Exploit Domain-Robust Optical Flow in Domain Adaptive Video Semantic Segmentation

Yuan Gao[1], **Zilei Wang**[*1], **Jiafan Zhuang**[2], **Yixin Zhang**[1,3], **Junjie Li**[1]

[1] University of Science and Technology of China
[2] Shantou University
[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{gyy, zhyx12, hnljj}@mail.ustc.edu.cn, zlwang@ustc.edu.cn, jfzhuang@stu.edu.cn

Abstract

Domain adaptive semantic segmentation aims to exploit the pixel-level annotated samples on source domain to assist the segmentation of unlabeled samples on target domain. For such a task, the key is to construct reliable supervision signals on target domain. However, existing methods can only provide unreliable supervision signals constructed by segmentation model (SegNet) that are generally domain-sensitive. In this work, we try to find a domain-robust clue to construct more reliable supervision signals. Particularly, we experimentally observe the domain-robustness of optical flow in video tasks as it mainly represents the motion characteristics of scenes. However, optical flow cannot be directly used as supervision signals of semantic segmentation since both of them essentially represent different information. To tackle this issue, we first propose a novel *Segmentation-to-Flow Module* (SFM) that converts semantic segmentation maps to optical flows, named the segmentation-based flow (SF), and then propose a *Segmentation-based Flow Consistency* (SFC) method to impose consistency between SF and optical flow, which can implicitly supervise the training of segmentation model. The extensive experiments on two challenging benchmarks demonstrate the effectiveness of our method, and it outperforms previous state-of-the-art methods with considerable performance improvement. Our code is available at https://github.com/EdenHazardan/SFC.

Introduction

As a fundamental task in computer vision, semantic segmentation has achieved remarkable progress and brings out various applications, such as autonomous driving (Siam et al. 2018), robotics (Milioto and Stachniss 2019), and disease diagnosis (Sumithra, Suhil, and Guru 2015). The success is mostly driven by a large amount of labeled data (Cordts et al. 2016; Lin et al. 2014; Everingham et al. 2010), but the involved data labeling is laborious and expensive (Zhang and Wang 2020; Guan et al. 2021; Melas-Kyriazi and Manrai 2021), which hinders practical developments. An appealing approach to this issue is to use synthetic data that can be automatically generated and annotated by render engines (Richter et al. 2016; Richter, Hayder, and Koltun 2017; Ros et al. 2016). However, the model trained on synthetic data cannot generalize well to real-world data because of the domain shift between the source (*synthetic*) and target (*real-world*) domains (Mei et al. 2020; Liu and Wang 2022; Yang and Soatto 2020). Domain adaptive semantic segmentation (DASS) techniques are proposed to tackle such a domain shift problem.

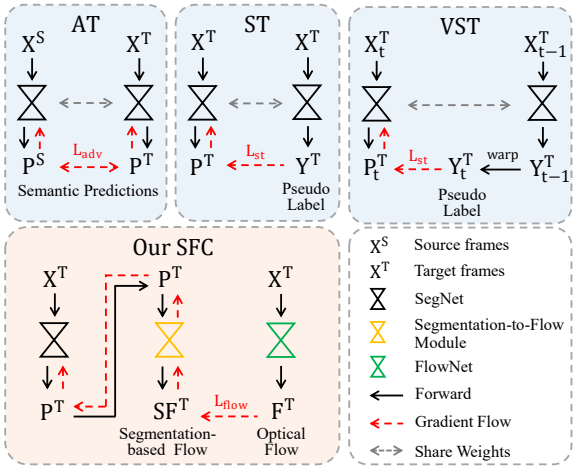


Figure 1: Illustration of supervision signals on target domain for different DASS methods. Different from existing methods, our SFC uses FlowNet (*i.e.*, optical flow) rather than SegNet to construct supervision signals. Best viewed in color.

DASS aims to adapt a model trained on source domain dataset with segmentation annotations to the unlabeled target domain. With the regular supervised learning on source domain, existing methods mainly focus on constructing effective segmentation supervision signals for target domain. In this way, existing methods can be roughly divided into two families, *i.e.*, adversarial training (AT) methods (Vu et al. 2019; Guan et al. 2021) and self-training (ST) methods (Melas-Kyriazi and Manrai 2021; Zhang et al. 2021). For video semantic segmentation, there are mainly video self-training (VST) methods (Xing et al. 2022; Guan et al. 2021). As shown in Figure 1, AT methods construct adversarial loss with segmentation predictions from both source and target domains, which guides the model to capture domain-invariant characteristics. Differently, ST methods

*Corresponding author

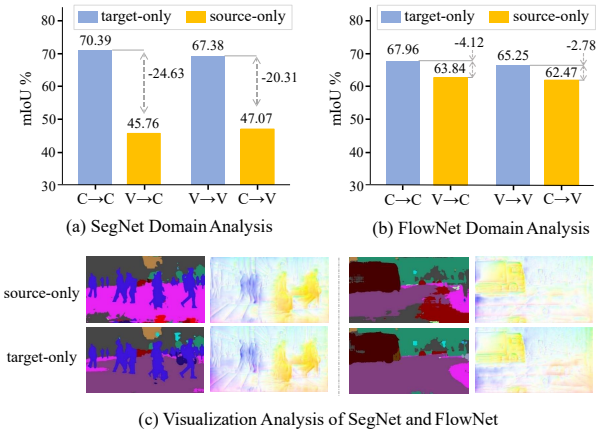


Figure 2: Experimental analysis of domain shift. Here Accel with SegNet and FlowNet is particularly adopted, and Cityscapes-Seq (C) and VIPER (V) datasets are used. For fair comparison, SegNet and FlowNet are conducted with the same initialization throughout all settings. Ablation study is performed by training one component while keeping the other fixed. We can observe the performance drop from *target-only learning* (blue) to *source-only learning* (yellow). In particular, (a) SegNet suffers from a severe performance drop; (b) FlowNet nearly maintains its performance. (c) We visualize the predicted segmentation maps and optical flows for two test samples on Cityscapes-Seq. Obviously, segmentation maps become worse due to domain shift while optical flows are almost remained. Best viewed in color.

