Unsupervised Pixel-Wise Weighted Adversarial Domain Adaptation

Haitao Tian[1,2](✉), Shiru Qu[1], and Pierre Payeur[2]

[1] Northwestern Polytechnical University, Xi'an, China
htian026@uottawa.ca
[2] University of Ottawa, Ottawa, Canada

Abstract. Fully convolutional networks have been leveraged extensively in semantic segmentation tasks. While possessing demonstrated competency for dense prediction, such supervised learning networks are restricted to labeled data during training and hence show poor generalization while confronting unseen domains. As a pivotal transfer learning technique, domain adaptation aims to alleviate discrepancies between distinct domain distributions to improve the performance of generalization in unsupervised manners. Although a family of domain adaptation methods have demonstrated significant effectiveness on cross-domain semantic segmentation tasks, the overlook of pixel-wise domain divergences leads to over-adaptation. To deal with this problem, we investigate effective pixel-wise inter-domain discrepancy metrics to regularize the training of adaptation networks at a pixel-wise level. We first leverage generation confidence encoded from the output space as a weighting map to impose more adaptation emphasis on deeply shifted regions. Furthermore, we employ discrimination confidence on the feature space to refine generation confidence into a more reliable weighting map. The formulation of generation and discrimination confidence does not introduce additional computations over the fundamental DA framework. In our experiments, the proposed pixel-wise weighted adaptation approach outperforms state-of-the-art methods on two cross-domain segmentation tasks and demonstrates effective alleviation of over-adaptation.

Keywords: Domain adaptation · Semantic segmentation · Pixel-wise adaptation · Generation confidence · Discrimination confidence

1 Introduction

Even though fully convolutional networks (FCNs) recently dominated the field of semantic segmentation [1–3], such networks are generally trained using a huge number of pixel-wise labeled data, which precludes its application in practical scenarios, such as autonomous driving and robotic navigation, where collecting labeled data in changing scenarios with large appearance gaps is extremely expensive. Recently developed photo-realistic synthetic street-scene datasets [4] offer an appealing workaround by simulating various scenarios for supervised network training. However, an FCN pre-trained on synthetic datasets will generally fail on real-world inference, which is referred to as

G. Bebis et al. (Eds.): ISVC 2021, LNCS 13017, pp. 586–600, 2021.
https://doi.org/10.1007/978-3-030-90439-5_46

the cross-domain shift [5, 6]. This domain shift cannot be eliminated even with a large number of synthetic data.

The domain adaptation (DA) technique is recognized for its capability of transferring learnt semantic patterns from a source domain to a target domain without using labeled data from that target domain. Therefore, DA has been well leveraged into cross-domain semantic segmentation tasks between synthetic datasets (source domain) and real-world datasets (target domain). Early research [7–9] developed a discrimination network (discriminator) on the feature space to align domain distributions. However, feature-level adaptation is limited when decoding high dimensional visual cues that have been suppressed in the feature space [10]. To overcome this issue, alternative approaches [10–12] turned to utilize structural information on the output space and hence obtained promising domain adaptation performance.

Despite such advances, pixel-wise divergence on inter-domain discrepancy is generally overlooked in conventional DA works. For instance, as shown in Fig. 1, while clear domain shift exits in between two domains, the source-trained network still preserves a capability to correctly classify parts of the pixels that are originally exhibiting light domain discrepancy in the target domain. However, conventional domain adaptation methods would have negative effectiveness on such pixels while treating all pixels under a same degree of domain shift, thereby segmenting parts of lightly shifted pixels incorrectly.

In this paper, we propose a pixel-wise weighted DA scheme for semantic segmentation. We align the cross-domain distributions by interpreting pixel-wise inter-domain discrepancy amongst the target domain pixels, so as to deploy weighted adaptation regularization that allows domain adaptation to pay more attention to deeply shifted regions than to lightly shifted ones. As such, each pixel contributes differently and separately to the adaptation loss. Moreover, given the uncertainty associated with generation confidence on discrepancy indication, we introduce discrimination confidence from an auxiliary discriminator to refine the generation confidence with the goal of further improvements.

The main contributions of this work are: 1) a pixel-wise weighted domain adaptation approach, which leverages a synthetic-data-trained segmentation network more effectively when applied on a real-world dataset; and 2) an auxiliary discriminator on the feature space, which is trained to measure the inter-domain discrepancy in a different perspective.

2 Related Work

2.1 Semantic Segmentation

Fully convolutional network [3] based models have outperformed traditional non-CNN models [13, 14] in the field of semantic segmentation. For instance, FCN- based Deeplab v2 [1] and PSPNet [2] show state-of-the-art performance on semantic segmentation tasks. Meanwhile, the development of pixel-wise labeled datasets (e.g., PASCAL VOC 2012 [15] and Cityscapes [16]) made FCNs usage possible in supervised training setting. But when confronting cross-domain segmentation tasks, the domain shift across datasets degrades the performance of many state-of-the-art FCNs. Expanding the volume of datasets is the first attempt to deal with this problem. But collecting datasets with widespread variability that cover various testing scenarios is extremely time-consuming in the real world. As such, provided that the domain shift problem can be properly handled, FCN-based segmentation networks can alternatively be trained on synthetic datasets [4, 17], which are easily generated by graphics engines, and transposed onto real-world datasets and scenarios.

