Is It Necessary to Transfer Temporal Knowledge for Domain Adaptive Video Semantic Segmentation?

Xinyi Wu[1], Zhenyao Wu[1], Jin Wan[2], Lili Ju[1,†] , and Song Wang[1,†]

[1] University of South Carolina
{xinyiw,zhenyao}@email.sc.edu ju@math.sc.edu songwang@cec.sc.edu
[2] Beijing Jiaotong University
jinwan@bjtu.edu.cn

Abstract. Video semantic segmentation is a fundamental and important task in computer vision, and it usually requires large-scale labeled data for training deep neural network models. To avoid laborious manual labeling, domain adaptive video segmentation approaches were recently introduced by transferring the knowledge from the source domain of self-labeled simulated videos to the target domain of unlabeled real-world videos. However, it leads to an interesting question – while video-to-video adaptation is a natural idea, **are the source data required to be videos?** In this paper, we argue that it is not necessary to transfer temporal knowledge since the temporal continuity of video segmentation in the target domain can be estimated and enforced without reference to videos in the source domain. This motivates a new framework of Image-to-Video Domain Adaptive Semantic Segmentation (**I2VDA**), where the source domain is a set of images without temporal information. Under this setting, we bridge the domain gap via adversarial training based only on the spatial knowledge, and develop a novel temporal augmentation strategy, through which the temporal consistency in the target domain is well-exploited and learned. In addition, we introduce a new training scheme by leveraging a proxy network to produce pseudo-labels on-the-fly, which is very effective to improve the stability of adversarial training. Experimental results on two synthetic-to-real scenarios show that the proposed I2VDA method can achieve even better performance on video semantic segmentation than existing state-of-the-art video-to-video domain adaption approaches.

1 Introduction

Generating a dense prediction map for each frame to indicate specific class of each pixel, video semantic segmentation is a fundamental task in computer vision with important applications in autonomous driving and robotics [7,6]. Just like image semantic segmentation [24,3,43], state-of-the-art supervised learning methods for video semantic segmentation require large-scale labeled training data, which

[†] Co-corresponding authors. Code is available at github.com/W-zx-Y/I2VDA.

is costly and laborious to annotate manually [9,10,46,22,30]. Semi-supervised training [29,30,47,2] can help relieve the manual-annotation burden but still requires to annotate sparsely sampled video frames from the same domain.

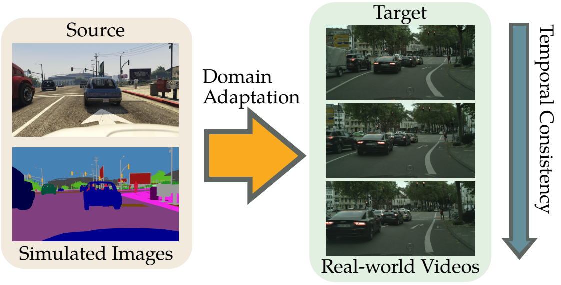


Fig. 1. An illustration of the framework setting for the proposed image-to-video domain adaptive semantic segmentation (I2VDA).

One way to avoid completely manual annotation is to train segmentation models on simulated data that are easily rendered by video game engines and therefore self-annotated, and then transfer the learned knowledge into real-world video data for improving semantic segmentation. Underlying this is actually an important concept of domain adaptation – from the source domain of simulated data to the target domain of real-world data – which was initially studied for image semantic segmentation [15,4,37,38,41,42], *e.g.*, from GTA5 [33] to Cityscapes [6] and from SYNTHIA [34] to Cityscapes with much success. This concept of domain adaptation also has been extended to tackle video semantic segmentation – a straightforward approach is to treat each video frame as an image and directly perform image-to-image domain adaptation to segment each frame independently [11]. By ignoring the temporal information along the videos, these approaches usually exhibit limited performance on video semantic segmentation.

Recent progress on video semantic segmentation witnesses two inspirational works [11,36] that coincidentally suggest video-to-video domain adaptation. Both of them employ adversarial learning of the video predictions between the source and target domains and therefore consider spatial-temporal information in both domains. While we can generate large-scale simulated videos to well reflect the source domain, it may lead to high complexity of the network and its training in the source domain. Motivated by such observation and with the goal to reduce the cost, we aim to develop a new concept of *image-to-video domain adaptive semantic segmentation* (I2VDA), where the source domain contains only simulated images and the target domain consists of real-world videos, as illustrated in Figure 1.

Videos contain spatial-temporal information and video-to-video domain adaption can exploit and pass both spatial and temporal knowledge from the source domain to the target domain. The fundamental hypothesis of the proposed image-to-video domain adaptation for semantic segmentation is that we only need to pass the spatial knowledge from the source domain to the target domain, not the temporal one. In principle, we have two major arguments for this hypothesis: 1) the between-frame continuity is the most important temporal knowledge for video semantic segmentation and such continuity can be well exploited from videos in the target domain, *e.g.,* the optical flow along each video; and 2) the temporal information between the source and target domains practically may not show a systematic domain gap that has to be filled by adaptation. On the other hand, using images, instead of videos, in the source domain can significantly reduce the required training-data size and the network complexity.

