross-domain perceptual supervision for image translation; 2) combining knowledge from both source and target domain for self-supervised learning.

3. Method

Given the source dataset $\mathcal{S}$ (synthetic data) with pixel-level segmentation labels $Y_{\mathcal{S}}$, and the target dataset $\mathcal{T}$ (real data) with no labels. The goal of unsupervised domain adaptation (UDA) is that by only using $\mathcal{S}$, $Y_{\mathcal{S}}$ and $\mathcal{T}$, the segmentation performance can be on par with the model trained on $\mathcal{T}$ with corresponding ground-truth labels $Y_{\mathcal{T}}$. Domain gap between $\mathcal{S}$ and $\mathcal{T}$ makes the task difficult for the network to learn transferable knowledge at once.

To address this problem, we propose a novel dual path learning framework named DPL. As shown in Figure 2.(a), DPL consists of two complementary and interactive paths: path-$\mathcal{S}$ (adaptive learning is performed in *source* domain) and path-$\mathcal{T}$ (adaptive learning is performed in *target* domain). How to allow one of both paths provide positive feedbacks to the other is the key to success. To achieve this goal, we propose two modules, namely dual path image translation (DPIT) and dual path adaptive segmentation (DPAS). DPIT aims to reduce the visual gap between different domains without introducing visual inconsistency. In our design, DPIT unites general unpaired image translation models with dual perceptual supervision from two single-domain segmentation models. Notice any unpaired image translation models can be used in DPIT, we use CycleGAN [47] as our default model due to its popularity and it provides bidirectional image translation inherently. We use $\mathcal{T}' = G_{\mathcal{T}\to\mathcal{S}}(\mathcal{T})$ and $\mathcal{S}' = G_{\mathcal{S}\to\mathcal{T}}(\mathcal{S})$ to denote translated

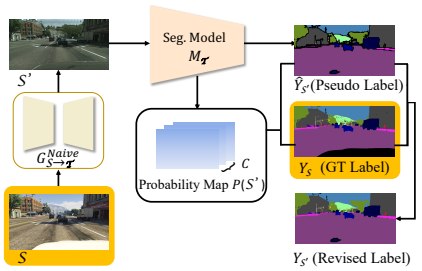


Figure 3: Illustration of label correction strategy. Inputs are highlighted by orange rectangles.

images in path-$\mathcal{S}$ and path-$\mathcal{T}$ respectively, where $G_{\mathcal{T}\to\mathcal{S}}$ and $G_{\mathcal{S}\to\mathcal{T}}$ are image translation models in the corresponding path. DPAS utilizes translated images from DPIT and the proposed dual path pseudo label generation (DPPLG) module to generate high-quality pseudo labels for target images, then segmentation models $M_{\mathcal{S}}$ (in path-$\mathcal{S}$) and $M_{\mathcal{T}}$ (in path-$\mathcal{T}$) are trained with both transferred knowledge in source domain and implicit supervision in target domain. The testing of DPL is extremely simple, we only retain $M_{\mathcal{T}}$ for inference as shown in Figure 2.(b).

The training process of DPL consists of two phases: single-path warm-up and DPL training. DPL benefits from well-initialized $M_{\mathcal{S}}$ and $M_{\mathcal{T}}$, since both DPIT and DPAS rely on the quality of segmentation models. A simple but efficient warm-up strategy can accelerate the convergence of DPL. Once the warm-up finishes, DPIT and DPAS are trained sequentially in DPL training phase.

In this section, we first describe our warm-up strategy in Section 3.1. Then, we introduce the key components of DPL: DPIT in Section 3.2 and DPAS in Section 3.3. Next, we revisit and summarize the whole training process in Section 3.4. Finally, testing pipeline of DPL is presented in Section 3.5.

3.1. Single Path Warm-up

Perceptual supervision in DPIT and pseudo label generation in DPAS rely on the quality of segmentation models. To accelerate convergence of DPL, a warm-up process for segmentation models $M_{\mathcal{S}}$ and $M_{\mathcal{T}}$ is required.

$M_{\mathcal{S}}$ Warm-up. The warm-up for $M_{\mathcal{S}}$ is easily conducted in a fully supervised way by using source dataset $\mathcal{S}$ with ground-truth labels $Y_{\mathcal{S}}$.

$M_{\mathcal{T}}$ Warm-up. It is difficult to directly train $M_{\mathcal{T}}$ in a supervised manner since no labels can be accessed in target dataset $\mathcal{T}$. A straightforward idea is to translate source images $\mathcal{S}$ to target domain by using naive CycleGAN, and then $M_{\mathcal{T}}$ is trained on translated images $\mathcal{S}'$ with approximate ground-truth labels $Y_{\mathcal{S}}$. Unfortunately, naive CycleGAN does not apply any constrains to preserve visual consistency between $\mathcal{S}$ and $\mathcal{S}'$, i.e., visual content may be changed when $\mathcal{S}$ is translated to $\mathcal{S}'$. Misalignment between $\mathcal{S}'$ and $Y_{\mathcal{S}}$ can disturb the training of $M_{\mathcal{T}}$.