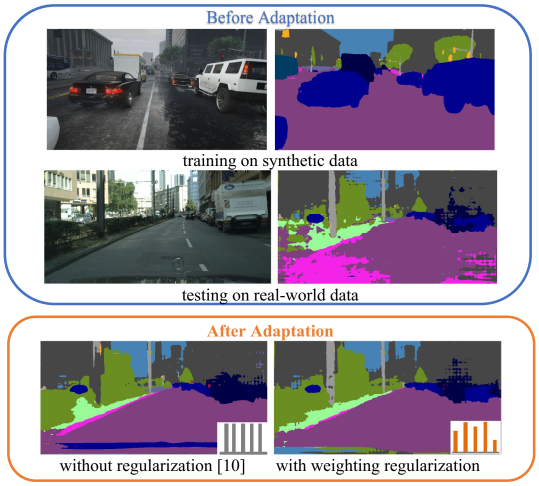

Before Adaptation
training on synthetic data
testing on real-world data
After Adaptation
without regularization [10]
with weighting regularization

Fig. 1. Illustration of proposed pixel-wise weighting adaptation. **Top 2 rows**: samples from source and target domain (lft), a source-domain pretrained model for semantic segmentation inferred on these samples without adaptation (right). **Bottom row**: prediction results on target domain are improved by output-space domain adaptation method [10], while visible over-adaptation results (negative segmentations) exist (left). The proposed pixel-wise weighting adaptation method obtains comparable improvement while it alleviates over-adaptation (right).

2.2 Adversarial Domain Adaptation

Adversarial domain adaptation [6] has been successfully leveraged in cross-domain image classification [18, 19] and object detection tasks [20, 21]. When applied to semantic segmentation [7–12, 22–25], it embeds a discrimination network (discriminator) into FCN-based segmentation networks as an adaptation component. By imposing the DA operation, the segmentation network not only learns discriminative representations but also invariant encodings from different domains. Specifically, the DA model consists of a generator and a discriminator. During training, as a generator, the segmentation network interacts with the discriminator in an adversarial manner [26] by which the discriminator is trained to upper bound the source and target domains distributions. The generator then minimizes this bound to eliminate the discrepancy between two distinct distributions. With this adversarial training process, the generator manages to produce invariant-discriminative features, such that the segmentation network can be applied on both domains.

3 Preliminaries

In this section, we introduce mathematical preliminaries on supervised segmentation model settings and unsupervised domain adaptation settings, upon which we introduce the pixel-wise weighted adaptation approach in the next section.

The goal of domain adaptation for semantic segmentation is to train an FCN-based segmentation network on the source domain and transpose it to the target domain without supervision. To this end, we set a source domain $\mathcal{D}_S = \{(\boldsymbol{X}_s^{[i]}, \boldsymbol{Y}_s^{[i]})\}_i^{n_S}$, where $\boldsymbol{X}_s^{[i]} \in \mathbb{R}^{W \times H \times 3}$ is the i th of n_S synthetic dataset images with a size of $W \times H \times 3$ in $\mathcal{D}_S$, and $\boldsymbol{Y}_s^{[i]} \in \mathbb{R}^{W \times H \times L}$ is a corresponding pixel-wise annotation label with L categories. We also set a target domain in the same way, i.e., $\mathcal{D}_T = \{(\boldsymbol{X}_t^{[i]})\}_i^{n_T}$, where $\boldsymbol{X}_t^{[i]} \in \mathbb{R}^{W \times H \times 3}$ is a real-world dataset image. Note that there are no available labels in $\mathcal{D}_T$. For clarity, $(\boldsymbol{X}_s, \boldsymbol{Y}_s)$ and $\boldsymbol{X}_t$ correspond to a random sample from $\mathcal{D}_S$ and $\mathcal{D}_T$ respectively in the rest of this paper.

3.1 Supervised Semantic Segmentation

For training a FCN-based segmentation network, a source domain sample $(\boldsymbol{X}_s, \boldsymbol{Y}_s)$ is fed into a feature encoder $\boldsymbol{F}$, then a dense classifier $\boldsymbol{C}$ takes high-dimensional feature encodings from $\boldsymbol{F}$ and produces a final soft-max probability distribution $\boldsymbol{P}_s = \boldsymbol{C}(\boldsymbol{F}(\boldsymbol{X}_s))$, where the structural prediction likelihood is $\boldsymbol{P}_s^* = \max_l(\boldsymbol{P}_s)$. The objective loss function is formed by the multi-class cross entropy with $\boldsymbol{P}_s$ and $\boldsymbol{Y}_s$, formulated as:

$$\mathcal{L}_{\text{seg}}(\text{F,C}) = -\mathbb{E}_{(\text{X}_\text{s},\text{Y}_\text{s}) \in \mathcal{D}_S}[\sum\nolimits_{(w,h)} \sum\nolimits_l \boldsymbol{I}_{\left[l = y_s^{(w,h)}\right]} \log \boldsymbol{C}\Big(\boldsymbol{F}(x_s^{(w,h)}\Big))^{(l)}] \quad (1)$$

where $\mathbb{E}[\cdot]$ is the statistical expectation and ground truth label $\boldsymbol{Y}_s$ is correspondingly encoded into one-hot vector.