In this paper, we verify the above fundamental hypothesis by developing a new image-to-video domain adaptive semantic segmentation method. In our method, we propose a novel temporal augmentation strategy to make use of the temporal consistency in the target domain and improve the target predictions. Moreover, the domain gap is bridged by the widely-used adversarial learning strategy which only considers the spatial features in the two domains. To relieve the instability of the adversarial learning, we further introduce a new training scheme that leverages a proxy network to generate pseudo labels for target predictions on-the-fly. We conduct extensive experiments to demonstrate the effectiveness of the proposed method and each of its strategy. The main contributions of this paper are summarized as follows:

- We propose and verify a new finding – for segmenting real videos, it is sufficient to perform domain adaptation from synthetic *images*, instead of synthetic *videos*, i.e., there is no need to adapt and transfer temporal information in practice.
- We introduce for the first time the setting of image-to-video domain adaptive semantic segmentation, *i.e.,* which uses labeled images as the source domain in domain adaptation for video semantic segmentation.
- We successfully develop an I2VDA method with two novel designs: 1) a temporal augmentation strategy to better exploit and learn diverse temporal consistency patterns in the target domain; and 2) a training scheme to achieve more stable adversarial training with the help of a proxy network.
- Experimental results on two synthetic-to-real scenarios demonstrate the effectiveness of the proposed method and verify our fundamental hypothesis. Without simulating/adapting temporal information in the source domain, our method still outperforms existing state-of-the-art video-to-video domain adaptation methods.

2 Related Works

Video semantic segmentation Existing video semantic segmentation approaches can be categorized into accuracy-oriented and efficiency-oriented ones. Optical-flow-based representation warping and multi-frame prediction fusion have been employed to achieve more robust and accurate results [10,46,22]. An alternative solution is to use the gated recurrent layers to extract the temporal information [9] or propagate labels to unlabeled frames by means of optical flow [30]. Many strategies have been studied to improve efficiency. For example, features in each frame can be reused by adjacent frames to reduce the overall cost [35,46,40]. Li *et al.* [19] further proposed to reduce both of computational cost and maximum latency by adaptive feature propagation and key-frame allocation. More recently, Liu *et al.* [23] proposed to train a compact model via temporal knowledge distillation for real-time inference.

All of the above video semantic segmentation methods need the labeling on densely or sparsely sampled frames from the target domain for training. In this paper, we instead use self-labeled simulated images for training and then adapt to the target domain for video semantic segmentation.

Domain adaptive image segmentation In recent years, many domain adaptation approaches have been proposed for image semantic segmentation to relieve the burden of dense pixel-level labeling. Hoffman *et al.* [15] introduced the first unsupervised domain adaptation method for transferring segmentation FCNs [24] by applying adversarial learning on feature representations, which has become a standard strategy for domain adaptive semantic segmentation [4,25]. More recently, the adversarial learning has been further extended to image level [14,5], output level [37,1] and entropy level [38,31] for this task.

In [48], Zou *et al.*first suggested the self-training in the form of a self-paced curriculum learning scheme for segmentation by generating and selecting pseudo labels based on confidence scores. Following this, many alter works on semantic segmentation directly integrate self-training [20,41,18] or refine it by confidence regularization [49], self-motivated curriculum learning [21], uncertainty estimation [44], instance adaptive selection [27], prototypical pseudo label denoising [42] and domain-aware meta-learning [12].

As mentioned earlier, while these image-to-image domain adaptation methods can be applied to video segmentation by processing each frame independently [11], their performance is usually limited by ignoring the temporal information in the videos.

Domain adaptive video segmentation Recently, Guan *et al.* [11] made the first attempt at video-to-video domain adaptive semantic segmentation, in which both cross-domain and intra-domain temporal consistencies are considered to regularize the learning. The former is achieved by the adversarial learning of the spatial-temporal information between the source and target domains and the latter by passing the confident part of the flow-propagated prediction between adjacent frames. Concurrently, Shin *et al.* [36] also introduced the concept of domain adaptive video segmentation and propose a two-stage solution – the adversarial learning at the clip level first, followed by the target-domain learning with the refined pseudo labels. As mentioned earlier, while our work also performs domain adaptive video segmentation, it differs from the above two works in terms of the source domain setting – they use videos but we instead use images.

3 Proposed Method

3.1 Problem setting

The goal of image-to-video domain adaptive semantic segmentation is to transfer only spatial knowledge from a labeled source domain $\mathtt{S}$ to an unlabeled target

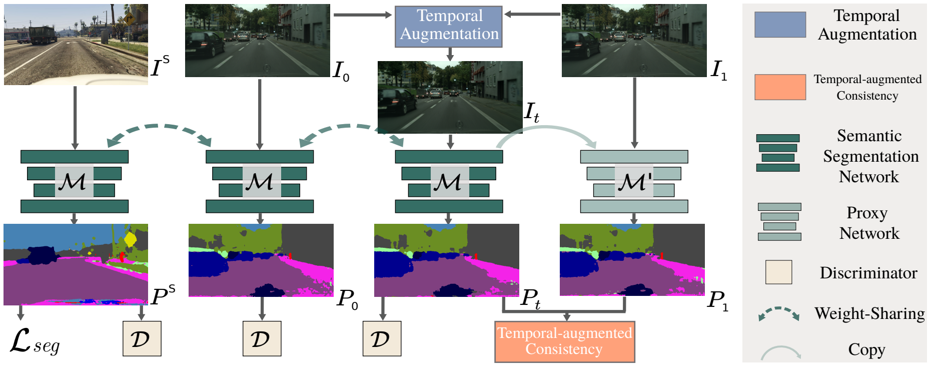


Fig. 2. The framework of the proposed image-to-video domain adaptive semantic segmentation. During training, the framework requires three inputs including a source image $I^{\mathtt{S}}$ and two consecutive frames I_0 and I_1 from a target video $I^{\mathtt{T}}$. First, an intermediate target frame I_t $(0 < t < 1)$ is synthesized using I_0 and I_1 via a frame interpolation with temporal augmentation. Then, $I^{\mathtt{S}}$, I_0 and I_t are fed into a weight-sharing semantic segmentation network $\mathcal{M}$ to obtain the corresponding predictions. A semantic segmentation loss $\mathcal{L}_{seg}$ is computed using the prediction of $I^{\mathtt{S}}$ and its label $GT^{\mathtt{S}}$. A discriminator $\mathcal{D}$ is employed to distinguish outputs from the source domain $\mathtt{S}$ and target domain $\mathtt{T}$. Besides, a proxy network $\mathcal{M}'$ takes I_1 as the input to generate its pseudo label which is used for ensuring the temporal consistency of the target predictions. Note that the parameters of $\mathcal{M}'$ are updated via copying from $\mathcal{M}$ instead of back propagation.

