# Domain Adaptive Video Segmentation via Temporal Consistency Regularization

Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu*
Singtel Cognitive and Artificial Intelligence Lab for Enterprises, Nanyang Technological University
{Dayan.Guan, Jiaxing.Huang, Aoran.Xiao, Shijian.Lu}@ntu.edu.sg

## Abstract

*Video semantic segmentation is an essential task for the analysis and understanding of videos. Recent efforts largely focus on supervised video segmentation by learning from fully annotated data, but the learnt models often experience clear performance drop while applied to videos of a different domain. This paper presents DA-VSN, a domain adaptive video segmentation network that addresses domain gaps in videos by temporal consistency regularization (TCR) for consecutive frames of target-domain videos. DA-VSN consists of two novel and complementary designs. The first is cross-domain TCR that guides the prediction of target frames to have similar temporal consistency as that of source frames (learnt from annotated source data) via adversarial learning. The second is intra-domain TCR that guides unconfident predictions of target frames to have similar temporal consistency as confident predictions of target frames. Extensive experiments demonstrate the superiority of our proposed domain adaptive video segmentation network which outperforms multiple baselines consistently by large margins.*

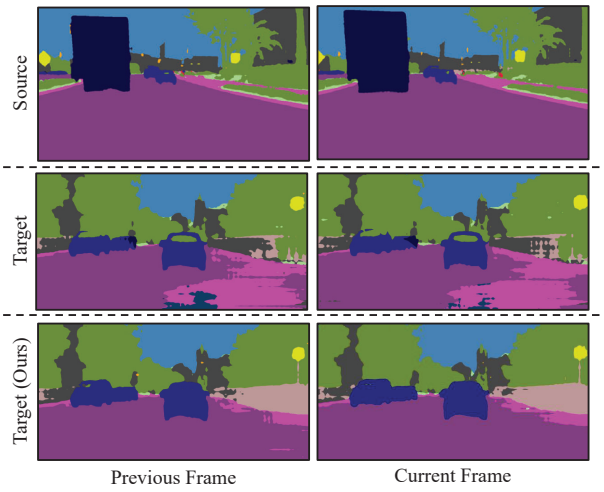


Figure 1. Temporal consistency helps in domain adaptive video segmentation: A video segmentation model trained in a *Source* domain often experiences clear performance drop while applied to videos of a *Target* domain. We employ temporal consistency, the inherent and universal nature of videos, as a constraint to regularize inter-domain and intra-domain adaptation for optimal video segmentation in target domain as in *Target (ours)*.

## 1. Introduction

Video semantic segmentation aims to assign pixel-wise semantic labels to video frames, and it has been attracting increasing attention as one essential task in video analysis and understanding [19, 53, 15, 45, 60]. With the advance of deep neural networks (DNNs), several studies have been conducted in recent years with very impressive video segmentation performance [65, 40, 20, 33, 37, 38, 26, 47]. However, most existing works require large amounts of densely annotated training videos which entail a prohibitively expensive and time-consuming annotation process [3, 14]. One approach to alleviate the data annotation constraint is to resort to self-annotated synthetic videos that are collected with computer-generated virtual scenes [62, 24], but models trained with the synthesized data often experience clear performance drops while applied to videos of natural scenes largely due to the *domain shift* as illustrated in Fig. 1.

Domain adaptive video segmentation is largely neglected in the literature despite its great values in both research and practical applications. It could be addressed from two approaches by leveraging existing research. The first approach is domain adaptive image segmentation [80, 69, 79, 58, 21, 74] which could treat each video frame independently to achieve domain adaptive video segmentation. However, domain adaptive image segmentation does not consider temporal information in videos which is very important in video semantic segmentation. The second approach is semi-supervised video segmentation [56, 78, 5] that exploits sparsely annotated video frames for segmenting unannotated frames of the same video. However, semi-supervised video segmentation was designed for consecu-

*Corresponding author.

tive video frames of the same domain and does not work well in domain adaptive video segmentation which usually involves clear domain shifts and un-consecutive video frames of different sources.

In this work, we design a domain adaptive video segmentation network (DA-VSN) that introduces temporal consistency regularization (TCR) to bridge the gaps between videos of different domains. The design is based on the observation that video segmentation model trained in a source domain tends to produce temporally consistent predictions over source-domain data but temporally inconsistent predictions over target-domain data (due to domain shifts) as illustrated in Fig. 1. We designed two complementary regularization modules in DA-VSN, namely, cross-domain TCR (C-TCR) and intra-domain TCR (I-TCR). C-TCR employs adversarial learning to minimize the discrepancy of temporal consistency between source and target domains. Specifically, it guides target-domain predictions to have similar temporal consistency of source-domain predictions which usually has decent quality by learning from fully-annotated source-domain data. I-TCR instead works from a different perspective by guiding unconfident target-domain predictions to have similar temporal consistency as confident target-domain predictions. In I-TCR, we leverage entropy to measure the prediction confidence which works effectively across multiple datasets.

The contributions of this work can be summarized in three major aspects. *First*, we proposed a new framework that introduces temporal consistency regularization (TCR) to address domain shifts in domain adaptive video segmentation. To the best of our knowledge, this is the first work that tackles the challenge of unsupervised domain adaptation in video semantic segmentation. *Second*, we designed inter-domain TCR and intra-domain TCR that improve domain adaptive video segmentation greatly by minimizing the discrepancy of temporal consistency across different domains and different video frames in target domain, respectively. *Third*, extensive experiments over two challenging synthetic-to-real benchmarks (VIPER [62] $\rightarrow$ Cityscapes-Seq [14] and SYNTHIA-Seq [63] $\rightarrow$ Cityscapes-Seq) show that the proposed DA-VSN achieves superior domain adaptive video segmentation as compared with multiple baselines.

## 2. Related Works

### 2.1. Video Semantic Segmentation

Video semantic segmentation aims to predict pixel-level semantics for each video frame. Most existing works exploit inter-frame temporal relations for robust and accurate segmentation [34, 77, 20, 46, 72, 42, 37, 26, 47]. For example, [77, 20] employs optical flow [18] to warp feature maps between frames. [42] leverages inter-frame feature propagation for efficient video segmentation with low latency. [37] presents an adaptive fusion policy for effective integration of predictions from different frames. [26] distributes several sub-networks over sequential frames and recomposes the extracted features via attention propagation. [47] presents a compact network that distills temporal consistency knowledge for per-frame inference.