To address this issue, we propose a novel label correction strategy as shown in Figure 3. The core principle is to find a revised label $Y_{\mathcal{S}'}$ for $\mathcal{S}'$ by considering both ground-truth labels $Y_{\mathcal{S}}$ and segmentation predictions of $\mathcal{S}'$. Specially, we feed $\mathcal{S}'$ into $M_{\mathcal{T}}$ (which is initialized as $M_{\mathcal{S}}$ at the beginning) to generate pseudo labels $\hat{Y}_{\mathcal{S}'}$. Then label correction module revises raw ground-truth labels $Y_{\mathcal{S}}$ by replacing pixel-wise labels in $Y_{\mathcal{S}}$ with high-confidence pixel-wise labels in $\hat{Y}_{\mathcal{S}'}$, which means the labels of content-changed areas have been approximately corrected by reliable predictions. Formally, define revised labels $Y_{\mathcal{S}'} = \{Y^{(i,j)}_{\mathcal{S}'}\}$ $(1 \le i \le H, 1 \le j \le W)$ as:

$$Y^{(i,j)}_{\mathcal{S}'} = \begin{cases} \hat{Y}^{(i,j)}_{\mathcal{S}'}, & \text{if } P^{(i,j,\hat{c})}(\mathcal{S}') - P^{(i,j,c)}(\mathcal{S}') > \delta \\ Y^{(i,j)}_{\mathcal{S}}, & \text{else,} \end{cases} \tag{1}$$

where H and W denote the height and width of the input image respectively, $P(\cdot)$ is probability map predicted by segmentation model, $\hat{c}$ and c denote the category index of $\hat{Y}^{(i,j)}_{\mathcal{S}'}$ and $Y^{(i,j)}_{\mathcal{S}}$ respectively, δ controls correction rate, we set $\delta = 0.3$ empirically.

In addition, we also use $M_{\mathcal{T}}$ to generate pseudo labels $\hat{Y}_{\mathcal{T}}$ for $\mathcal{T}$. Now we have paired training data $(\mathcal{S}', Y_{\mathcal{S}'})$ and $(\mathcal{T}, \hat{Y}_{\mathcal{T}})$ which approximately lie in target domain for $M_{\mathcal{T}}$ training. The overall loss is defined as:

$$\begin{aligned} \mathcal{L}_{M_{\mathcal{T}}} = \mathcal{L}_{seg}(\mathcal{S}', Y_{\mathcal{S}'}) + \mathcal{L}_{seg}(\mathcal{T}, \hat{Y}_{\mathcal{T}}) \\ + \lambda_{adv}\mathcal{L}_{adv}(\mathcal{S}', \mathcal{T}), \end{aligned} \tag{2}$$

where $\mathcal{L}_{adv}$ represents typical adversarial loss as used in [32, 16, 41] to further align target domain, $\mathcal{L}_{seg}$ indicates the commonly used per-pixel segmentation loss:

$$\mathcal{L}_{seg}(I, Y) = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} Y^{(i,j,c)} \log P^{(i,j,c)}(I), \tag{3}$$

where I and Y denote input image (raw image or translated image) and corresponding labels (ground-truth labels or pseudo labels), respectively.

Once warm-up procedure is finished, we obtain preliminary segmentation models which are approximately aligned in the corresponding domain. These well-initialized models facilitate the training of DPIT and DPAS, which will be described in next sections.

3.2. Dual Path Image Translation

Image-to-image translation aims to reduce the gap in visual appearance (e.g., object textures and lighting) between source and target domain. As discussed in Section 1, unavoidable visual inconsistency caused by image translation may mislead the subsequent adaptive segmentation learning, and thus extra constraints to maintain visual consistency are required.

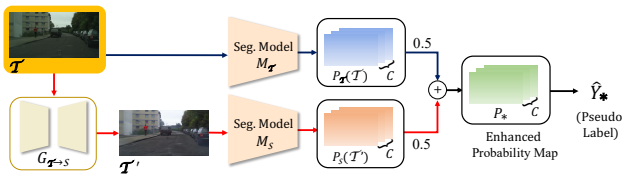


Figure 4: Illustration of dual path pseudo label generation (DPPLG). Input is highlighted by orange rectangle.

BDL [16] introduces a perceptual loss to maintain visual consistency between paired images (i.e., raw images and corresponding translated images). Perceptual loss measures distance of perceptual features[1] extracted from a well-trained segmentation model. In BDL, domain adaptation is only performed in target domain, as a result, the perceptual loss of paired images ($\mathcal{S}$, $\mathcal{S}'$) and ($\mathcal{T}$, $\mathcal{T}'$) is computed with identical segmentation model. Notice paired images are from two different domains ($\mathcal{S}$ and $\mathcal{T}'$ are in source domain while $\mathcal{T}$ and $\mathcal{S}'$ are in target domain), using segmentation model aligned in a single domain to extract features for perceptual loss computation may be suboptimal.