domain $\mathtt{T}$. Same as the setting of domain adaptive video segmentation [11,36], the target domain is in the format of video sequences $I^{\mathtt{T}} := \{I_0^{\mathtt{T}}, I_1^{\mathtt{T}}, ..., I_n^{\mathtt{T}}, ...\}$ with $I^{\mathtt{T}} \in \mathtt{T}$. In contrast, the source domain consists of a set of image-label pairs that are not in chronological order, $(I^{\mathtt{S}}, GT^{\mathtt{S}}) \in \mathtt{S}$.

3.2 Framework overview

Our work bridges the spatial domain gap between the source and the target via adversarial learning and further considers the augmented temporal consistency in the target domain to achieve accurate predictions for the videos. In addition, a novel training scheme is introduced to improve the stability of the adversarial training. The proposed image-to-video domain adaptive semantic segmentation framework is illustrated in Figure 2. The main components include flow estimation network $\mathcal{F}$ (for temporal augmentation and consistency learning), semantic segmentation network $\mathcal{M}$ and its proxy $\mathcal{M}'$, and discriminator $\mathcal{D}$.

Flow estimation network In our work, the flow estimation network $\mathcal{F}$ is used to obtain the optical flow between two consecutive frames and the computed optical flow is used for two purposes: 1) synthesizing an intermediate frame I_t given two consecutive target frames I_0 and I_1; and 2) warping the predictions to ensure temporal consistency in the target domain. Here, we use pre-trained FlowNet2 [16] as $\mathcal{F}$ to estimate the optical flow.

Semantic segmentation network We adopt the widely-used Deeplab-v2 [3] with a backbone of ResNet-101 [13] (pre-trained on ImageNet [8]) as the semantic segmentation network $\mathcal{M}$. During training, $\mathcal{M}$ is used in a training mode to generate the predictions for I^{S}, I_0 and I_t, which are denoted as P^{S}, P_0 and P_t, respectively. Note that these predictions are upsampled to the same resolution as the input images. In addition, the proxy network $\mathcal{M}'$ has the same architecture as $\mathcal{M}$, which instead is used in an evaluation mode to generate pseudo labels given I_1 as the input. The parameters of $\mathcal{M}'$ are updated via a copy from $\mathcal{M}$ at a certain frequency.

Discriminator To perform the adversarial learning, we employ the discriminator $\mathcal{D}$ to distinguish whether the prediction is from the source domain or the target one by following [37].

3.3 The temporal augmentation strategy

From our perspective, the source does not require to be an ordered video sequence, but the temporal patterns such as frame rate and the speed of the ego-vehicle in the target domain do matter for performance improvements. As stated in [26], the temporal constraint is sensitive to object occlusions and lost frames. Here we propose a novel temporal augmentation strategy to achieve robust temporal consistency in the target domain, which is implemented based on a well-studied task – video frame interpolation [17]. Different from images, videos have the unique temporal dimension where more choices on data augmentation strategies can be applied other than those only focusing on the spatial dimension, *e.g.,* random flipping and rotation. In [47], Zhu *et al.*proposed to synthesize more image-label pairs by transforming a past frame and its corresponding label via video prediction technique for video semantic segmentation. This method can tackle the general video semantic segmentation task where only sparsely sampled video frames are labeled – the labels can be propagated to the unlabeled or synthesized frames. However, it is not applicable to our setting because of no labels in the target videos.

We carefully design a temporal augmentation strategy that is suitable for robust unlabeled video representation to improve the diversity of temporal consistency in the target domain. Specifically, given two consecutive target frames I_0 and I_1, we first extract the bi-directional optical flows using the pre-trained $\mathcal{F}$ as follows:

$$F_{0\to 1} = \mathcal{F}(I_0, I_1), \quad F_{1\to 0} = \mathcal{F}(I_1, I_0). \tag{1}$$

By assuming that the optical flow field is locally smooth as [17], $F_{t\to 0}$ and $F_{t\to 1}$, for some $t \in (0, 1)$ randomly generated in each training iteration, can be approximated by:

$$F_{t\to 0} \approx tF_{1\to 0}, \quad F_{t\to 1} \approx (1-t)F_{0\to 1}. \tag{2}$$

https://github.com/NVIDIA/flownet2-pytorch

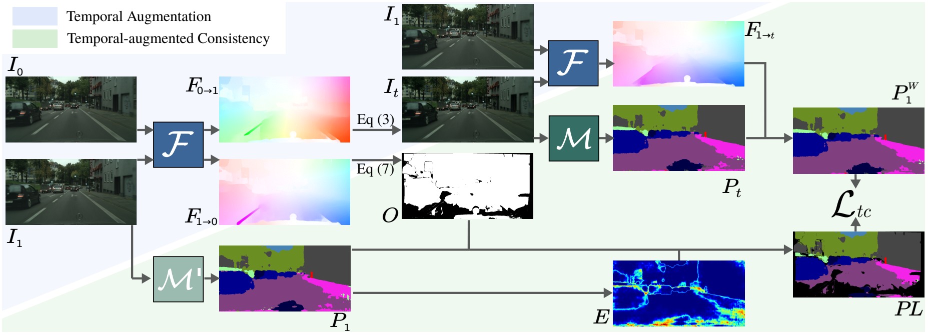


Fig. 3. An illustration of the proposed temporal augmentation strategy (Sec. 3.3) and temporal-augmented consistency learning (Sec. 3.4) in the target domain.