In addition, semi-supervised video segmentation has been investigated which exploits sparsely annotated video frames for segmenting unannotated frames of the same videos. Two typical approaches have been studied. The first approach is based on label propagation that warps labels of sparsely-annotated frames to generate pseudo labels for unannotated frames via self-supervised learning [70, 41, 36], patch matching [1, 4], motion cues [66, 78] or optical flow [55, 77, 56, 17]. The other approach is based on self-training that generates pseudo labels through a distillation across multiple augmentations [5].

Both supervised and semi-supervised video segmentation work on frames of the same video or same domain that have little domain gaps. Our proposed domain adaptive video segmentation exploits off-the-shelf video annotations from a *source domain* for the segmentation of videos of a different *target domain* without requiring any annotations of target-domain videos.

### 2.2. Domain Adaptive Video Classification

Domain adaptive video classification has been explored to investigate domain discrepancy in action classification problem. One category of works focuses on the specific action recognition task that aims to classify a video clip into a particular category of human actions via temporal alignment [8], temporal attention [57, 13], or self-supervised video representation learning [54, 13]. Another category of works focus on action segmentation that simultaneously segments a video in time and classifies each segmented video clip with an action class via temporal alignment [9] or self-supervised video representation learning [10].

This work focuses on a new problem of domain adaptive semantic segmentation of videos, a new and much more challenging domain adaptation task as compared with domain adaptive video classification. Note that existing domain adaptive video classification methods do not work for the semantic segmentation task as they cannot generate pixel-level dense predictions for each frame in videos.

### 2.3. Domain Adaptive Image Segmentation

Domain adaptive image segmentation has been widely investigated to address the image annotation challenge and domain shift issues. Most existing methods take two typical approaches, namely, adversarial learning based [25, 67, 69, 68, 48, 32, 58, 21, 31] and self training based [80, 79, 43, 44, 7, 74, 39, 76, 52]. The adversarial learning based meth-

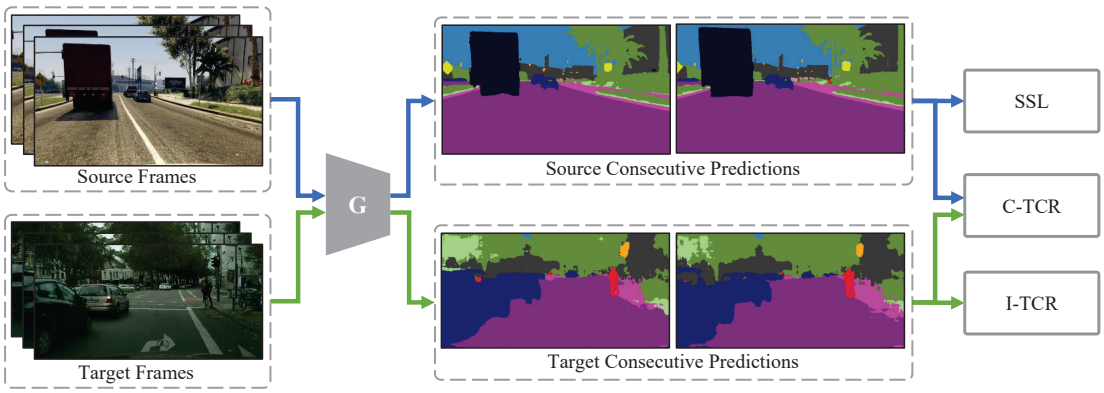


Figure 2. The framework of the proposed domain adaptive video segmentation network (DA-VSN): DA-VSN introduces temporal consistency regularization (TCR) to minimize the divergence between source and target domains. It consists of a video semantic segmentation model *G* that generates segmentation predictions, a source-domain supervised learning module (*SSL*) that learns knowledge from source domain, a cross-domain TCR component (*C-TCR*) that guides target predictions to have similar temporal consistency as source predictions, and an intra-domain TCR component (*I-TCR*) that guides unconfident target predictions to have similar temporal consistency as confident target predictions.

ods perform domain alignment by adopting a discriminator that strives to differentiate the segmentation in the space of inputs [25, 75, 12, 30, 39, 31], features [25, 11, 27, 73, 22, 28, 48] or outputs [67, 69, 49, 68, 32, 50, 71, 21, 58, 29]. The self-training based methods exploit self-training to predict pseudo labels for target-domain data and then exploit the predicted pseudo labels to fine-tune the segmentation model iteratively.

Though a number of domain adaptive image segmentation techniques have been reported in recent years, they do not consider temporal information which is critically important in video segmentation. We introduce temporal consistency of videos as a constraint and exploit it to regularize the learning in domain adaptive video segmentation.

## 3. Method

### 3.1. Problem Definition

Given source-domain video frames $X^{\mathbb{S}}$ with the corresponding labels $Y^{\mathbb{S}}$ and target-domain video frames $X^{\mathbb{T}}$ without labels, the goal of domain adaptive video segmentation is to learn a model $G$ that can produce accurate predictions $P^{\mathbb{T}}$ in target domain. According to the domain adaptation theory in [2], the target error in domain adaptation is bounded by three terms including a shared error of the ideal joint hypothesis on the source and target domains, an empirical source-domain error, and a divergence measure between source and target domains.

This work focuses on the third term and presents a domain adaptive video semantic segmentation network (DA-VSN) for minimizing the divergence between source and target domains. We design a novel temporal consistency regularization (TCR) technique for consecutive frames in target domain, which consists of two complementary components including a cross-domain TCR (C-TCR) component and an intra-domain TCR (I-TCR) component as illustrated in Fig. 2. C-TCR targets cross-domain alignment by encouraging target predictions to have similar temporal consistency as source predictions (accurate via supervised learning), while I-TCR aims for intra-domain adaptation by forcing unconfident predictions to have similar temporal consistency as confident predictions in target domain, more details to be described in the ensuing two subsections.