Now we introduce our dual path image translation (DPIT) as illustrated in Figure 2.(a). DPIT is an bidirectional image translation model with cross-domain perceptual supervision. We use $G_{\mathcal{S}\to\mathcal{T}}$ and $G_{\mathcal{T}\to\mathcal{S}}$ to denote image translation in Path-$\mathcal{T}$ and Path-$\mathcal{S}$ respectively. CycleGAN is served as our default model since it provides bidirectional image translation inherently, however, any unpaired image translation algorithms can be used in DPIT. Different from BDL, DPIT makes use of two paths aligned in opposite domains and extracts perceptual features for paired images from their corresponding path to better maintain visual consistency. Concretely, DPIT utilizes $M_{\mathcal{S}}$ to extract perceptual features for $\mathcal{S}$ and $\mathcal{T}'$, and $M_{\mathcal{T}}$ to extract perceptual features for $\mathcal{T}$ and $\mathcal{S}'$, respectively. Then we can formulate our dual perceptual loss $\mathcal{L}_{DualPer}$ as:

$$\begin{aligned}\mathcal{L}_{DualPer}(\mathcal{S},\mathcal{S}',\mathcal{T},\mathcal{T}') = \mathcal{L}_{Per}(F_{\mathcal{T}}(\mathcal{S}'),F_{\mathcal{S}}(\mathcal{S})) \\ + \mathcal{L}_{Per}(F_{\mathcal{T}}(\mathcal{T}),F_{\mathcal{S}}(\mathcal{T}')),\end{aligned} \tag{4}$$

where $\mathcal{L}_{Per}$ is perceptual loss as in [16], $F_{\mathcal{S}}(\cdot)$ and $F_{\mathcal{T}}(\cdot)$ represent perceptual feature extracted by $M_{\mathcal{S}}$ and $M_{\mathcal{T}}$ respectively.

Besides the supervision of dual perceptual loss, DPIT is also supervised by general adversarial and reconstruction loss. The overall loss of DPIT can be formulated as:

$$\begin{aligned}\mathcal{L}_{DPIT} = \mathcal{L}^{\mathcal{S}}_{GAN}(\mathcal{S},\mathcal{T}') + \mathcal{L}^{\mathcal{T}}_{GAN}(\mathcal{S}',\mathcal{T}) \\ + \lambda_{Recon}\mathcal{L}^{\mathcal{S}}_{Recon}(\mathcal{S},G_{\mathcal{T}\to\mathcal{S}}(\mathcal{S}')) \\ + \lambda_{Recon}\mathcal{L}^{\mathcal{T}}_{Recon}(\mathcal{T},G_{\mathcal{S}\to\mathcal{T}}(\mathcal{T}')) \\ + \lambda_{DualPer}\mathcal{L}_{DualPer}(\mathcal{S},\mathcal{S}',\mathcal{T},\mathcal{T}'),\end{aligned} \tag{5}$$

[1]Perceptual feature denotes the probability map before softmax layer of segmentation model.

where $\mathcal{L}^{\mathcal{S}}_{GAN}$ ($\mathcal{L}^{\mathcal{T}}_{GAN}$) and $\mathcal{L}^{\mathcal{S}}_{Recon}$ ($\mathcal{L}^{\mathcal{T}}_{Recon}$) are GAN loss and reconstruction loss as in [47], λ_{Recon} and $\lambda_{DualPer}$ denote the weights of reconstruction loss and dual perceptual loss respectively. We set $\lambda_{Recon} = 10$ and $\lambda_{DualPer} = 0.1$ by default.

3.3. Dual Path Adaptive Segmentation

Once DPIT is symmetrically trained, translated images $\mathcal{S}' = G_{\mathcal{S}\to\mathcal{T}}(\mathcal{S})$ and $\mathcal{T}' = G_{\mathcal{T}\to\mathcal{S}}(\mathcal{T})$ are fed into dual path adaptive segmentation (DPAS) module for subsequent learning. As shown in figure 2.(a), DPAS utilizes self-supervised learning with combination of well-trained image translation for adaptive segmentation learning, i.e., segmentation models are trained on both source images (or translated source images) with ground-truth labels and target images (or translated target images) with pseudo labels. The core of DPAS is to generate high-quality pseudo labels of target images by combining predicted results from two paths. The training process of DPAS can be formulated as two alternative steps: 1) dual path pseudo label generation; 2) dual path segmentation training.

Dual Path Pseudo Label Generation. The labels of target dataset are unavailable in unsupervised domain adaptation tasks. Self-supervised learning (SSL) has been demonstrated great success when the labels of dataset are insufficient or noisy. The way to generate pseudo labels plays an important role in SSL. As described in Section 1, in path-$\mathcal{T}$, visual inconsistency brings in misalignment between translated source images $\mathcal{S}'$ and uncorrected ground-truth labels $Y_{\mathcal{S}}$, which disturbs the training of $M_{\mathcal{T}}$. Similar issue exists in path-$\mathcal{S}$ (see Figure 1). Inspired by the observation that two paths from opposite domains are almost complementary, we take full advantages of two paths and present a novel dual path pseudo label generation (DPPLG) strategy to generate high-quality pseudo labels as shown in Figure 4.