Then, an intermediate frame I_t can be formulated as:

$$I_t = \alpha\mathcal{W}(I_0, F_{t\to0}) + (1-\alpha)\mathcal{W}(I_1, F_{t\to1}), \tag{3}$$

where the parameter α controls the contribution of I_0 and I_1 and is set to 0.5 in all experiments, and $\mathcal{W}(\cdot,\cdot)$ is a backward warping function implemented using the bilinear interpolation [45,17].

The blue region of Figure 3 illustrated the process of the proposed temporal augmentation strategy. Next, we will show how to use the produced synthesized frame to achieve better temporal-augmented consistency learning in the target domain.

3.4 The temporal-augmented consistency learning

Temporal consistency learning is a commonly-used constraint for video-level tasks [28,30,23,39,11]. In this work, we extend this idea and propose the temporal-augmented consistency learning leveraging the synthesized frame I_t obtained via Eq. (3). The goal of this operation is to not only improve the prediction consistency between consecutive frames, but more importantly, fulfil the on-the-fly self-training to stablize the adversarial training. As illustrated in the green part of Figure 3, the temporal-augmented consistency loss computed between a propagated prediction P_1^W of I_t and a corresponding pseudo label PL. Below we detail how to achieve this temporal-augmented consistency learning.

Firstly, the target frame I_0 and the synthesized frame I_t are fed into the segmentation network $\mathcal{M}$ to obtain the corresponding segmentation predictions P_0, $P_t \in \mathbb{R}^{B\times C\times H\times W}$, where B, C, H and W are the batch size, the number of categories, the height and the width of the input image, respectively.

The prediction P_t is then propagated forward to the moment 1 to generate

$$P_1^W = \mathcal{W}(P_t, F_{1\to t}), \tag{4}$$

where $F_{1\to t}$ denotes the optical flow from the moment 1 to moment t and is computed by:

$$F_{1\to t} = \mathcal{F}(I_1, I_t). \tag{5}$$

Simultaneously, the pseudo label of P_1^W is generated via another path. The proxy network $\mathcal{M}'$ first takes the other target frame I_1 as input and output the prediction P_1 (More details related to the usage of $\mathcal{M}'$ are introduced later in Sec. 3.5). Then the prediction P_1 is rectified according to its own confidence and only the predictions with the high confidence will be kept as the pseudo labels. Following [38], we first compute the entropy map $E \in [0,1]^{B\times H\times W}$ via:

$$E = -\frac{1}{\log(C)} \sum_{k=1}^{C} \left(P_1^{(k)} \cdot \log(P_1^{(k)}) \right). \tag{6}$$

Since the synthesized frame I_t is not perfect, especially in the occlusion region, we further exclude the occlusion region in P_1 during the temporal-augmented consistency learning. Specifically, the occlusion region $O \in \mathbb{R}^{B\times H\times W}$ is defined as:

$$O = \begin{cases} 1, & \text{if } \mathcal{W}(F_{1\to 0}, F_{0\to 1}) + F_{0\to 1} < \eta. \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where η is a hyper-parameter (set to 1 in all experiments). The final rectified pseudo label PL is then given by:

$$PL = \begin{cases} \text{ArgMax}(P_1), & \text{if } E < \delta \text{ and } O = 1. \\ i, & \text{otherwise,} \end{cases} \tag{8}$$

where the threshold $\delta = 0.8$, and i is the ignored class label which is not considered during training. The rectified pseudo label PL is used to guide the prediction P_1^W which is achieved by minimizing the following **temporal-augmented consistency loss** $\mathcal{L}_{tc}$:

$$\mathcal{L}_{tc} = CE(P_1^W, PL). \tag{9}$$

Different from [23,11] which compute the L1 distance for temporal consistency, we employ the cross-entropy (CE) instead. Note that this is a non-trivial design, since Eq. (9) is also used to achieve the on-the-fly self-training. The CE loss is a common choice for self-training-based approaches [20,41,18] in domain adaptive semantic segmentation.

3.5 Proxy network for on-the-fly self-training

The usage of the proxy network is motivated by two observations: 1) the instability of the adversarial training strategy in existing domain adaptation approaches [37,38]; and 2) the self-training technique requires multiple training stages but is not able to improve the performance on the target domain. Therefore, in this paper we propose to employ a proxy network $\mathcal{M}'$ to implicitly generate the pseudo labels for P_t on-the-fly. Specifically, $\mathcal{M}'$ gets starting to work after a few training iterations and it is used only in an evaluation mode. The parameters of $\mathcal{M}'$ will be updated via copying from M at every a fixed number of iterations.

3.6 Pipeline and other training objectives

In summary, we describe the whole training pipeline in Algorithm 1 with the involved loss functions listed and discussed below.

Algorithm 1 – I2VDA

```
Input: Source images $\{I^{\mathrm{S}}\}$, source labels $\{GT^{\mathrm{S}}\}$, two consecutive frames $\{I_0, I_1\}$ from
    target videos, the base segmentor $\mathcal{M}$ with parameter $\theta_{\mathcal{M}}$, the proxy network $\mathcal{M}'$
    with parameter $\theta_{\mathcal{M}'}$ and discriminator $\mathcal{D}$ with parameter $\theta_{\mathcal{D}}$, the max number of
    training iterations MAX_ITER, the copying frequency ITER_COPY, the training
    iterations ITER_LAUNCH before launching $\mathcal{M}'$.
Output: Optimal $\theta_{\mathcal{M}}$
iter = 0
for iter<MAX_ITER do
    Synthesize $I_t$ via Eq. (3);
    Feed $I^{\mathrm{S}}$, $I_0$ and $I_t$ into $\mathcal{M}$ to obtain the predictions;
    if iter % ITER_COPY then
        Update the parameters: $\theta_{\mathcal{M}'} \leftarrow \theta_{\mathcal{M}}$;
    end if
    if iter < ITER_LAUNCH then
        Update $\theta_{\mathcal{M}}$ using $(\mathcal{L}_{seg} + 0.01\mathcal{L}_{adv})$;
    else
        Feed $I_1$ into $\mathcal{M}'$ and obtain $PL$;
        Compute $\mathcal{L}_{tc}$ using Eq. (9);
        Update $\theta_{\mathcal{M}}$ using $(\mathcal{L}_{seg} + 0.01\mathcal{L}_{adv} + \mathcal{L}_{tc})$;
    end if
    Update $\theta_{\mathcal{D}}$ using $\mathcal{L}_d$ defined in Eq. (12); iter += 1;
end for
Return $\theta_{\mathcal{M}}$;
```