As for a target sample, $\boldsymbol{X}_t \in \mathcal{D}_T$, the source-trained segmentation network under the parameter distribution of $\mathcal{D}_S$ can also generate a direct probability distribution $\boldsymbol{P}_t$, while it would hardly reflect the genuine label distribution of $\mathcal{D}_T$ because of the presence of domain shift. Meanwhile the absence of pixel annotation in $\mathcal{D}_T$ does not allow for fine-tuning on the trained segmentation network.

3.2 Unsupervised Output-Space Based Domain Adaptation

Different from feature-space based DA models that embed a discriminative network on the feature space, the output-space adaptation scheme considers the entire FCN-based segmentation model as a generator ($\boldsymbol{G} = \text{C}[\boldsymbol{F}(\cdot)]$), and embeds a discriminator $\boldsymbol{D}$ at the end of $\boldsymbol{G}$ in order to take advantage of structural information available on the output space.

During training, the $\boldsymbol{G}$ and $\boldsymbol{D}$ are updated alternatively by adversarial optimization on binary cross entropy loss (2). In detail, the discriminator $\boldsymbol{D}$ is firstly trained to classify each element on the soft-max outputs $\boldsymbol{P_s}$ and $\boldsymbol{P}_t$ into its original domain label (we denote source domain label as 0, and target domain as 1), aiming to best represent inter-domain distribution discrepancy. $\boldsymbol{D}$ is updated with back propagation by minimizing (2), formulated as:

$$\begin{aligned} &\mathcal{L}_{adv}(\boldsymbol{G},\boldsymbol{D}) = -\mathbb{E}_{\mathrm{X_s} \in \boldsymbol{D}_S}[\sum\nolimits_{(w,h)} \log(1-\boldsymbol{D}(\boldsymbol{G}(x_s^{(w,h)})))] \\ &-\mathbb{E}_{\mathrm{X_t} \in \boldsymbol{D}_T}[\sum\nolimits_{(w,h)} \log\Big(\boldsymbol{D}\Big(\boldsymbol{G}(x_t^{(w,h)}\Big)\Big)] \end{aligned} \tag{2}$$

During the discriminative network training, $\boldsymbol{G}$ is fixed for just taking part in the forward propagation.

Second, with the supervision from $\boldsymbol{D}$, the generator $\boldsymbol{G}$ is trained to produce "source-style" soft-max output $\boldsymbol{P_t}$ to eliminate the inter-domain discrepancy. $\boldsymbol{G}$ is trained by maximizing (2), while $\boldsymbol{D}$ is only allowed for forward propagation in this stage.

Combining with supervised training using (1), the segmentation network $\boldsymbol{G}$ manages to generate discriminative and invariant distributions among pixels within the source and target domains.

4 Pixel-Wise Weighted Adaptation Method

In this section, building upon the conventional output-space DA model, the proposed pixel-wise weighted adaptation approach is introduced in two stages. It first investigates the use of the generation confidence to develop a pixel-wise weighting method for output-space based adaptation. Second, it employs an auxiliary discriminator on the feature space to measure the inter-domain discrepancy from a different perspective, which contributes to further improvements.

4.1 Generation Confidence Weighted Adaptation

In Tsai *et al.* [10], although the pixel-wise adaptation process is carried out by an output-space based DA network, each element on the output space is considered under a same inter-domain discrepancy. In other words, such an approach implements a global adaptation for each target domain sample without any consideration on the pixel-wise inter-domain discrepancy. To deal with the problem, the proposed pixel-wise weighted adaptation focuses on emphasizing precise regularization on the adaptation process by using pixel-wise inter-domain discrepancy.

From the perspective of self-supervised adaptation [27, 28], the structural prediction likelihood $\boldsymbol{P}_s^*$ for a source domain sample shows the segmentation confidence of an FCN when classifying pixels into correct categories. It further infers that, for a target domain sample, $\boldsymbol{P}_t^*$ indicates the generation confidence in the context of DA, which tells the confidence of a FCN generating distributions on the target domain that are consistent with the source domain. In other words, a large value in $\boldsymbol{P}_t^*$ corresponds to small inter-domain discrepancy as the generator can segment the target-domain pixel with high generation confidence. As such, pixel-wise inter-domain discrepancy could effectively be estimated by the generation confidence.