Note the shared error in the first term (the difference in labeling functions across domains) is usually small as proven in [2]. The empirical source-domain error in the second term actually comes from the supervised learning in the source domain. For domain adaptive video segmentation, we directly adopt video semantic segmentation loss $\mathcal{L}_{ssl}(G)$ [77, 37, 26, 47] as the source-domain supervised learning loss.

### 3.2. Cross-domain Regularization

Cross-domain temporal consistency regularization (C-TCR) aims to guide target predictions to have similar temporal consistency of source predictions which is determined by minimizing the supervised source loss $\mathcal{L}_{ssl}$ and usually has decent quality. We design a dual-discriminator structure for optimal spatial-temporal alignment of source and target *video-clips* as illustrated in Fig. 3. As Fig. 3 shows, one discriminator $D_s$ focuses on spatial alignment of a single video frame of different domains (as in domain adaptive image segmentation) and the other discriminator $D_{st}$ focuses on temporal alignment of consecutive videos frames of different domains. Since $D_{st}$ inevitably involves spatial information, we introduce a divergence loss between $D_s$ and $D_{st}$ to

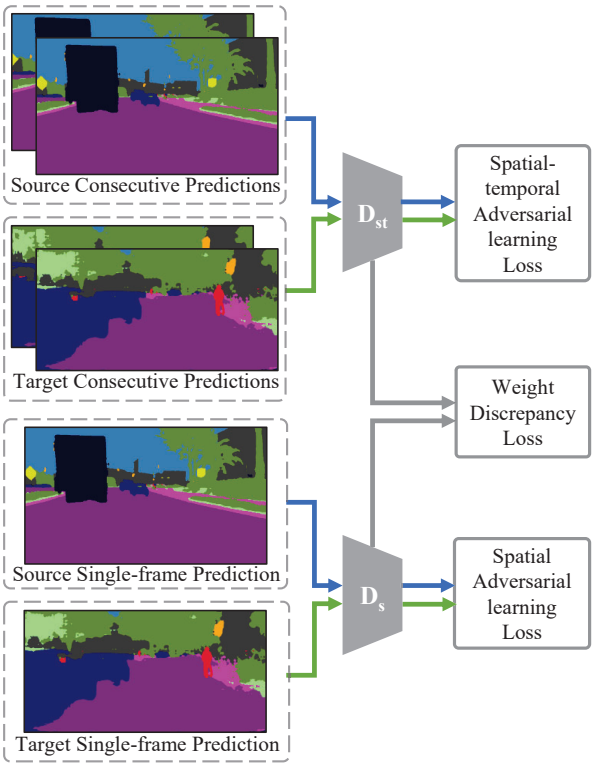


Figure 3. The framework of the proposed cross-domain temporal consistency regularization (C-TCR): C-TCR performs temporal alignment to minimize the divergence of temporal consistency between source and target domains. It introduces a spatial-temporal discriminator $D_{st}$ to align consecutive predictions (encoding spatial-temporal information) and a spatial discriminator $D_s$ to align single-frame predictions (encoding spatial information). A spatial-temporal adversarial learning loss $\mathcal{L}_{sta}$ and a spatial loss $\mathcal{L}_{sa}$ are introduced to optimize the discriminators and segmentation model. To enhance temporal alignment, we introduce a weight discrepancy loss $\mathcal{L}_{wd}$ to force $D_{st}$ to be independent from $D_s$ so that $D_{st}$ can focus more on temporal alignment.

force $D_{st}$ to focus on the alignment in temporal space.

For spatial alignment, we forward the current frame $x_k^{\mathbb{S}}$ to obtain the current prediction $p_k^{\mathbb{S}}$, and simultaneously forward $x_{k-1}^{\mathbb{S}}$ to to obtain the consecutive prediction $p_{k-1}^{\mathbb{S}}$. The spatial discriminator $D_s$ aligns frame-level predictions $p_k^{\mathbb{S}}$ and $p_k^{\mathbb{T}}$ and its objective $\mathcal{L}_{sa}$ can be formulated as follows:

$$\mathcal{L}_{sa}(G, D_s) = \log(D_s(p_k^{\mathbb{S}})) + \log(1 - D_s(p_k^{\mathbb{T}})). \quad (1)$$

For temporal alignment, we stack the current prediction $p_k^{\mathbb{S}}$ and the consecutive predictions $p_{k-1}^{\mathbb{S}}$ as $p_{k-1:k}^{\mathbb{S}}$ which encode spatial-temporal information in source domain. This same process is applied to target domain which produces two consecutive target predictions $p_{k-1:k}^{\mathbb{T}}$ that encode spatial-temporal information in target domain. The spatial-temporal discriminator $D_{st}$ then aligns $p_{k-1:k}^{\mathbb{S}}$ and $p_{k-1:k}^{\mathbb{T}}$ and its objective $\mathcal{L}_{sta}$ can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{sta}(G, D_{st}) = &\log(D_{sta}(p_{k-1:k}^{\mathbb{S}})) \\ &+ \log(1 - D_{sta}(p_{k-1:k}^{\mathbb{T}})).\end{aligned} \quad (2)$$

We enforce the divergence of the weights of $D_{st}$ and $D_s$ so that the spatial-temporal discriminator $D_{st}$ can focus more on temporal alignment. The weight divergence of the two discriminators can be reduced by minimizing their cosine similarity as follows:

$$\mathcal{L}_{wd}(D_{st}, D_s) = \frac{1}{J}\sum_{j=1}^{J} \frac{\overrightarrow{w}_{st}^{j} \cdot \overrightarrow{w}_{s}^{j}}{\left\|\overrightarrow{w}_{st}^{j}\right\| \left\|\overrightarrow{w}_{s}^{j}\right\|}, \quad (3)$$

where $J$ is the number of convolutional layers in each discriminator, $\overrightarrow{w}_{st}^{j}$ and $\overrightarrow{w}_{s}^{j}$ are obtained by flattening the weights of the $j$-th convolutional layer in the discriminators $D_{st}$ and $D_s$, respectively.