We compute the **Semantic segmentation loss** based on CE to train $\mathcal{M}$ to learn knowledge from the source domain:

$$\mathcal{L}_{seg} = CE(P^{\mathrm{S}}, GT^{\mathrm{S}}). \tag{10}$$

Minimizing the **adversarial loss** can close the gap between the source and target predictions so that the target prediction can fool the discriminator. The adversarial loss $\mathcal{L}_{adv}$ is defined as:

$$\mathcal{L}_{adv} = (\mathcal{D}(P_0) - r)^2 + (\mathcal{D}(P_t) - r)^2, \tag{11}$$

where r is the label indicating the source domain which has the same resolution as the output of the discriminator. The final loss of semantic segmentation network can be expressed as $\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{tc} + 0.01\mathcal{L}_{adv}$. Besides, the goal of the discriminator is to distinguish between the source and target predictions which is trained with the following objective function:

$$\mathcal{L}_d = (\mathcal{D}(P^{\mathrm{S}}) - r)^2 + \frac{1}{2}(\mathcal{D}(P_0) - f)^2 + \frac{1}{2}(\mathcal{D}(P_t) - f)^2, \tag{12}$$

where f is the label indicating the target domain with the same resolution as the output of discriminator.

4 Experimental results

4.1 Datasets

VIPER [32] dataset comprises 254,064 fully annotated video frames for training, validation and testing rendered from a computer game. We use 13,367 images marked as *0 with their labels as one of our source datasets. The frame resolution is $1,920 \times 1,080$. Following [11], 15 classes are considered for adaptation.

SYNTHIA [34] dataset is a synthetic dataset that consists of photo-realistic video frames rendered from a virtual city. It contains 8,000 labeled frames with a resolution of $1,280 \times 720$. We use the 850 labeled images from SYNTHIA-SEQS-04-DAWN as another source dataset. Note that we remove the temporal constraint by randomly shuffling the frames in time. Following [11], 11 classes are considered for adaptation.

Cityscapes [6] dataset focuses on semantic understanding of real urban street scenes. It contains 5,000 images with fine annotations that are split into 2,975/ 500/1,525 for training/validation/testing. Each annotated image is the 20^{th} image from a 30 frame video snippets. The resolution of each image is $2,048\times1,024$. We use it as the target domain in this work.

4.2 Experimental settings

We implement the proposed I2VDA method using Pytorch. Following [37], our semantic segmentation network $\mathcal{M}$ is trained using Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and its initial learning rate is 2.5×10^{-4}. The discriminator $\mathcal{D}$ is optimized using Adam with a β of $(0.9, 0.99)$ and its initial learning rate is 1.0×10^{-4}. We employ the polynomial decay with a power of 0.9 on the learning rates of both $\mathcal{M}$ and $\mathcal{D}$. The images in VIPER [32], SYNTHIA [34], Cityscapes [6] are resized to 896×512, 1280×768 and 1024×512, respectively. We don't perform any spatial-level data augmentation strategy during training and testing. Each experiment in this paper is run for 50,000 iterations with a batch size of 2 on two Tesla V100 GPUs. Especially for testing, we only feed each frame independently into $\mathcal{M}$ to achieve the prediction without using optical flow. The mean intersection-over-union (mIoU) is used as the main evaluation metric, for which the higher the better. We also report video-specific metric of "Temporal Consistency" (TC) [23], which is again the higher the better.

https://playing-for-benchmarks.org/download/

https://synthia-dataset.net/downloads/

Table 1. Quantitative comparison results on the VIPER → Cityscapes domain adaptive video segmentation task. The best results are presented in **bold**, with the second best results underlined.

Methods	road	sidewalk	building	fence	traffic light	traffic sign	vegetation	terrain	sky	person	car	truck	bus	motorcycle	bicycle	**mIoU (%)**
Source only	56.7	18.7	78.7	6.0	22.0	15.6	81.6	18.3	80.4	59.9	66.3	4.5	16.8	20.4	10.3	37.1
AdvEnt [38]	78.5	31.0	81.5	22.1	29.2	26.6	81.8	13.7	80.5	58.3	64.0	6.9	38.4	4.6	1.3	41.2
CBST [48]	48.1	20.2	**84.8**	12.0	20.6	19.2	83.8	18.4	**84.9**	59.2	71.5	3.2	38.0	23.8	**37.7**	41.7
IDA [31]	78.7	33.9	82.3	22.7	28.5	26.7	82.5	15.6	79.7	58.1	64.2	6.4	41.2	6.2	3.1	42.0
CRST [49]	56.0	23.1	82.1	11.6	18.7	17.2	85.5	17.5	82.3	60.8	73.6	3.6	38.9	30.5	35.0	42.4
FDA [41]	70.3	27.7	81.3	17.6	25.8	20.0	83.7	31.3	82.9	57.1	72.2	22.4	49.0	17.2	7.5	44.4
DA-VSN [11]	**86.8**	**36.7**	83.5	22.9	30.2	27.7	83.6	26.7	80.3	60.0	79.1	20.3	47.2	21.2	11.4	47.8
I2VDA	84.8	36.1	84.0	**28.0**	**36.5**	**36.0**	**85.9**	**32.5**	74.0	**63.2**	**81.9**	**33.0**	**51.8**	**39.9**	0.1	**51.2**