Concretely, let $P_{\mathcal{S}}(\cdot) = \text{Softmax}(F_{\mathcal{S}}(\cdot))$ and $P_{\mathcal{T}}(\cdot) = \text{Softmax}(F_{\mathcal{T}}(\cdot))$ denote probability map predicted by $M_{\mathcal{S}}$ and $M_{\mathcal{T}}$, respectively. In path-$\mathcal{T}$, target images can be directly fed into $M_{\mathcal{T}}$ to generate $P_{\mathcal{T}}(\mathcal{T})$. In contrast, path-$\mathcal{S}$ requires image translation to generate $\mathcal{T}' = G_{\mathcal{T}\to\mathcal{S}}(\mathcal{T})$, then $P_{\mathcal{S}}(\mathcal{T}')$ can be obtained by feeding $\mathcal{T}'$ into $M_{\mathcal{S}}$. Finally, enhanced probability map P_* which is used for generating pseudo labels of target images can be obtained by a weighted sum of two separate probability maps $P_{\mathcal{T}}(\mathcal{T})$ and $P_{\mathcal{S}}(\mathcal{T}')$:

$$P_* = \frac{1}{2}P_{\mathcal{T}}(\mathcal{T}) + \frac{1}{2}P_{\mathcal{S}}(\mathcal{T}'), \tag{6}$$

Following common practice [16, 12], we use max probability threshold (MPT) to select the pixels with higher confidence in P_* as pseudo labels of unlabeled target images. Concretely, define pseudo labels $\hat{Y}_* = \{\hat{Y}_*^{(i,j,c)}\}$ $(1 \le i \le H, 1 \le j \le W, 1 \le c \le C)$ as:

$$\hat{Y}_*^{(i,j,c)} = \begin{cases} 1, & \text{if} \quad c = argmax\,(P_*^{(i,j,c)}) \\ & \text{and } P_*^{(i,j,c)} > \lambda \\ 0, & \text{else,} \end{cases} \tag{7}$$

where λ denotes threshold to filter pixels with low prediction confidence. We set $\lambda = 0.9$ as default according to [16].

Though path-$\mathcal{S}$ and path-$\mathcal{T}$ can use respective pseudo labels generated by themselves, we will demonstrate the benefits by using shared pseudo label $\hat{Y}_*$ in Section 4.4.

Dual Path Segmentation Training. Now we introduce the process of dual path segmentation training. Concretely, for path-$\mathcal{T}$, the objective is to train a well generalized segmentation model $M_\mathcal{T}$ in target domain. Training data for $M_\mathcal{T}$ includes two part, translated source images $\mathcal{S}' = G_{\mathcal{S}\to\mathcal{T}}(\mathcal{S})$ with ground-truth labels $Y_\mathcal{S}$, and raw target images $\mathcal{T}$ with pseudo labels $\hat{Y}_*$ generated by DPPLG. In contrast, path-$\mathcal{S}$ requires good generalization in source domain. Similarly, $M_\mathcal{S}$ is trained on source images $\mathcal{S}$ with ground-truth labels $Y_\mathcal{S}$ and translated images $\mathcal{T}' = G_{\mathcal{T}\to\mathcal{S}}(\mathcal{T})$ with shared pseudo labels $\hat{Y}_*$. Besides the supervision from segmentation loss, we also utilize a discriminator on top of the features of the segmentation model to further decrease the domain gap as in [9, 16]. The overall loss function of dual path segmentation can be defined as:

$$\begin{aligned} \mathcal{L}_{DualSeg} &= \mathcal{L}_{seg}^{\mathcal{T}}(\mathcal{S}', Y_\mathcal{S}) + \mathcal{L}_{seg}^{\mathcal{T}}(\mathcal{T}, \hat{Y}_*) \\ &+ \mathcal{L}_{seg}^{\mathcal{S}}(\mathcal{S}, Y_\mathcal{S}) + \mathcal{L}_{seg}^{\mathcal{S}}(\mathcal{T}', \hat{Y}_*) \\ &+ \lambda_{adv}(\mathcal{L}_{adv}^{\mathcal{T}}(\mathcal{S}', \mathcal{T}) + \mathcal{L}_{adv}^{\mathcal{S}}(\mathcal{S}, \mathcal{T}')), \end{aligned} \tag{8}$$

where $\mathcal{L}_{adv}^{\mathcal{S}}$ and $\mathcal{L}_{adv}^{\mathcal{T}}$ denote typical adversarial loss, $\mathcal{L}_{seg}^{\mathcal{S}}$ and $\mathcal{L}_{seg}^{\mathcal{T}}$ are per-pixel segmentation loss as defined in Equation 3, λ_{adv} controls contribution of adversarial loss.

3.4. Training pipeline

Algorithm 1 summarizes the whole training process of DPL. First, $M_\mathcal{S}$ and $M_\mathcal{T}$ are initialized by the proposed warm-up strategy. Next, we train DPIT to provide well-translated images for subsequent learning. At last, following the common practice that self-supervised learning is conducted in an iterative way [16, 49, 12], DPAS is trained N times for domain adaptation. We use superscript (n) to refer to the n-th iteration.