Combining the losses in Eqs 1, 2, 3, the C-TCR loss $\mathcal{L}_{ctcr}$ can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{ctcr}(G, D_{st}, D_s) = &\mathcal{L}_{sta}(G, D_{st}) \\ &+ \lambda_{sa}\mathcal{L}_{sa}(G, D_s) \\ &+ \lambda_{wd}\mathcal{L}_{wd}(D_{st}, D_s),\end{aligned} \quad (4)$$

where $\lambda_{sa}$ and $\lambda_{wd}$ are the balancing weights.

### 3.3. Intra-domain Regularization

The intra-domain temporal consistency regularization (I-TCR) aims to minimize the divergence between source and target domains by suppressing the temporal inconsistency across different target frames. As illustrated in Fig. 4, I-TCR guides unconfident target predictions to have similar temporal consistency as confident target predictions. Specifically, it first propagates predictions (of previous frames) forward by using frame-to-frame optical flow estimates, and then forces unconfident predictions in the current frame to be consistent with confident predictions propagated from the previous frame.

In the target domain, we first forward $x_k^{\mathbb{T}}$ to obtain the current prediction $p_k^{\mathbb{T}}$, and similarly forward $x_{k-1}^{\mathbb{T}}$ to obtain the previous prediction $p_{k-1}^{\mathbb{T}}$. We then adopt FlowNet [35] to estimate the optical flow $f_{x_{k-1}^{\mathbb{T}} \to x_k^{\mathbb{T}}}$ from $x_{k-1}^{\mathbb{T}}$ to $x_k^{\mathbb{T}}$. With the estimated frame-to-frame optical flow $f_{x_{k-1}^{\mathbb{T}} \to x_k^{\mathbb{T}}}$, the prediction $p_{k-1}^{\mathbb{T}}$ can be warped to generate the propagated prediction $\hat{p}_{k-1}^{\mathbb{T}}$.

To force unconfident predictions $p_k^{\mathbb{T}}$ in the current frame to be consistent with confident predictions $\hat{p}_{k-1}^{\mathbb{T}}$ propagated from the previous frames, we employ an entropy function $E$ [64] to estimate the prediction confidence and use the confident prediction $\hat{p}_{k-1}^{\mathbb{T}}$ (*i.e.*, with low entropy) to optimize

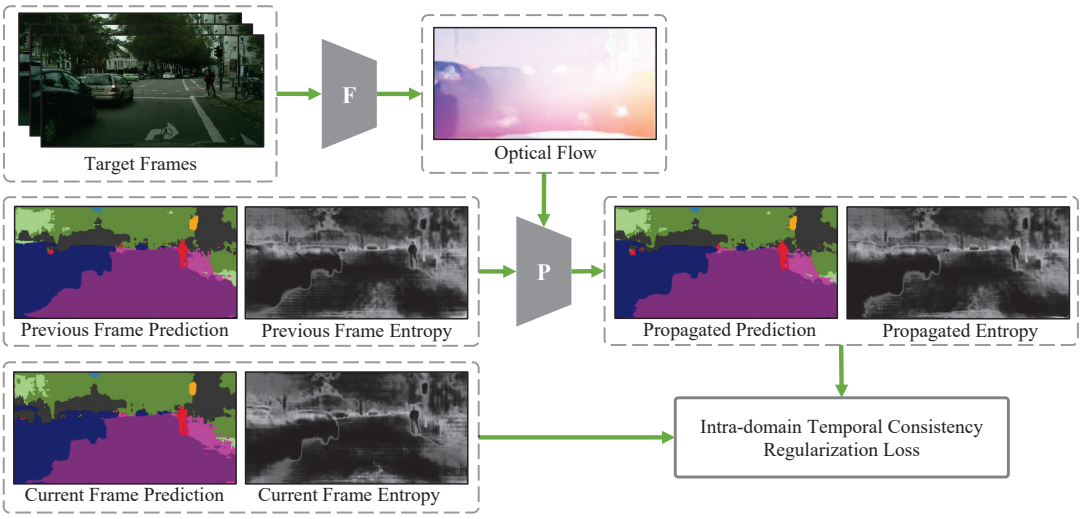


Figure 4. The framework of the proposed intra-domain temporal consistency regularization (I-TCR): I-TCR guides unconfident target-domain predictions to have similar temporal consistency as confident predictions, where the prediction confidence is measured by entropy. It consists of a network $F$ to estimate optical flow and a propagation operation $P$ to warp the previous frame prediction and its entropy based on the estimated optical flow. The objective of I-TCR is to force the current frame prediction with high entropy (*i.e.*, low confidence) to be consistent with the propagated prediction with low entropy (*i.e.*, high confidence). Note that only networks trained by the current frame requires gradient.

unconfident prediction $p_k^{\mathbb{T}}$ (*i.e.*, with high entropy). Given $p_k^{\mathbb{T}}$ and $\hat{p}_{k-1}^{\mathbb{T}}$ from target video frames $X^{\mathbb{T}}$, the I-TCR loss $\mathcal{L}_{itcr}$ can be formulated as follows:

$$\mathcal{L}_{itcr}(G) = S(E(p_k^{\mathbb{T}}) - E(\hat{p}_{k-1}^{\mathbb{T}}))|p_k^{\mathbb{T}} - \hat{p}_{k-1}^{\mathbb{T}}|. \quad (5)$$

where $S$ is a signum function which returns 1 if the input is positive or 0 otherwise.

DA-VSN jointly optimizes the source-domain supervised learning (*i.e.*, SSL) and the target-domain unsupervised learning (*i.e.*, C-TCR and I-TCR) as follows:

$$\min_{G} \max_{D_{st}, D_s} \mathcal{L}_{ssl}(G) + \lambda_u \mathcal{L}_{ctcr}(G, D_{st}, D_s) + \lambda_u \mathcal{L}_{itcr}(G), \quad (6)$$

where $\lambda_u$ is the weight for balancing the supervised and unsupervised learning in source and target domains.