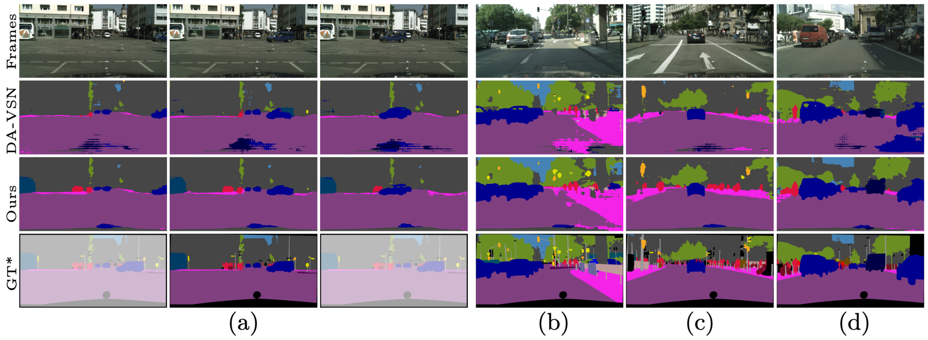


Fig. 4. Qualitative comparison results on the VIPER → Cityscapes domain adaptive video segmentation task. (a) The first three columns show the predictions of three consecutive frames. *Only one frame has ground truth in each video (30 frames). (b)-(d) show three other independent results from the Cityscapes validation set.

4.3 Comparison with state-of-the-art methods

VIPER → Cityscapes We first compare our I2VDA method with the existing state-of-the-art methods, including [11,41,49,31,48,48,38] as in [11], for the VIPER → Cityscapes scenario. The quantitative results are reported in Table 1. We find our method significantly outperforms (51.2% mIoU) all the others that are trained with VIPER videos, *i.e.,* use additional unlabeled frames that are adjacent to the labeled one for temporal modeling. Besides, these video-to-video domain adaptation approaches require two images and a pre-computed optical flow as inputs during testing, while our method performs only a per-frame inference without optical-flow computation. On the video-level evaluation, our method achieves 66.01% on TC metric, while DAVSN [11] obtain 63.82%. This shows that our proposed method can generate more consistent prediction across frames. The video samples that we provide in the supplemental materials also

Table 2. Quantitative comparison results on the SYNTHIA → Cityscapes domain adaptive video segmentation task.

Methods	road	sidewalk	building	pole	traffic light	traffic sign	vegetation	sky	person	rider	car	**mIoU (%)**
Source only	56.3	26.6	75.6	25.5	5.7	15.6	71.0	58.5	41.7	17.1	27.9	38.3
AdvEnt [38]	85.7	21.3	70.9	21.8	4.8	15.3	59.5	62.4	46.8	16.3	64.6	42.7
CBST [48]	64.1	30.5	78.2	**28.9**	14.3	21.3	75.8	62.6	46.9	20.2	33.9	43.3
IDA [31]	87.0	23.2	71.3	22.1	4.1	14.9	58.8	67.5	45.2	17.0	73.4	44.0
CRST [49]	70.4	31.4	**79.1**	27.6	11.5	20.7	**78.0**	67.2	**49.5**	17.1	39.6	44.7
FDA [41]	84.1	32.8	67.6	28.1	5.5	20.3	61.1	64.8	43.1	19.0	70.6	45.2
DA-VSN [11]	89.4	31.0	77.4	26.1	9.1	20.4	75.4	74.6	42.9	16.1	**82.4**	49.5
I2VDA (Ours)	**89.9**	**40.5**	77.6	27.3	**18.7**	**23.6**	76.1	**76.3**	48.5	**22.4**	82.1	**53.0**

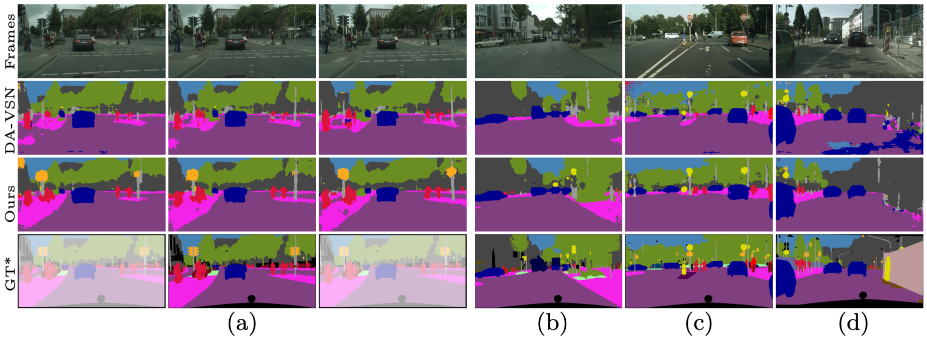


Fig. 5. Qualitative comparison results on the SYNTHIA → Cityscapes domain adaptive video segmentation task. (a) The first three columns show the predictions of three consecutive frames. *Only one frame has ground truth in each video (30 frames). (b)-(d) show three other independent results from the Cityscapes validation set.

verify this conclusion. We also present sample qualitative results for VIPER → Cityscapes scenario in Figure 4. It can be observed that our method visually achieves better performance than the second best approach DA-VSN (but the best among all existing ones). Although our method does not contain temporal modeling during testing, our predictions in Figure 4 (a) still show better temporal consistency than those by DA-VSN. In the spatial level, our segmentation results look also more accurate, *e.g*, bus in (a), person in (a) and (c), car in (a)-(d).