Based on the above observation, we introduce a pixel-wise weighted adversarial adaptation framework that minimizes distribution distance across domains on the output space according to the pixel-wise inter-domain discrepancy. We consider the pixel-wise inter-domain discrepancy as a weighting map to drive the adaptation model to pay more attention on deeply shifted pixels. To further protect well aligned regions from over-adaptation, we set the pixel-wise discrepancy to a small value, here empirically set at 0.01, on well aligned regions (where the generation confidence is higher than a threshold, T). As such, the generation-confidence weighting map $d^{(w,h)}$ is formulated as:

$$d^{(w,h)} = \begin{cases} 0.01 & \text{, if } p_t^{(w,h)*} > T, \\ e^{-p_t(w,h)*} & \text{, otherwise.} \end{cases} \tag{3}$$

The generation-confidence weighting map (shown on the upper-left of Fig. 2) allows domain adaptation to pay more attention to deeply shifted regions than to lightly shifted ones. Specifically, the weighting map $d^{(w,h)}$ is utilized to pixel-wise weight conventional adversarial adaptation loss (2). As such, each pixel contributes differently and separately to the adaptation loss. Following [11], we also employ a small adaptive weight (see details in Section V.B) to stabilize the adversarial training process. Hence, the proposed pixel-wise weighted adaptation approach can be formulated as:

$$\begin{aligned} \mathcal{L}_{adv\mathrm{I}}\,(\boldsymbol{G},\boldsymbol{D}) = -\mathbb{E}_{\mathrm{X_s}\in\boldsymbol{D}_S}[\sum_{(w,h)} \log(1-\boldsymbol{D}(\boldsymbol{G}(x_s^{(w,h)})))] \\ -\mathbb{E}_{\mathrm{X_t}\in\boldsymbol{D}_T}[\textstyle\sum_{(w,h)} (d^{(w,h)}+\mu)\cdot\log(\boldsymbol{D}\Big(\boldsymbol{G}(x_t^{(w,h)}\Big)))] \end{aligned} \tag{4}$$

4.2 Discrimination Confidence Weighted Aadaptation

In this section, we introduce discrimination confidence to integrate with generation confidence for a more reliable adaptation weighting map. It is based on the observation that target domain pixels that are far away from source domain decision boundaries would be hardly interpreted by generation confidence due to the unreliable value of $p_t^{(w,h)*}$ generated by the source FCN on the output space. In this way, a part of the discrepancy-agnostic pixels would be wrongly weighted according to the second condition of (3). This factor influences the generation confidence in the adaptation procedure, which eventually is prone to over-adaptation on the target domain.

To handle the above problem, it is proposed to rather use an auxiliary discriminator $\boldsymbol{D}^{\mathrm{aux}}$ on the feature space to estimate domain divergence from a different perspective. As such, a complementary regularization is introduced that aims to further refine the generation-confidence weighting map. Specifically, we first introduce a discriminator on the feature space of the FCN for distinguishing the difference grid of the feature encodings between source and target domain images. Second, we involve the discriminator $\boldsymbol{D}^{\mathrm{aux}}$ into an auxiliary adversarial training with feature encoder $\boldsymbol{F}$ using the binary cross entropy loss, as formulated in (5).

$$\begin{aligned} \mathcal{L}_{\mathrm{aux}}\Big(\boldsymbol{F},\boldsymbol{D}^{\mathrm{aux}}\Big) = -\mathbb{E}_{\mathrm{X_s}\in\boldsymbol{D}_S}[\textstyle\sum_{(w,h)} \log(1-\boldsymbol{D}^{\mathrm{aux}}(\boldsymbol{F}(x_s^{(w,h)})))] \\ -\mathbb{E}_{\mathrm{X_t}\in\boldsymbol{D}_T}[\sum_{(w,h)} \log(\boldsymbol{D}^{\mathrm{aux}}\Big(\boldsymbol{F}(x_t^{(w,h)}\Big)))] \end{aligned} \tag{5}$$

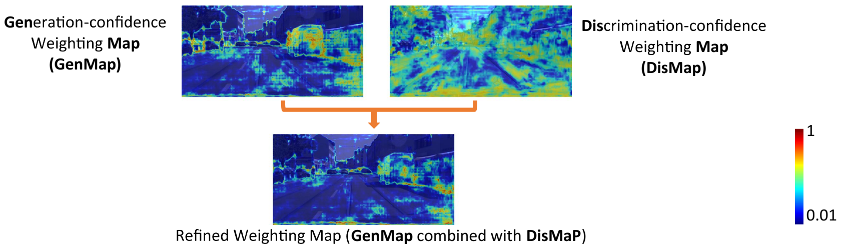

Generation-confidence Weighting Map (GenMap)
Discrimination-confidence Weighting Map (DisMap)
1
0.01
Refined Weighting Map (GenMap combined with DisMaP)

Fig. 2. Illustration of proposed weighting maps. The generation-confidence weighting map is encoded by the segmentation network that is trained on the source domain and tested on a target domain image. The discrimination-confidence weighting map is encoded by the auxiliary discremination network that is trained on both domains. The two weighting maps interpret the inter-domain discrepancy from different perspectives. The combination of the two maps, Eq. (6), provides a refined weighting map for pixel-wise adaptation regularization.