# 4. Experiments

## 4.1. Experimental Setup

**Datasets:** Our experiments involve two challenging synthetic-to-real domain adaptive video semantic segmentation tasks: VIPER [62] $\rightarrow$ Cityscapes-Seq [14] and SYNTHIA-Seq [63] $\rightarrow$ Cityscapes-Seq. **Cityscapes-Seq** is a standard benchmark for supervised video semantic segmentation and we use it as the target-domain dataset. It contains $2,975$ and $500$ video sequences for training and evaluation, where each sequence consists of 30 realistic frames with one ground-truth label provided for the $20^{th}$ frame. **VIPER** is used as one source-domain dataset, which contains $133,670$ synthesized video frames with segmentation labels produced by game engines. **SYNTHIA-Seq** is used as the other source-domain dataset, which contains $8,000$ synthesized video frames with automatically generated segmentation annotations. The frame resolution is $1024 \times 2048$, $1080 \times 1920$ and $760 \times 1280$ in Cityscapes-Seq, VIPER and SYNTHIA-Seq, respectively.

**Implementation Details:** We adopt ACCEL [37] as the video semantic segmentation architecture. It consists of two segmentation branches, an optical flow network and a score fusion layer. Each segmentation branch generates single-frame prediction using Deeplab network [6] whose backbone is ResNet-101 [23] pre-trained on ImageNet [16]. The optical flow network propagates prediction in the previous frame via FlowNet [35] and the score fusion layer adaptively integrates predictions in previous and current frames using a $1 \times 1$ convolutional layer. All the discriminators in our experiments are designed as in DCGAN [61]. For the efficiency of training and inference, we apply bicubic interpolation to resize every video frame in Cityscapes-Seq and VIPER to $512 \times 1024$ and $720 \times 1280$, respectively. Our experiments are built on PyTorch [59] and the size of memory usage is below 12 GB. All the models are trained using SGD optimizer with a momentum of $0.9$ and a weight decay of $10^{-4}$. The learning rate is set at $10^{-4}$ and has a polynomial decay with a power of $0.9$. The balancing weights $\lambda_{sa}$, $\lambda_{wd}$, and $\lambda_u$ are set as 1, 1 and 0.001, respectively. The

| VIPER → Cityscapes-Seq | | | | | |
|---|---|---|---|---|---|
| Method | $\mathcal{L}_{ssl}$ | $\mathcal{L}_{sa}$ | $\mathcal{L}_{sta}$ | $\mathcal{L}_{wd}$ | mIoU |
| Source only | ✓ | | | | 37.1 |
| SA | ✓ | ✓ | | | 41.6 |
| STA | ✓ | | ✓ | | 43.7 |
| JT | ✓ | ✓ | ✓ | | 44.2 |
| **C-TCR** | ✓ | ✓ | ✓ | ✓ | **46.5** |

Table 1. Ablation study of C-TCR over domain adaptive segmentation task VIPER → Cityscapes-Seq: spatial alignment (SA) and spatial-temporal alignment (STA) both outperform 'Source only' greatly. Simple joint training (JT) with STA and SA yields marginal gains over STA, showing that additional spatial alignment does not help much. C-TCR outperforms JT clearly by introducing weight discrepancy loss $L_{wd}$ which forces STA to be independent from SA and focuses more on temporal alignment.

| VIPER → Cityscapes-Seq | | | | |
|---|---|---|---|---|
| Method | $\mathcal{L}_{ssl}$ | $\mathcal{L}_{ctcr}$ | $\mathcal{L}_{itcr}$ | mIoU |
| Source only | ✓ | | | 37.1 |
| C-TCR | ✓ | ✓ | | 46.5 |
| I-TCR | ✓ | | ✓ | 45.9 |
| **DA-VSN** | ✓ | ✓ | ✓ | **47.8** |

Table 2. Ablation study of DA-VSN over domain adaptive segmentation task VIPER → Cityscapes-Seq: Cross-domain TCR (C-TCR) and intra-domain TCR (I-TCR) both outperform 'Source only' by large margins. In addition, the combination of C-TCR and I-TCR in DA-VSN outperforms either C-TCR or I-TCR clearly, demonstrating the synergic relation of the two designs.

mean intersection-over-union (mIoU) is adopted to evaluate all methods.

### 4.2. Ablation Studies

We conduct comprehensive ablation studies to examine the effectiveness of our designs. and Tables 1 and 2 show experimental results. As shown in Table 1, both spatial alignment (SA) and spatial-temporal alignment (STA) outperform 'Source only' consistently, which verifies the effectiveness of the alignment in spatial and temporal spaces. Specifically, the performance gain of STA is larger than SA, which validates that temporal alignment is important in domain adaptive video segmentation by guiding the target predictions to have similar temporal consistency of source predictions. Joint training (JT) of STA and SA outperforms STA with a marginal performance gain, largely because the spatial-temporal alignment captures spatial alignment already. Cross-domain temporal consistency regularization (C-TCR) improves JT clearly by introducing weight discrepancy loss $L_{wd}$ between discriminators in STA and SA which forces STA to focus on alignment in the temporal space. It also validates the significance of temporal alignment in domain adaptive video semantic segmentation. Similar to C-TCR, intra-domain TCR (I-TCR) outperforms 'Source only' with a large margin as shown in Table 2. This shows the importance of intra-domain adaptation that suppresses temporal inconsistency across target-domain frames. Lastly, DA-VSN produces the best video segmentation, which demonstrates that C-TCR and I-TCR complement with each other.

### 4.3. Comparison with Baselines

Since few works study domain adaptive video semantic segmentation, we quantitatively compare DA-VSN with multiple domain adaptation baselines [69, 80, 58, 79, 21, 74] that achieved superior performance in domain adaptive image segmentation. We apply these approaches to the domain adaptive video segmentation task by simply replacing their image segmentation model by video segmentation model and performing domain alignment as in [69, 80, 58, 79, 21, 74]. The comparisons are performed over two synthetic-to-real domain adaptive video segmentation tasks as shown in Tables 3 and 4. As the two tables show, the proposed method outperforms all the domain adaptation baselines consistently with large margins.