SYNTHIA → Cityscapes The quantitative comparison results for SYNTHIA → Cityscapes scenario are reported in Table 2, where our method still achieves the best performance and surpasses the second best (DA-VSN) by 2.6% mIoU. Sample qualitative comparison results for this adaptation scenario are shown in Figure 5, and our method still achieves more consistent and accurate segmentation results.

4.4 Ablation studies

On the framework design To verify the effectiveness of each component of the I2VDA framework, we conduct a comprehensive ablation study with several model variants. The results under the VIPER → Cityscapes scenario are reported in Table 3. The variant in the first row serves as the baseline which trains the semantic segmentation network with the labeled source domain and employs the adversarial learning to close the domain gap [37], and the last row is the I2VDA method with full settings. We find that the proposed temporal augmentation strategy and temporal consistency learning are both very effective and can achieve 4.2% and 5.6% gains, respectively, over the baseline. Another observation is that the temporal augmentation strategy only obtain 1.6% mIoU gain on its own, but it will play a much greater role (51.2% vs. 44.0%) when combined with the temporal consistency learning. In addition, rows 4-5 show the effectiveness of some designs inside the temporal consistency learning including the consideration of occlusion and entropy.

Table 3. Ablation study on the I2VDA framework designs under the VIPER → Cityscapes scenario.

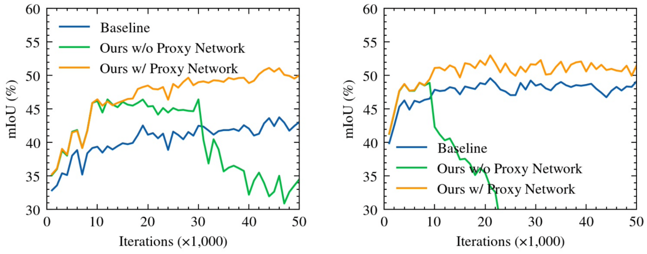
Variants	mIoU (%)
Baseline	44.0
w/o Temporal Augmentation	47.0
w/o Temporal-augmented Consistency	45.6
w/o Occlusion O in Eq. (8)	49.7
w/o Entropy Map E in Eq. (8)	49.6
Full I2VDA settings	**51.2**

On the proxy network The proxy network also plays an important role in the temporal-augmented consistency learning. We conduct experiments on the choice of copying frequency (ITER_COPY) and the training iterations before launching (ITER_LAUNCH). From Table 4, we find that copying every 8,000 iterations and launching the proxy network after 8,000 iterations achieves the best performance. In addition, as shown in Figure 6, the use of the proxy network does improve the training stability effectively. The baseline here is the same as the one in Table 3.

Table 4. Ablation study on ITER_COPY and ITER_LAUNCH for the proxy network under the VIPER → Cityscapes scenario. The ITER_LAUNCH is fixed to 8,000 for the first sub-table and the ITER_LAUNCH is fixed to 8,000 for the second sub-table.

ITER_COPY	1k	8k	15k
mIoU(%)	48.9	**51.2**	49.6

ITER_LAUNCH	1k	8k	15k
mIoU(%)	48.3	**51.2**	50.2



(a) VIPER → Cityscapes (b) SYNTHIA → Cityscapes

Fig. 6. The mIoU performance vs. varying adaptation iterations.

Table 5. Ablation study on η, δ, t and α under the VIPER → Cityscapes scenario.

η	0.1	1.0	2.0
mIoU(%)	50.4	**51.2**	50.2

δ	0.1	0.3	0.5
mIoU(%)	48.5	**51.2**	49.7

t	0	0.25	0.5	0.75	1	random
mIoU	47.0	47.0	48.4	48.1	47.4	**51.2**

α	0.1	0.3	0.5	0.7	0.9
mIoU	49.2	49.7	**51.2**	49.6	48.8

On some hyper-parameters We also conduct experiments to explore the choice of hyper-parameters involved in the temporal-augmented consistency learning. The results are reported in Table 5 where we find that our method achieves better performance when $\eta = 1.0$, $\delta = 0.3$ and $\alpha = 0.5$ and using randomly generated t.

5 Conclusion

In this paper, we found that it is not necessary to transfer temporal knowlege for domain-adaptive video semantic segmentation and have introduced for the first time the setting of image-to-video domain adaptive semantic segmentation which transfers knowledge from simulated images to real-world videos. Our I2VDA method reduces the domain gap between the source and target via adversarial training on only spatial knowledge. On the other hand, our method enhances the temporal consistency learning in the target domain by performing the temporal augmentation via frame interpolation to explore more temporal patterns and leveraging the proxy network to provide the pseudo labels on-the-fly to improve the stability of adversarial training. Experimental results on two synthetic-to-real scenarios showed that our method can outperform existing state-of-the-art video-to-video domain adaptation methods.

Acknowledgment: Dr. Lili Ju's work is partially supported by U.S. Department of Energy, Office of Advanced Scientific Computing Research through Applied Mathematics program under grant DE-SC0022254. This work used GPUs provided by the NSF MRI-2018966.