Algorithm 1 Training process of DPL

Input: $\mathcal{S}, Y_\mathcal{S}, \mathcal{T}$
Output: $M_\mathcal{T}^{(N)}$, $M_\mathcal{S}^{(N)}$
warm-up $M_\mathcal{S}^{(0)}$, $M_\mathcal{T}^{(0)}$
train DPIT with Equation 5
for $n \leftarrow 1$ to N **do** DPAS
  generate $\hat{Y}_*^{(n)}$ with Equation 7
  train $M_\mathcal{T}^{(n)}$ and $M_\mathcal{S}^{(n)}$ with Equation 8
end for

3.5. Testing Pipeline

As shown in Figure 2.(b), the inference of DPL is extremely simple, we only retain $M_\mathcal{T}$ when testing on target images. Though DPL already shows the superiority over the state-of-the-art methods, we explore an optional dual path testing pipeline named DPL-Dual to boost performance by considering predictions from two paths. Concretely, we first generate probability map $P_\mathcal{T}(\mathcal{T})$ and $P_\mathcal{S}(\mathcal{T}')$ from two well-trained segmentation models $M_\mathcal{T}$ and $M_\mathcal{S}$ respectively, then an average function is used to generate final probability map $P_F = (P_\mathcal{S}(\mathcal{T}') + P_\mathcal{T}(\mathcal{T}))/2$. Though DPL-Dual promotes the performance, extra computation is introduced. We recommend DPL-Dual as an optional inference pipeline when computation cost is secondary.

4. Experiments

4.1. Datasets

Following common practice, We evaluate our framework in two common scenarios, GTA5 [26]→Cityscapes [6] and SYNTHIA [27]→Cityscapes. GTA5 consists of 24,996 images with the resolution of 1914×1052 and we use the 19 common categories between GTA5 and Cityscapes for training and testing. For SYNTHIA dataset, we use the SYTHIA-RAND-CITYSCAPES set which contains 9,400 images with resolution 1280×760 and 16 common categories with Cityscapes. Cityscapes is split into training set, validation set and testing set. Training set contains 2,975 images with resolution 2048×1024. Following common practice, we report the results on the validation set which contains 500 images with same resolution. All ablation studies are performed on GTA5→Cityscapes, and comparison with state-of-the-art is performed on both GTA5→Cityscapes and SYNTHIA→Cityscapes. We use category-wise IoU and mIoU to evaluate the performance.

4.2. Network Architecture

Following common practice, we use DeepLab-V2 [4] with ResNet-101 [8] and FCN-8s [18] with VGG16 [29] as our semantic segmentation models. The discriminator used in adversarial learning is similar to [24], which has 5 convolutional layers with kernel size 4×4 with channel number {64, 128, 256, 512, 1} and stride of 2. For each of convolutional layer except the last one, a leaky ReLU [40] layer parameterized by 0.2 is followed. The discriminator is implemented over the softmax output of segmentation model. For DPIT, following [16], we adopt the architecture of CycleGAN with 9 blocks and use the proposed dual perceptual loss to maintain visual consistency.

4.3. Implementation Details

When training DPIT, the input image is randomly cropped to the size 512×256 and it is trained for 40 epochs.

Table 1: Comparison of different image translation models.

Image translation module	mIoU($M_{\mathcal{S}}^{(1)}$)	mIoU($M_{\mathcal{T}}^{(1)}$)
CycleGAN	41.4	48.5
SPIT	48.6	51.1
DPIT	**49.6**	**51.8**

Table 2: Comparison of different pseudo label generation strategies.

Pseudo label generation strategy	mIoU($M_{\mathcal{S}}^{(1)}$)	mIoU($M_{\mathcal{T}}^{(1)}$)
SPPLG	46.0	50.0
DPPLG-Max	49.2	50.6
DPPLG-Joint	49.1	50.3
DPPLG-Weighted	**49.6**	**51.8**

The learning rate of first 20 epochs is 0.0002 and decreases to 0 linearly after 20 epochs. Following [47, 16], in Equation 5, λ_{Recon} is set to 10, $\lambda_{DualPer}$ is set to 0.1, respectively. For DPAS training, the input images are resized to the size 1024×512 with batch size 4. For DeepLab-V2 with ResNet-101, we adopt SGD as optimizer and set initial learning rate with 5×10^{-4}, which is decreased with 'poly' learning rate policy with power as 0.9. For FCN-8s with VGG16, we use Adam optimizer with momentum $\{0.9, 0.99\}$ and initial learning rate is set to 2×10^{-5}. The learning rate is decreased with 'step' policy with step size 50000 and drop factor 0.1. For adversarial learning, λ_{adv} is set to 1×10^{-3} for DeepLab-V2 and 1×10^{-4} for FCN-8s in Equation 2 and 8. The discriminator is trained with Adam optimizer with the initial learning rate 2×10^{-4}. The momentum parameters are set as 0.9 and 0.99. All ablation studies are conducted on the first iteration ($N = 1$). We set $N = 4$ when comparing with state-of-the-art methods.

4.4. Experiments

Dual Path Image Translation Improves Translation Quality. DPIT encourages visual consistency through dual perceptual loss computed by segmentation models $M_{\mathcal{S}}$ and $M_{\mathcal{T}}$. To demonstrate the effectiveness of DPIT, we compare it with: 1) naive CycleGAN, in which no perceptual loss is used to maintain visual consistency; 2) Single Path Image Translation (SPIT) used in BDL [16], which applies CycleGAN and perceptual loss computed by single segmentation model aligned in target domain. Notice the only difference in this ablation study is that different image translation methods are used in DPL. Table 1 shows the comparison. By using perceptual loss to maintain visual consistency, both SPIT and DPIT can significantly improve the adaptation performance compared with naive CycleGAN. Our DPIT surpasses SPIT in both segmentation models ($M_{\mathcal{S}}$ and $M_{\mathcal{T}}$) demonstrates that extracting aligned perceptual features can further alleviate visual inconsistency caused by image translation.