We also perform qualitative comparisons over the video segmentation task VIPER → Cityscapes-Seq. We compare the proposed DA-VSN with the best-performing baseline FDA [74] as illustrated in Fig. 5. We can see that the qualitative results are consistent with the quantitative results in Table 3. Specifically, our method can generate better segmentation results with higher temporal consistency across consecutive video frames. The excellent segmentation performance is largely attributed to the proposed temporal consistency regularization which minimizes the divergence of temporal consistency across different domains and different target-domain video frames.

### 4.4. Discussion

**Feature Visualization:** In the Section 4.3, we have demonstrated that the proposed DA-VSN has achieved superior performance in domain adaptive video segmentation as compared with multiple baselines. To further study the properties of DA-VSN, we use t-SNE [51] to visualize the distribution of target-domain temporal feature representations from different domain adaptive video segmentation methods, where the inter-class and intra-class variances are computed for quantitative analysis. As shown in Fig. 6, DA-VSN produces the most discriminative target-domain temporal features with the largest inter-class variance and the smallest intra-class variance, as compared with 'Source only' and FDA [74].

**Complementary Studies:** We also investigate whether the proposed DA-VSN can complement with multiple domain adaptation baselines [80, 58, 79, 74] (as described in

| VIPER → Cityscapes-Seq | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | road | side. | buil. | fence | light | sign | vege. | terr. | sky | pers. | car | truck | bus | mot. | bike | 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 [69] | 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 [80] | 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 [58] | 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 [79] | 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 |
| SVMin [21] | 51.1 | 14.3 | 80.8 | 11.9 | **30.9** | 23.1 | 83.5 | **37.7** | 74.5 | 59.5 | **79.7** | **36.4** | **53.2** | 20.0 | 4.2 | 44.1 |
| CrCDA [32] | 78.1 | 33.3 | 82.2 | 21.3 | 29.1 | 26.8 | 82.9 | 28.5 | 80.7 | 59.0 | 73.8 | 16.5 | 41.4 | 7.8 | 2.5 | 44.3 |
| RDA [31] | 72.0 | 25.9 | 80.8 | 15.1 | 27.2 | 20.3 | 82.6 | 31.4 | 82.2 | 56.3 | 75.5 | 22.8 | 48.3 | 19.1 | 6.7 | 44.4 |
| FDA [74] | 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 (Ours)** | **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** |

Table 3. Quantitative comparisons of DA-VSN with multiple baselines over domain adaptive video segmentation task VIPER → Cityscapes-Seq: DA-VSN outperforms all domain adaptation baselines consistently by large margins.

| SYNTHIA-Seq → Cityscapes-Seq | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | road | side. | buil. | pole | light | sign | vege. | sky | pers. | 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 [69] | 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 [80] | 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 [58] | 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 [79] | 70.4 | 31.4 | **79.1** | 27.6 | 11.5 | 20.7 | 78.0 | 67.2 | **49.5** | 17.1 | 39.6 | 44.7 |
| SVMin [21] | 84.9 | 0.5 | 77.9 | **29.6** | 7.4 | 15.0 | **78.6** | 73.2 | 46.9 | 6.2 | 73.8 | 44.9 |
| CrCDA [32] | 86.5 | 26.3 | 74.8 | 24.5 | 5.0 | 15.5 | 63.5 | 64.4 | 46.0 | 15.8 | 72.8 | 45.0 |
| RDA [31] | 84.7 | 26.4 | 73.9 | 23.8 | 7.1 | 18.6 | 66.7 | 68.0 | 48.6 | 9.3 | 68.8 | 45.1 |
| FDA [74] | 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 (Ours)** | **89.4** | 31.0 | 77.4 | 26.1 | 9.1 | 20.4 | 75.4 | **74.6** | 42.9 | 16.1 | **82.4** | **49.5** |

Table 4. Quantitative comparisons of DA-VSN with multiple baselines over domain adaptive video segmentation task SYNTHIA-Seq → Cityscapes-Seq: DA-VSN outperforms all domain adaptation baselines consistently by large margins.

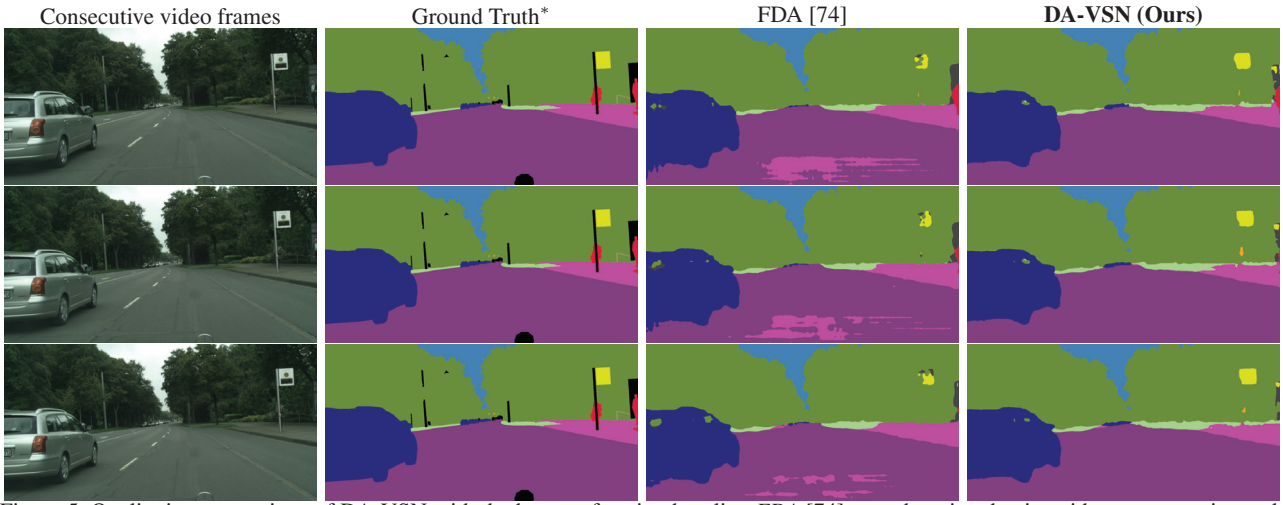


Figure 5. Qualitative comparison of DA-VSN with the best-performing baseline *FDA* [74] over domain adaptive video segmentation task "VIPER → Cityscapes-Seq": DA-VSN produces more accurate pixel-wise segmentation predictions with higher temporal consistency across consecutive video frames as shown in rows 1-3. Since Cityscapes-Seq only provides ground-truth label of one frame per 30 consecutive frames, we show the same ground truth for all rows. The *Ground Truth** denotes the ground truth annotated for the video frame in Row 2. Best viewed in color.