In the auxiliary adversarial training, $\boldsymbol{D}^{\text{aux}}$ allocates a membership of the target domain, $\boldsymbol{D}^{\text{aux}}(\boldsymbol{F}(x^{(h,w)})) \in [0,1]$, to each feature grid. It further infers that the encodings of $\boldsymbol{D}^{\text{aux}}$ represent the confidence of a source-trained feature encoder when generating consistent-with-target distributions. As such, pixel-wise domain discrepancy in the target domain is estimated by the discrimination confidence as $\boldsymbol{D}^{\text{aux}}(\boldsymbol{F}({x_t}^{(h,w)}))$, from a different perspective. We also visualize the discrimination-confidence weighting map on the upper-right of Fig. 2. Lastly, the generation-confidence weighting map is refined by the discrimination-confidence weighting map, which is illustrated on the bottom of Fig. 2. The weighted adversarial adaptation loss is therefore updated as in (6):

$$\begin{aligned}\mathcal{L}_{adv\mathrm{II}}(\boldsymbol{G},\boldsymbol{D}) = -\mathbb{E}_{\mathrm{X_s}\in\boldsymbol{D}_S}[\textstyle\sum_{(w,h)} \log(1-\boldsymbol{D}(\boldsymbol{G}(x_s^{(w,h)})))] \\ -\mathbb{E}_{\mathrm{X_t}\in\boldsymbol{D}_T}[\textstyle\sum_{(w,h)} (d^{(w,h)}+\mu)\cdot\boldsymbol{D}^{\text{aux}}(\boldsymbol{F}(x_t^{(w,h)}))\cdot\log(\boldsymbol{D}(\boldsymbol{G}(x_t^{(w,h)})))]\end{aligned} \tag{6}$$

4.3 Network Overview and Optimization

The proposed pixel-wise weighted adaptation network is illustrated in Fig. 3, which is formed of three components: a generator $\boldsymbol{G}$, a discriminator $\boldsymbol{D}$ (embedded on the softmax output space) and an auxiliary discriminator $\boldsymbol{D}^{\text{aux}}$(embedded on the final feature layer). The network training is carried out by the interaction between (1), (5) and (6), which are alternatively optimized according to the stages below:

- **Segmenter Updating**. The segmentation network is initially trained in a supervised way utilizing labeled source domain data $(\boldsymbol{X}_s, \boldsymbol{Y}_s) \in \mathcal{D}_S$. Parameters in $\boldsymbol{F}$ and $\boldsymbol{C}$ are updated by minimizing the loss function (1) as follows:

$$\min_{F,C}[\mathcal{L}_{seg}(\boldsymbol{F},\boldsymbol{C})] \tag{7}$$

- **Generator updating**. At this stage, unlabeled target domain data $\boldsymbol{X}_t \in \boldsymbol{D}_T$ are used to optimize the generator $\boldsymbol{G}$ and feature encoder $\boldsymbol{F}$. Note that, unlike the previous stage, $\boldsymbol{G}$ and $\boldsymbol{F}$ are updated to confuse the $\boldsymbol{D}$ and $\boldsymbol{D}^{\text{aux}}$ by which the target-generated features and soft-max output will be considered as derived more likely from the source domain. The $\boldsymbol{D}$ and $\boldsymbol{D}^{\text{aux}}$ are fixed at this step, and $\boldsymbol{G}$ and $\boldsymbol{F}$ are updated by maximizing the loss functions (5) and (6) as in (8), where λ_{adv} is the trade-off weight used to balance the supervised training in (7) and adversarial training in (8).

$$\max_{G,F} \lambda_{\text{adv}} [\mathcal{L}_{adv\text{II}}(\boldsymbol{G}, \boldsymbol{D}) + \mathcal{L}_{aux}(\boldsymbol{F}, \boldsymbol{D}^{\text{aux}})] \tag{8}$$

- **Discriminator updating**. Data and domain labels in $\mathcal{D}_S$ and $\mathcal{D}_T$ are used to update $\boldsymbol{D}$ and $\boldsymbol{D}^{\text{aux}}$ simultaneously to ensure the capability of distinguishing soft-max output and feature encodings respectively. At this stage, the $\boldsymbol{F}$ and $\boldsymbol{C}$ are fixed. The loss functions (5) and (6) are minimized as in (9).

$$\min_{D,D^{aux}} [\mathcal{L}_{adv\text{II}}(\boldsymbol{G}, \boldsymbol{D}) + \mathcal{L}_{aux}(\boldsymbol{F}, \boldsymbol{D}^{\text{aux}})] \tag{9}$$

5 Experiments

In this section, the proposed method is evaluated on two classical synthetic-to-real semantic segmentation tasks. Meanwhile we analyze the results with qualitative and quantitative comparisons to the state-of-the-art.

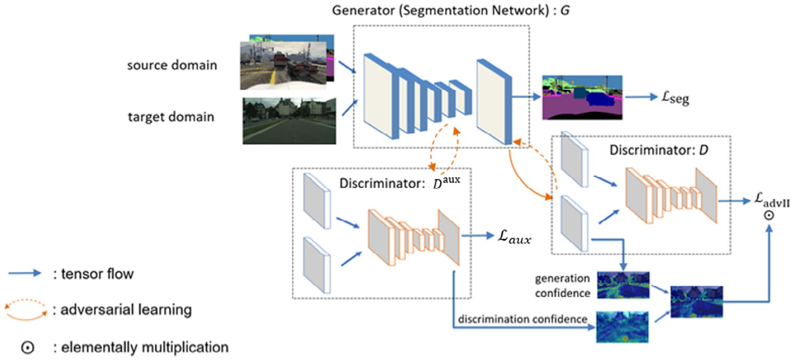


Fig. 3. A conceptual overview of the proposed weighted adaptation network. Entire network is composed of 3 different fully convolutional networks: segmentation (generation) network, $\boldsymbol{G}$, and 2 discrimination networks, $\boldsymbol{D}$ and $\boldsymbol{D}^{aux}$. The weighted adaptation is deployed through an adversarial learning with the pixel-wise weighting map.