The Effectiveness of Dual Path Pseudo Label Generation. In our proposed DPPLG module, predictions from two paths jointly participate in the generation of pseudo labels. We compare DPPLG with single path pseudo label generation (SPPLG) method, i.e., path-$\mathcal{S}$ and path-$\mathcal{T}$ generate respective pseudo labels by themselves. Meanwhile, we study three different strategies of DPPLG: 1) DPPLG-Max, which selects the prediction with maximum probability of two paths; 2) DPPLG-Joint, in which two paths generate pseudo labels separately and intersections are selected as final pseudo labels; 3) DPPLG-Weighted, which is the default strategy as described in Section 3.3. Table 2 shows the results. All of the DPPLG strategies have better performance than SPPLG, which means the joint decision of two complementary paths can improve the quality of pseudo labels. We use DPPLG-Weighted as our pseudo label generation strategy due to the preeminent experimental result.

Table 3: Ablation study on stage-wise DPAS.

$M_{\mathcal{S}}$	mIoU	$M_{\mathcal{T}}$	mIoU
$M_{\mathcal{S}}^{(0)}$	43.7	$M_{\mathcal{T}}^{(0)}$	48.5
$M_{\mathcal{S}}^{(1)}$	49.6	$M_{\mathcal{T}}^{(1)}$	51.8
$M_{\mathcal{S}}^{(2)}$	50.6	$M_{\mathcal{T}}^{(2)}$	52.4
$M_{\mathcal{S}}^{(3)}$	**50.7**	$M_{\mathcal{T}}^{(3)}$	52.6
$M_{\mathcal{S}}^{(4)}$	**50.7**	$M_{\mathcal{T}}^{(4)}$	**52.8**

Table 4: Ablation study on $M_{\mathcal{T}}$ warm-up.

Model	δ	mIoU
$M_{\mathcal{T}}$	0.2	47.4
$M_{\mathcal{T}}$	0.3	**48.5**
$M_{\mathcal{T}}$	0.5	47.3
$M_{\mathcal{T}}$ w/ $Y_{\mathcal{S}}$	-	46.2
$M_{\mathcal{T}}$ w/ $\hat{Y}_{\mathcal{S}'}$	-	44.3

The Effectiveness of Dual Path Adaptive Segmentation. We show the stage-wise results of DPAS in Table 3. When warm-up is finished, $M_{\mathcal{S}}^{(0)}$ and $M_{\mathcal{T}}^{(0)}$ achieve mIoU of 43.7 and 48.5, respectively. After first iteration, $M_{\mathcal{S}}^{(1)}$ achieves 49.6 (+13.5% improvement), and $M_{\mathcal{T}}^{(1)}$ achieves 51.8 (+6.8% improvement). The big improvements on two segmentation models demonstrates that the interactions between two complementary paths facilitate the adaptive learning mutually. Though subsequent iterations ($M_{\mathcal{S}}^{(2)}$-$M_{\mathcal{S}}^{(4)}$ and $M_{\mathcal{T}}^{(2)}$-$M_{\mathcal{T}}^{(4)}$) can still promote the performance, the improvement is limited.

Ablation Study on Label Correction Strategy. In Section 3.1, we propose a label correction strategy for $M_{\mathcal{T}}$ warm-up. Now we study different warm-up strategies as well as hyper parameters in Table 4. Recall that label correction is used to find a revised label $Y_{\mathcal{S}'}$ by considering both ground-truth labels $Y_{\mathcal{S}}$ and pseudo labels $\hat{Y}_{\mathcal{S}'}$ (see Equation 1). We ablate two extreme cases: 1) directly leverage ground-truth labels $Y_{\mathcal{S}}$ without label correction; 2) directly leverage pseudo labels $\hat{Y}_{\mathcal{S}'}$ without label correction. Results in Table 4 shows the superiority of our label correction module. We also study different δ which controls correction rate, from the table, we find δ is a less-sensitive hyper parameter which can be set as 0.3 by default.

Comparison with State-of-the-art Methods. We evaluate DPL and DPL-Dual with state-of-the-art methods on two common scenarios, GTA5→Cityscapes and SYNTHIA→Cityscapes. For each scenario, we report the results on two segmentation models, ResNet101 and VGG16. Table 5 shows the results on the scenario GTA5→Cityscapes,

Table 5: Comparison with state-of-the-art methods on GTA5→Cityscapes scenario. Red: best result. Blue: second best result.