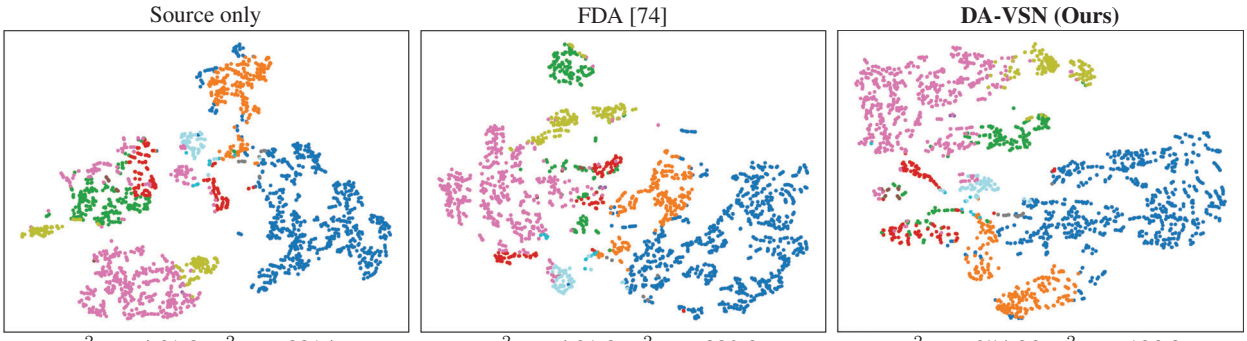


Figure 6. Visualization the distribution of temporal feature representations in the target domain via t-SNE [51]: We calculate inter-class variance $\sigma^2_{\text{inter}}$ and intra-class variance $\sigma^2_{\text{intra}}$ of temporal features, *i.e.*, the stacked feature maps from two consecutive frames. It can be observed that the proposed DA-VSN outperforms 'Source only' model and FDA [74] clearly. Evaluation is conducted on the domain adaptive video segmentation task "VIPER → Cityscapes-Seq". Note that different colors denote different classes and best viewed in color.

| **VIPER → Cityscapes-Seq** | | | |
|---|---|---|---|
| Method | Base | + DA-VSN | Gain |
| FDA [74] | 44.4 | 48.5 | +4.1 |
| IDA [58] | 42.0 | 49.9 | +7.9 |
| CBST [80] | 41.7 | 50.2 | +8.5 |
| CRST [79] | 42.4 | 51.3 | +8.9 |

Table 5. The proposed DA-VSN complements with multiple domain adaption baselines over domain adaptive video segmentation task VIPER → Cityscapes-Se: *DA-VSN* can be easily incorporated into state-of-the-art domain adaptive image segmentation methods [80, 79, 58, 74] with consistent performance improvement.

| **VIPER → Cityscapes-Seq** | | | |
|---|---|---|---|
| Architectures | Source only | DA-VSN | Gain |
| NetWarp [20] | 36.5 | 47.2 | +10.7 |
| TDNet [26] | 37.6 | 47.9 | +10.3 |
| ESVS [47] | 38.2 | 48.1 | +9.9 |

Table 6. DA-VSN can work with different video semantic segmentation architectures: *DA-VSN* can work with different video segmentation architectures (e.g. Netwarp [20], TDNet [26] and ESVS [47]) with consistent performance improvement as compared with *Source only* over the domain adaptive video segmentation task "VIPER → Cityscapes-Seq".

Section 4.3) over domain adaptive video segmentation task. To conduct this experiment, we integrate our proposed temporal consistency regularization components (DA-VSN) into these baselines and Table 5 shows the segmentation results of the newly trained models. It can be seen that the incorporation of DA-VSN improves video segmentation performance greatly across all the baselines, which shows that DA-VSN is complementary to the domain adaptation methods that minimize domain discrepancy via image translation (*e.g.*, FDA [74]), adversarial learning (*e.g.*, AdvEnt [69]) and self-training (*e.g.*, CBST [80] and CRST [79]).

**Different Video Segmentation Architectures:** We further study whether DA-VSN can work well with different video semantic segmentation architectures. Three widely adopated video segmentation architectures (*i.e.*, Netwarp [20], TDNet [26] and ESVS [47]) are used in this experiments. As shown in Table 6, the proposed DA-VSN outperforms the 'Source only' consistently with large margins. This experiment shows that our method performs excellently with different video semantic segmentation architectures that exploits temporal relations via feature propagation [20], attention propagation [26], and temporal consistency constraint [47].

## 5. Conclusion

This paper presents a domain adaptive video segmentation network that introduces cross-domain temporal consistency regularization (TCR) and intra-domain TCR to address domain shift in videos. Specifically, cross-domain TCR performs spatial and temporal alignment that guides the target video predictions to have similar temporal consistency as the source video predictions. Intra-domain TCR directly minimizes the discrepancy of temporal consistency across different target video frames. Extensive experiments demonstrate the superiority of our method in domain adaptive video segmentation. In the future, we will adapt the idea of temporal consistency regularization to other video domain adaptation tasks such as video instance segmentation and video panoptic segmentation.

**Acknowledgement.** This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from Singapore Telecommunications Limited (Singtel), through Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU).

## References


[1] Vijay Badrinarayanan, Fabio Galasso, and Roberto Cipolla. Label propagation in video sequences. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3265–3272. IEEE, 2010.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[3] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *European conference on computer vision*, pages 44–57. Springer, 2008.

[4] Ignas Budvytis, Patrick Sauer, Thomas Roddick, Kesar Breen, and Roberto Cipolla. Large scale labelled video data augmentation for semantic segmentation in driving scenarios. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 230–237, 2017.

[5] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *European Conference on Computer Vision*, pages 695–714. Springer, 2020.

[6] 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, 2017.

[7] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2090–2099, 2019.