5.1 Datasets

Cityscapes [16] is a real-world street scene dataset collected by dash cameras mounted on a moving car wandering in European cities. It contains 2,975 training images and 500 validation images, with high resolution (2048 × 1024), and pixel-wise labels in 34 categories of street objects. The training set (without labels) is considered as the target domain. **GTA5** [17] is a synthetic street scene dataset extracted from a realistically rendered computer game: Grand Theft Auto V. As rendered and annotated by a graphics engine, it forms a large dataset with 24,966 images with high resolution (1914 × 1052), and pixel-wise labels in 19 of the 34 categories of Cityscape. The entire image set with ground truth labels are used as the source domain in the task "GTA5 to Cityscapes". **SYNTHIA** [4] contains 9,400 synthetic images with 16 of the 19 categories of GTA5. The resolution of each image is 760 × 1280. While this dataset is also rendered by a graphics engine, it is less realistic than GTA5 and uses different viewing angles. That presents more severe domain shift that will challenge the adaptation capability of the proposed model. This dataset is used as the source domain in the task "SYNTHIA to Cityscapes".

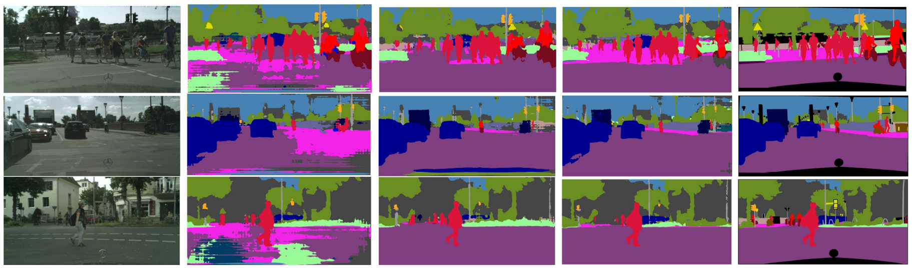
5.2 Implementation Details

The proposed network is deployed using PyTorch on a NVIDIA GTX 1080Ti GPU. Following the settings in MRNet [12], we use ResNet-101 with memory module as the FCN backbone for generator $\boldsymbol{G}$ (the segmentation network). Discriminator $\boldsymbol{D}$ comprises four convolutional layers and a classifier layer with stride 2 and kernel size 4 followed by a leaky ReLU. The auxiliary discriminator $\boldsymbol{D}^{\mathrm{aux}}$ comprises three convolution layers with stride 1 and kernel size 1 followed by a leaky ReLU, and a classifier layer with kernel size 1. Before feeding into $\boldsymbol{D}$ and $\boldsymbol{D}^{\mathrm{aux}}$, the encodings from output space and feature space are up-sampled into the input image size of $W \times H$.

The performance of the proposed method is evaluated under the metric of mean Intersection over Union (mIoU) proposed in [15]. During hyper-parameter learning, we first consider threshold T that is used to protect well aligned (high generation confidence) pixels from over-adaptation. Following self-training strategies proposed in [25, 27], we set T to 0.9 to involve sufficient well aligned pixels. In our best model, the adaptive weight μ is set to 0.1. We then consider hyper-parameter λ_{adv} as a trade-off for the supervised training in (7) and the adversarial training in (8). We follow [10] to set it as 0.0005.

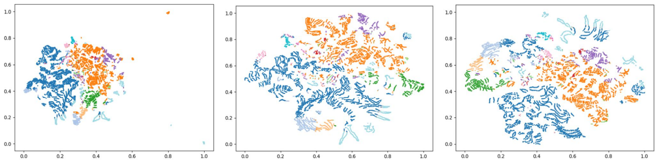
5.3 Adaptation from GTA5 to Cityscapes

The overall quantitative experimental results over 19 classes are detailed in Table 1. The "Source-only" model (trained on data from the source domain (GTA5) only and inferred on the target domain (Cityscapes) reaches 36.6% on mean IoU. The proposed method outperforms the "source-only" model by 10.6% on mean IoU after imposing domain adaptation training. We also compare against the method proposed by Tsai *et al.* [10] considered here as a vanilla method for output-space based DA network. The proposed method brings a 5.8% improvement over [10] on mean IoU and demonstrates efficient



Target image Source only CLAN [11] Proposed Manual annotation

Fig. 4. Qualitative results obtained for semantic segmentation without and with domain adaptation. The third column is the adaptation results shown in CLAN [11]. The fourth column is the proposed pixel-wise weighted adaptation results, and the fifth column is the ground truth manual annotation.