Segmentation Model	Method	road	sidewalk	building	wall	fence	pole	t-light	t-sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bicycle	mIoU
ResNet101[8]	BDL[16]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
	SIM [36]	90.6	44.7	84.8	34.3	28.7	31.6	35.0	37.6	84.7	43.3	85.3	57.0	31.5	83.8	42.6	48.5	1.9	30.4	39.0	49.2
	FADA[34]	92.5	47.5	85.1	37.6	32.8	33.4	33.8	18.4	85.3	37.7	83.5	63.2	39.7	87.5	32.9	47.8	1.6	34.9	39.5	49.2
	Label-Driven[41]	90.8	41.4	84.7	35.1	27.5	31.2	38.0	32.8	85.6	42.1	84.9	59.6	34.4	85.0	42.8	52.7	3.4	30.9	38.1	49.5
	Kim et al. [15]	92.9	55.0	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	87.1	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2
	FDA-MBT [42]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
	TPLD [12]	94.2	60.5	82.8	36.6	16.6	39.3	29.0	25.5	85.6	44.9	84.4	60.6	27.4	84.1	37.0	47.0	31.2	36.1	50.3	51.2
	DPL	92.5	52.8	86.0	38.5	31.7	36.2	47.3	34.9	85.5	39.9	85.2	62.9	33.9	86.8	37.2	45.3	20.1	44.1	42.4	52.8
	DPL-Dual	92.8	54.4	86.2	41.6	32.7	36.4	49.0	34.0	85.8	41.3	86.0	63.2	34.2	87.2	39.3	44.5	18.7	42.6	43.1	53.3
VGG16[29]	TPLD [12]	83.5	49.9	72.3	17.6	10.7	29.6	28.3	9.0	78.2	20.1	25.7	47.4	13.3	79.6	3.3	19.3	1.3	14.3	33.5	34.1
	BDL [16]	89.2	40.9	81.2	29.1	19.2	14.2	29.0	19.6	83.7	35.9	80.7	54.7	23.3	82.7	25.8	28.0	2.3	25.7	19.9	41.3
	FDA-MBT [42]	86.1	35.1	80.6	30.8	20.4	27.5	30.0	26.0	82.1	30.3	73.6	52.5	21.7	81.7	24.0	30.5	29.9	14.6	24.0	42.2
	Kim et al. [15]	92.5	54.5	83.9	34.5	25.5	31.0	30.4	18.0	84.1	39.6	83.9	53.6	19.3	81.7	21.1	13.6	17.7	12.3	6.5	42.3
	SIM [36]	88.1	35.8	83.1	25.8	23.9	29.2	28.8	28.6	83.0	36.7	82.3	53.7	22.8	82.3	26.4	38.6	0.0	19.6	17.1	42.4
	Label-Driven[41]	90.1	41.2	82.2	30.3	21.3	18.3	33.5	23.0	84.1	37.5	81.4	54.2	24.3	83.0	27.6	32.0	8.1	29.7	26.9	43.6
	FADA[34]	92.3	51.1	83.7	33.1	29.1	28.5	28.0	21.0	82.6	32.6	85.3	55.2	28.8	83.5	24.4	37.4	0.0	21.1	15.2	43.8
	DPL	88.9	43.6	83.4	33.8	24.7	28.0	37.6	26.2	84.1	40.3	81.5	54.9	25.0	83.0	27.7	48.6	4.8	29.1	32.0	46.2
	DPL-Dual	89.2	44.0	83.5	35.0	24.7	27.8	38.3	25.3	84.2	39.5	81.6	54.7	25.8	83.3	29.3	49.0	5.2	30.2	32.6	46.5

Table 6: Comparison with state-of-the-art methods on SYNTHIA→Cityscapes scenario. Red: best result. Blue: second best result.

Segmentation Model	Method	road	sidewalk	building	wall	fence	pole	t-light	t-sign	vegetation	sky	person	rider	car	bus	motorbike	bicycle	mIoU (16)	mIoU (13)
ResNet101[8]	Kim et al. [15]	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
	BDL[16]	86.0	46.7	80.3	-	-	-	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	-	51.4
	SIM [36]	83.0	44.0	80.3	-	-	-	17.1	15.8	80.5	81.8	59.9	33.1	70.2	37.3	28.5	45.8	-	52.1
	FDA-MBT [42]	79.3	35.0	73.2	-	-	-	19.9	24.0	61.7	82.6	61.4	31.1	83.9	40.8	38.4	51.1	-	52.5
	FADA[34]	84.5	40.1	83.1	4.8	0.0	34.3	20.1	27.2	84.8	84.0	53.5	22.6	85.4	43.7	26.8	27.8	45.2	52.5
	Label-Driven[41]	85.1	44.5	81.0	-	-	-	16.4	15.2	80.1	84.8	59.4	31.9	73.2	41.0	32.6	44.7	-	53.1
	TPLD [12]	80.9	44.3	82.2	19.9	0.3	40.6	20.5	30.1	77.2	80.9	60.6	25.5	84.8	41.1	24.7	43.7	47.3	53.5
	DPL	87.4	45.5	82.7	14.8	0.7	33.0	21.9	20.0	82.9	85.1	56.4	21.7	82.1	39.5	30.8	45.2	46.9	53.9
	DPL-Dual	87.5	45.7	82.8	13.3	0.6	33.2	22.0	20.1	83.1	86.0	56.6	21.9	83.1	40.3	29.8	45.7	47.0	54.2
VGG16 [29]	CrCDA [11]	74.5	30.5	78.6	6.6	0.7	21.2	2.3	8.4	77.4	79.1	45.9	16.5	73.1	24.1	9.6	14.2	35.2	41.1
	TPLD [12]	81.3	34.5	73.3	11.9	0.0	26.9	0.2	6.3	79.9	71.2	55.1	14.2	73.6	5.7	0.5	41.7	36.0	41.3
	Kim et al. [15]	89.8	48.6	78.9	-	-	-	0.0	4.7	80.6	81.7	36.2	13.0	74.4	22.5	6.5	32.8	-	43.8
	BDL [16]	72.0	30.3	74.5	0.1	0.3	24.6	10.2	25.2	80.5	80.0	54.7	23.2	72.7	24.0	7.5	44.9	39.0	46.1
	FADA [34]	80.4	35.9	80.9	2.5	0.3	30.4	7.9	22.3	81.8	83.6	48.9	16.8	77.7	31.1	13.5	17.9	39.5	46.1
	FDA-MBT [42]	84.2	35.1	78.0	6.1	0.4	27.0	8.5	22.1	77.2	79.6	55.5	19.9	74.8	24.9	14.3	40.7	40.5	47.3
	Label-Driven [41]	73.7	29.6	77.6	1.0	0.4	26.0	14.7	26.6	80.6	81.8	57.2	24.5	76.1	27.6	13.6	46.6	41.1	48.5
	DPL	82.7	37.3	80.1	1.6	0.9	29.5	20.5	33.1	81.7	82.9	55.6	20.2	79.2	26.3	6.8	45.5	42.7	50.2
	DPL-Dual	83.5	38.2	80.4	1.3	1.1	29.1	20.2	32.7	81.8	83.6	55.9	20.3	79.4	26.6	7.4	46.2	43.0	50.5