[8] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6321–6330, 2019.

[9] Min-Hung Chen, Baopu Li, Yingze Bao, and Ghassan AlRegib. Action segmentation with mixed temporal domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 605–614, 2020.

[10] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020.

[11] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018.

[12] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840, 2019.

[13] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *European Conference on Computer Vision*, pages 678–695. Springer, 2020.

[14] 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.

[15] Camille Couprie, Clément Farabet, Yann LeCun, and Laurent Najman. Causal graph-based video segmentation. In *2013 IEEE International Conference on Image Processing*, pages 4249–4253. IEEE, 2013.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] Mingyu Ding, Zhe Wang, Bolei Zhou, Jianping Shi, Zhiwu Lu, and Ping Luo. Every frame counts: joint learning of video segmentation and optical flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10713–10720, 2020.

[18] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[19] Georgios Floros and Bastian Leibe. Joint 2d-3d temporally consistent semantic segmentation of street scenes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2823–2830. IEEE, 2012.

[20] Raghudeep Gadde, Varun Jampani, and Peter V Gehler. Semantic video cnns through representation warping. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4453–4462, 2017.

[21] Dayan Guan, Jiaxing Huang, Shijian Lu, and Aoran Xiao. Scale variance minimization for unsupervised domain adaptation in image segmentation. *Pattern Recognition*, 112:107764, 2021.

[22] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia*, 2021.

[23] 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.

[24] Daniel Hernandez-Juarez, Lukas Schneider, Antonio Espinosa, David Vázquez, Antonio M López, Uwe Franke, Marc Pollefeys, and Juan C Moure. Slanted stixels: Representing san francisco's steepest streets. *arXiv preprint arXiv:1707.05397*, 2017.

[25] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adap-

tation. In *International Conference on Machine Learning*, pages 1989–1998, 2018.

[26] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8818–8827, 2020.

[27] Jiaxing Huang, Dayan Guan, Shijian Lu, and Aoran Xiao. Mlan: Multi-level adversarial network for domain adaptive semantic segmentation. *arXiv preprint arXiv:2103.12991*, 2021.

[28] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Category contrast for unsupervised domain adaptation in visual tasks. *arXiv preprint arXiv:2106.02885*, 2021.

[29] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Cross-view regularization for domain adaptive panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10133–10144, 2021.

[30] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021.

[31] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. *arXiv preprint arXiv:2106.02874*, 2021.

[32] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *European conference on computer vision*, pages 705–722. Springer, 2020.

[33] Po-Yu Huang, Wan-Ting Hsu, Chun-Yueh Chiu, Ting-Fan Wu, and Min Sun. Efficient uncertainty estimation for semantic segmentation in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018.

[34] Junhwa Hur and Stefan Roth. Joint optical flow and temporally consistent semantic segmentation. In *European Conference on Computer Vision*, pages 163–177. Springer, 2016.

[35] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[36] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *Advances in Neural Information Processing Systems*, 2020.

[37] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8866–8875, 2019.

[38] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9859–9868, 2020.

[39] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. *arXiv preprint arXiv:2003.00867*, 2020.

[40] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. Feature space optimization for semantic video segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3168–3175, 2016.

[41] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2020.

[42] Yule Li, Jianping Shi, and Dahua Lin. Low-latency video semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5997–6005, 2018.

[43] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.

[44] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6758–6767, 2019.

[45] Buyu Liu and Xuming He. Multiclass semantic video segmentation with object-level active inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4286–4294, 2015.

[46] Si Liu, Changhu Wang, Ruihe Qian, Han Yu, Renda Bao, and Yao Sun. Surveillance video parsing with single frame supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–421, 2017.

[47] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *European Conference on Computer Vision*, pages 352–368. Springer, 2020.

[48] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6778–6787, 2019.

[49] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[50] Fengmao Lv, Tao Liang, Xiang Chen, and Guosheng Lin. Cross-domain semantic segmentation via domain-invariant interactive relation transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4334–4343, 2020.

[51] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[52] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *Computer Vision–ECCV 2020: 16th European*

*Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 415–430. Springer, 2020.

[53] Ondrej Miksik, Daniel Munoz, J Andrew Bagnell, and Martial Hebert. Efficient temporal consistency for streaming video scene analysis. pages 133–139. IEEE, 2013.

[54] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 122–132, 2020.

[55] Siva Karthik Mustikovela, Michael Ying Yang, and Carsten Rother. Can ground truth label propagation from video help semantic segmentation? In *European Conference on Computer Vision*, pages 804–820. Springer, 2016.

[56] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6819–6828, 2018.

[57] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11815–11822, 2020.

[58] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. *arXiv preprint arXiv:2004.07703*, 2020.

[59] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[60] Viorica Patraucean, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differentiable memory. *arXiv preprint arXiv:1511.06309*, 2015.

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

[62] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2213–2222, 2017.

[63] 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.

[64] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[65] Evan Shelhamer, Kate Rakelly, Judy Hoffman, and Trevor Darrell. Clockwork convnets for video semantic segmentation. In *European Conference on Computer Vision*, pages 852–868. Springer, 2016.

[66] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Weakly-supervised semantic segmentation using motion cues. In *European Conference on Computer Vision*, pages 388–404. Springer, 2016.

[67] 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.

[68] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1456–1465, 2019.

[69] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[70] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.

[71] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerior 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. *arXiv preprint arXiv:2003.08040*, 2020.

[72] Yu-Syuan Xu, Tsu-Jui Fu, Hsuan-Kung Yang, and Chun-Yi Lee. Dynamic video segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6556–6565, 2018.

[73] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Don Xie, Zongqiao Yu, Xiaowei Guo, Feiyue Huang, and Wen Gao. Part-aware progressive unsupervised domain adaptation for person re-identification. *IEEE Transactions on Multimedia*, 23:1681–1695, 2020.

[74] 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.

[75] Fangneng Zhan, Chuhui Xue, and Shijian Lu. Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9105–9115, 2019.

[76] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, pages 1–15, 2021.

[77] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2349–2358, 2017.

[78] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8865, 2019.

[79] 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.

[80] 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.