References

1. Chang, W.L., Wang, H.P., Peng, W.H., Chiu, W.C.: All about structure: Adapting structural information across domains for boosting semantic segmentation. In: CVPR. pp. 1900–1909 (2019) 4
2. Chen, L.C., Lopes, R.G., Cheng, B., Collins, M.D., Cubuk, E.D., Zoph, B., Adam, H., Shlens, J.: Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In: ECCV. pp. 695–714. Springer (2020) 2
3. 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. IEEE TPAMI **40**(4), 834–848 (2017) 1, 6
4. Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Frank Wang, Y.C., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: ICCV. pp. 1992–2001 (2017) 2, 4
5. Chen, Y.C., Lin, Y.Y., Yang, M.H., Huang, J.B.: Crdoco: Pixel-level domain transfer with cross-domain consistency. In: CVPR. pp. 1791–1800 (2019) 4
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016) 1, 2, 10
7. Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Indoor semantic segmentation using depth information. arXiv preprint arXiv:1301.3572 (2013) 1
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009) 6
9. Fayyaz, M., Saffar, M.H., Sabokrou, M., Fathy, M., Klette, R., Huang, F.: Stfcn: spatio-temporal fcn for semantic video segmentation. arXiv preprint arXiv:1608.05971 (2016) 2, 3
10. Gadde, R., Jampani, V., Gehler, P.V.: Semantic video cnns through representation warping. In: ICCV. pp. 4453–4462 (2017) 2, 3
11. Guan, D., Huang, J., Xiao, A., Lu, S.: Domain adaptive video segmentation via temporal consistency regularization. In: ICCV. pp. 8053–8064 (2021) 2, 4, 5, 7, 8, 10, 11, 12
12. Guo, X., Yang, C., Li, B., Yuan, Y.: Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In: CVPR. pp. 3927–3936 (2021) 4
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 6
14. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: Int. Conf. Machine Learning. pp. 1989–1998. PMLR (2018) 4
15. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016) 2, 4
16. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR. pp. 2462–2470 (2017) 6
17. Jiang, H., Sun, D., Jampani, V., Yang, M.H., Learned-Miller, E., Kautz, J.: Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In: CVPR. pp. 9000–9008 (2018) 6, 7
18. Kim, M., Byun, H.: Learning texture invariant representation for domain adaptation of semantic segmentation. In: CVPR. pp. 12975–12984 (2020) 4, 8

19. Li, Y., Shi, J., Lin, D.: Low-latency video semantic segmentation. In: CVPR. pp. 5997–6005 (2018) 4
20. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: CVPR. pp. 6936–6945 (2019) 4, 8
21. Lian, Q., Lv, F., Duan, L., Gong, B.: Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In: ICCV. pp. 6758–6767 (2019) 4
22. Liu, S., Wang, C., Qian, R., Yu, H., Bao, R., Sun, Y.: Surveillance video parsing with single frame supervision. In: CVPR. pp. 413–421 (2017) 2, 3
23. Liu, Y., Shen, C., Yu, C., Wang, J.: Efficient semantic video segmentation with per-frame inference. In: ECCV. pp. 352–368. Springer (2020) 4, 7, 8, 10
24. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015) 1, 4
25. Luo, Y., Liu, P., Guan, T., Yu, J., Yang, Y.: Significance-aware information bottleneck for domain adaptive semantic segmentation. In: ICCV. pp. 6778–6787 (2019) 4
26. Maninis, K.K., Caelles, S., Chen, Y., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: Video object segmentation without temporal information. IEEE TPAMI **41**(6), 1515–1530 (2018) 6
27. Mei, K., Zhu, C., Zou, J., Zhang, S.: Instance adaptive self-training for unsupervised domain adaptation. In: ECCV. pp. 415–430. Springer (2020) 4
28. Miksik, O., Munoz, D., Bagnell, J.A., Hebert, M.: Efficient temporal consistency for streaming video scene analysis. In: International Conference on Robotics and Automation. pp. 133–139. IEEE (2013) 7
29. Mustikovela, S.K., Yang, M.Y., Rother, C.: Can ground truth label propagation from video help semantic segmentation? In: ECCV. pp. 804–820. Springer (2016) 2
30. Nilsson, D., Sminchisescu, C.: Semantic video segmentation by gated recurrent flow propagation. In: CVPR. pp. 6819–6828 (2018) 2, 3, 7
31. Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S.: Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: CVPR. pp. 3764–3773 (2020) 4, 11, 12
32. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: ICCV. pp. 2213–2222 (2017) 10
33. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV. pp. 102–118. Springer (2016) 2
34. 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: CVPR. pp. 3234–3243 (2016) 2, 10
35. Shelhamer, E., Rakelly, K., Hoffman, J., Darrell, T.: Clockwork convnets for video semantic segmentation. In: ECCV. pp. 852–868. Springer (2016) 4
36. Shin, I., Park, K., Woo, S., Kweon, I.S.: Unsupervised domain adaptation for video semantic segmentation. arXiv preprint arXiv:2107.11052 (2021) 2, 4, 5
37. 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: CVPR. pp. 7472–7481 (2018) 2, 4, 6, 8, 10, 13
38. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR. pp. 2517–2526 (2019) 2, 4, 8, 11, 12

39. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. IEEE TIP **28**(11), 5596–5609 (2019) 7
40. Xu, Y.S., Fu, T.J., Yang, H.K., Lee, C.Y.: Dynamic video segmentation network. In: CVPR. pp. 6556–6565 (2018) 4
41. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: CVPR. pp. 4085–4095 (2020) 2, 4, 8, 11, 12
42. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: CVPR. pp. 12414–12424 (2021) 2, 4
43. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017) 1
44. Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. IJCV **129**(4), 1106–1120 (2021) 4
45. Zhou, T., Tulsiani, S., Sun, W., Malik, J., Efros, A.A.: View synthesis by appearance flow. In: ECCV. pp. 286–301. Springer (2016) 7
46. Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y.: Deep feature flow for video recognition. In: CVPR. pp. 2349–2358 (2017) 2, 3, 4
47. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: CVPR. pp. 8856–8865 (2019) 2, 6
48. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV. pp. 289–305 (2018) 4, 11, 12
49. Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J.: Confidence regularized self-training. In: ICCV. pp. 5982–5991 (2019) 4, 11, 12