DPL achieves state-of-the-art performance on both models (with mIoU of 52.8 on ResNet101 and 46.2 on VGG16). DPL-Dual further achieves mIoU of 53.3 on ResNet101 and 46.5 on VGG16. Domain gap between SYNTHIA and Cityscapes is much larger than that of GTA5 and Cityscapes, and their categories are not fully overlapped. We list both of the results for the 13-category and 16-category for a fair comparison with state-of-the-art methods. Results are shown in Table 6, mIoU (13) and mIoU (16) represent adaptation methods are evaluated on 13 common categories and 16 common categories, respectively. Once again, under 13-category metric, DPL achieves state-of-the-art result on both ResNet101 and VGG16, DPL-Dual further boosts performance. For 16-categories metric, the performance of DPL with ResNet101 is slightly worse since the domain shift is much larger in {*wall, fence, pole*} categories, and DPL with VGG16 still surpasses state-of-the-art with mIoU 42.7, DPL-Dual further promotes the performance to 43.0.

5. Conclusion

In this paper, we propose a novel dual path learning framework named DPL, which utilizes two complementary and interactive paths for domain adaptation of segmentation. Novel technologies such as dual path image translation and dual path adaptive segmentation are presented to make two paths interactive and promote each other. Meanwhile, a novel label correction strategy is proposed in the warm-up stage. The inference of DPL is extremely simple, only one segmentation model well aligned target domain is used. Experiments on common scenarios GTA5→Cityscapes and SYNTHIA→Cityscapes demonstrate the superiority of our DPL over the state-of-the-art methods.

References

[1] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. Dunit: Detection-based unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4787–4796, 2020. 2

[2] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1900–1909, 2019. 1

[3] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[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. 1, 6

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 2

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2, 6

[7] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 8

[9] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 1, 2, 6

[10] 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. 2

[11] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. *arXiv preprint arXiv:2007.02424*, 2020. 2, 8

[12] Fei Pan Inkyu, Sanghyun Woo and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *In European Conference on Computer Vision (ECCV), 2020.* 1, 3, 5, 6, 8

[13] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. 2

[14] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[15] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12975–12984, 2020. 2, 8

[16] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. *arXiv preprint arXiv:1904.10620*, 2019. 2, 3, 4, 5, 6, 7, 8

[17] 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. 3

[18] 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. 6

[19] Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Transferable representation learning with deep adaptation networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):3071–3085, 2018. 2

[20] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision (IJCV)*, 2021. 1

[21] 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. 1, 2

[22] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 3

[23] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015. 2

[24] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 6

[25] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 1

[26] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer

games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016. 1, 2, 6

[27] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 1, 2, 6

[28] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. *arXiv preprint arXiv:1712.02560*, 2017. 2

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6, 8

[30] M Naseer Subhani and Mohsen Ali. Learning from scale-invariant examples for domain adaptation in semantic segmentation. *arXiv preprint arXiv:2007.14449*, 2020. 3

[31] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020. 1

[32] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 2, 4

[33] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 2

[34] Haoran Wang, Tong Shen, Wei Zhang, Lingyu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. *arXiv preprint arXiv:2007.09222*, 2020. 8

[35] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019. 1

[36] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12635–12644, 2020. 8

[37] 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. 2

[38] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 1

[39] Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. *arXiv preprint arXiv:1911.04252*, 2019. 1

[40] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 6

[41] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. *arXiv preprint arXiv:2003.04614*, 2020. 2, 3, 4, 8

[42] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 8

[43] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. 2020. 1

[44] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 435–445, 2019. 1

[45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1

[46] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Tan, Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, 2019. 2

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 3, 5, 7

[48] Xiaojin Zhu. Semi-supervised learning tutorial. In *International Conference on Machine Learning (ICML)*, pages 1–135, 2007. 3

[49] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019. 1, 3, 6

[50] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 1, 2, 3