Source only CLAN [11] proposed

Fig. 5. Feature clusters visualization by t-SNE [34]. Each color represents a different class cluster.

alleviation of over-adaptation on several classes (e.g., “pole” and “bike”) as a result of the proposed pixel-wise weighted adaptation regularization.

Besides, we also present experimental comparisons with three state-of-the-art DA models [11, 12], and [29] in Table 1. Luo *et al.* [11] introduce a weighted adaptation method based on a different discrepancy measurement. The proposed method outperforms [11] on mean IoU by a margin of 4%. It shows a more efficient adaptation operation under similar complexity of the model structure (i.e., two discriminators vs. two generators). Zheng and Yang [12] recently proposed an output-space adaptation method that introduces a memory module for semantic segmentation. Conversely, Pan *et al.* [29] utilize entropy information in the intra-domain space for adaptation operation. As shown in Table 1, the proposed method 1eads to further adaptation improvements for [12] and [29] in general mean IoU, demonstrating the complementary effectiveness of the proposed adaptation method on the conventional methods that were elaborated on different DA strategies. Nevertheless, the proposed method happens to underperform on some classes (e.g., “train” and “truck”). We infer that a possible unbalance may develop in the pixel-wise weighted adaptation when several categories are involved simultaneously, which deserves further investigation.

At last, we illustrate qualitative experimental evaluation of the proposed method. First, in Fig. 4, it can be observed that the "Source only" model experiences a severe drop on segmentation performance when applied on the target domain. Compared to the "Source only" model, the proposed method provides a clear improvement in the segmentation results. Second, in comparison to [11], there are visible improvements on specific classes in the segmentation results, demonstrating the adaptation capability of the proposed model on those classes. Lastly, we visualize the feature space encodings by using t-SNE [30] in Fig. 5. It reveals how the proposed method provides more separable feature clusters for the target domain, which facilitates the domain adaptation process.

5.4 Adaptation from SYNTHIA to Cityscapes

We next demonstrate the efficacy of the proposed method on the task "SYNTHIA to Cityscapes". As detailed in Table 2, the proposed method brings a 12.3% improvement over the "Source only" model compared to a 7.3% improvement obtained in [10]. It is also worth noting that the leading performance of [29] on mean IoU in the task "*GTA5 to Cityscapes*" is here surpassed by the proposed method and by [12], which indicates that, even though all three methods leverage different underlying principles, the proposed method could obtain a more reliable performance in both cross-domain adaptation tasks. Regarding specific classes, the proposed method outperforms [12] on seven classes, such as "sky", "person" and "moto.", and outperforms [29] on nine classes, such as "side.", "person" and "bike".

5.5 Ablation Studies

Three experimental ablation studies are conducted with respect to the two proposed weighting maps (illustrated in Fig. 2), to evaluate the efficacy of each component introduced in this work. First, we examine the generation-confidence weighting map when disregarding the protection of well aligned pixels by setting T = 1 (**GenMap**$^{T=1}$). As shown in Table 3, **GenMap**$^{T=1}$ provides an improvement of 1.0% on mean IoU compared to the DA method developed upon the baseline [12] only. It indicates that while the generation-confidence weighting map might be limited by the implicit uncertainty mentioned in section IV.B, it still remains capable to implement pixel-wise adaptation regularization. The ablation study also conducts domain adaptation based on the generation confidence weighting map with a lower threshold at T = 0.9 (**GenMap**$^{T=0.9}$), and on combined generation and discrimination confidence weighting maps (**GenMap**$^{T=0.9}$+ **DisMap**), under the experimental settings detailed in section V.B. As shown in Table 3, the adaptation model with **GenMap**$^{T=0.9}$ reaches mean IoU of 46.8%, which indicates that an improvement can be achieved through applying thresholding (T = 0.9). Furthermore, the combined adaptation model **GenMap**$^{T=0.9}$+ **DisMap** reaches 47.2%, demonstrating that **DisMap** is a valid complementary strategy to further refine the generation confidence to reach higher performance.

Table 1. Experimental results of the adaptation task “GTA5 to Cityscapes” under the metric of mean IoU.

-GTA5 to Cityscapes-	Road	Side	Build	Wall	Fence	Pole	Light	Sign	Vege	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Moto	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
AdaptSegNet [10]	86.5	25.9	79.8	22.1	20.0	23.0	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	**7.2**	29.5	32.5	41.4
CLAN [11]	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
MRNet [12]	89.1	23.9	82.2	19.5	20.1	33.5	**42.2**	**39.7**	85.3	33.7	76.4	60.2	**33.7**	86.0	**36.1**	43.3	5.9	22.8	30.8	45.5
IntraDA [29]	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
Proposed	**91.4**	**55.1**	80.3	**30.6**	18.1	**35.0**	35.0	27.9	84.3	29.8	74.6	**60.9**	28.0	83.0	28.2	42.5	0.1	**35.8**	**48.9**	**47.2**

Table 2. Experimental results of the adaptation task “SYNTHIA to Cityscapes” under the metric of mean IoU.

-SYNTHIA to Cityscapes-	road	side	build	light	sign	vege	sky	person	rider	car	bus	moto	bike	mIoU
Source only	55.6	23.8	74.6	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	38.6
AdaptSegNet [10]	79.2	37.2	78.8	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.1	45.9
CLAN [11]	81.3	37.0	80.1	16.1	**13.7**	78.2	81.5	53.4	21.1	73.0	32.9	22.6	30.7	47.8
MRNet [12]	82.0	36.5	**80.4**	**18.0**	13.4	**81.1**	80.8	61.3	21.7	**84.4**	32.4	14.8	**45.7**	50.2
IntraDA [29]	84.3	37.7	79.5	9.2	8.4	80.0	84.1	57.2	23.0	78.0	**38.1**	20.3	36.5	48.9
Proposed	**85.1**	**41.2**	79.2	10.1	13.1	79.0	**85.6**	**61.7**	**26.6**	77.4	36.4	**23.4**	42.6	**50.9**

Table 3. Ablation studies in the task “GTA5 to Cityscapes”.

Methods	MRNet [12]	GenMap$^{T=1}$	GenMap$^{T=0.9}$	DisMap	mIoU
DA without pixel-wise weighting	✓				45.5
Pixel-wise weighted DA ($T = 1$)	✓	✓			46.5
Pixel-wise weighted DA ($T = 0.9$)	✓		✓		46.8
Pixel-wise weighted DA (refined w/ DisMap)	✓		✓	✓	47.2

6 Conclusion

In this work, we propose a pixel-wise weighted adversarial adaptation framework. We first utilize generation confidence to regularize the output-space adaptation process at a finer level of details. Second, we introduce an auxiliary feature space based discriminator as a complementary discrepancy indicator to refine generation confidence, which contributes additional improvements on adaptation. The experimental results demonstrate that the proposed method is effective for over-adaptation alleviation and outperforms leading state-of-the-art methods for semantic segmentation. Future work will involve designing more efficient domain discrepancy metrics and combining DA schemes to improve the performance of the proposed method for application on cross-domain semantic segmentation.

Acknowledgments. This work was supported by the China Scholarship Council (CSC) under Grant: 201906290105, and by Natural Sciences and Engineering Research Council of Canada under the Discovery grants program.

References

1. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. Trans. Patt. Recog. Mach. Intell. (PAMI) **40**(4), 834–848 (2018)

2. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890 (2017)
3. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015)
4. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3234–3243 (2016)
5. Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T.: A deeper look at dataset bias. CoRR, vol. arXiv:1505.01257 (2015)
6. Gong, B., Sha, F., Grauman, K.: Overcoming dataset bias: an unsupervised domain adaptation approach. In: NIPS Workshop on Large Scale Visual Recognition and Retrieval (LSVRR) (2012)
7. Hoffman, J., Wang, D., Yu, F., Darrell, T.: FCNS in the wild: Pixel-level adversarial and constraint-based adaptation. CoRR, abs/1612.02649 (2016)
8. Hoffman, J., et al.: Cycada: Cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning (ICML) (2018)
9. Chen, Y. H., et al.: No more discrimination: cross city adaptation of road scene segmenters. In: IEEE International Conference on Computer Vision (ICCV) (2017)
10. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7472–7481 (2018)
11. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: category-level adversaries for semantics consistent domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2507–2516 (2019)
12. Zheng, Z., Yang, Y.: Unsupervised scene adaptation with memory regularization in vivo. In: International Joint Conference on Artificial Intelligence (IJCAI) (2020)
13. Zhang, C., Wang, L., Yang, R.: Semantic segmentation of urban scenes using dense depth maps. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 708–721. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_51
14. Tighe, J., Lazebnik, S.: SuperParsing: scalable nonparametric image parsing with superpixels. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 352–365. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15555-0_26
15. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. (IJCV) **111**(1), 98–136 (2015)
16. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223 (2016)
17. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: ground truth from computer games. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 102–118. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_7
18. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning (ICML) (2015)
19. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 213–226. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_16
20. Sun, B., Saenko, K.: From virtual to reality: fast adaptation of virtual object detectors to real domains. In: BMVA British Machine Vision Conference (BMVC) (2014)
21. Vazquez, D., Lopez, A.M., Marin, J., Ponsa, D., Geronimo, D.: Virtual and real world adaptation for pedestrian detection. Trans. Pattern Recog. Mach. Intell. (PAMI) **36**(4), 797–809 (2014)

22. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R.: Learning from synthetic data: addressing domain shift for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3752–3761 (2018)
23. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3723–3732 (2018)
24. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional Learning for Domain Adaptation of Semantic Segmentation," in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6936–6945 (2019)
25. Biasetton, M., Michieli, U., Agresti, G., Zanuttigh, P.: Unsupervised domain adaptation for semantic segmentation of urban scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
26. Goodfellow, J. et al.: Generative adversarial nets. In: NIPS (2014)
27. Zou, Y., Yu, Z., Vijaya Kumar, B.V.K., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11207, pp. 297–313. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_18
28. Zou, Y., Yu, Z., Liu, X., Kumar, B.V., Wang, J.: Confidence regularized self-training. In: IEEE International Conference on Computer Vision (ICCV), pp. 5982–5991 (2019)
29. Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S: Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
30. Maaten, L.V., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11), 2579–2605 (2008)