provide direct semantic supervision with generated pseudo labels by SegNet. More recently, DA-VSN (Guan et al. 2021) and TPS (Xing et al. 2022) extend ST methods to video semantic segmentation, which utilizes optical flow to warp the pseudo labels from previous to current frames.

Although impressive performance has been achieved, existing methods suffer from unreliable supervision signals on target domain as SegNet is essentially domain-sensitive. Specifically, SegNet inevitably contains source domain-specific characteristics due to performing supervised learning on source domain. For AT methods, such domain-specific characteristics make it difficult for the model to extract discriminative features on target domain. For ST methods, it is hard to generate accurate pseudo labels on target domain without high-quality target features. In a word, the constructed supervision signals are always inaccurate for target domain due to domain sensitiveness of SegNet. Recently, some works have attempted to alleviate this issue. For example, GVB (Cui et al. 2020) proposes a gradually vanishing bridge mechanism to reduce domain-specific characteristics in adversarial learning. Some works are proposed to rectify noisy pseudo labels by adopting meta-learning (Guo et al. 2021) or exploiting confidence score (Mei et al. 2020) and uncertainty (Pan et al. 2020). However, these works can only alleviate the issue of unreliable supervision signals from domain-sensitive SegNet.

Unlike existing methods, we try to find a domain-robust clue to construct reliable supervision signals. Intuitively, al-

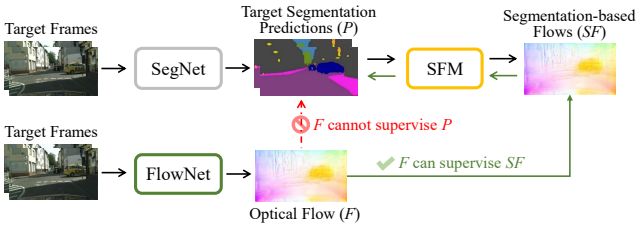


Figure 3: Illustration of our proposed Segmentation-based Flow Consistency (SFC).

though the render engines are difficult to generate realistic images (*e.g.*, layouts, colors, and illumination conditions) (Cheng et al. 2021), they can synthesize physically reasonable videos (Richter, Hayder, and Koltun 2017; Ros et al. 2016). That is, the motion patterns in synthesis videos and real-world scenarios are almost consistent. In video tasks, motion patterns are generally modeled by optical flow with pixel-wise displacements between two consecutive frames. Thus we consider that optical flow is domain-robust in this work. Here we particularly verify it in domain adaptive video semantic segmentation (DAVSS) task through some analysis experiments. Specifically, we adopt a popular video semantic segmentation (VSS) model (*i.e.*, Accel (Jain, Wang, and Gonzalez 2019)), which contains a SegNet for segmentation predictions and a FlowNet for optical flow estimation. To reveal the influence of domain shift, we investigate the performance drop from *target-only learning* to *source-only learning*. As shown in Figure 2(a) and 2(b), we can observe that SegNet suffers from a severe performance drop while FlowNet nearly maintains its performance. Besides, we visualize the results of predicted segmentation maps and optical flows in Figure 2(c). We can see that optical flows are almost the same for different settings while segmentation predictions change greatly. Then a natural question arises: *can we exploit the domain-robust optical flow to construct more reliable supervision signals?*

Existing DAVSS methods (Guan et al. 2021; Xing et al. 2022) use optical flow to warp the pseudo labels from previous to current frames, and such cross-frame pseudo label supervision can suppress the temporal inconsistency across different frames. However, they do not utilize the domain-robustness of optical flow to eliminate domain shift, and the supervision signals (*i.e.*, pseudo labels) are still constructed by domain-sensitive SegNet. In this work, we propose to use optical flow to construct reliable supervision signals for segmentation. But optical flow cannot be directly used to supervise semantic segmentation since they are different information. To tackle this issue, we first propose a novel Segmentation-to-Flow Module (SFM) to convert segmentation maps to optical flows, as shown in Figure 3, which is denoted by Segmentation-based Flow (SF). Here the quality of SF highly depends on the accuracy of input semantic segmentation. On this basis, we propose a Segmentation-based Flow Consistency (SFC) method to supervise the segmentation on target domain, where the consistency constraint between SF and optical flow is imposed. Compared to previous AT and ST methods that generate adversarial signals

or pseudo labels by domain-sensitive SegNet, our SFC can provide more reliable supervision signals by skillfully exploiting domain-robust optical flows.

We summarize the contributions of this work as follows:

- We propose to construct more reliable supervision signals by domain-robust clue, which can better eliminate domain shift in domain adaptation tasks. In particular, we discover the domain-robustness of optical flow and further exploit it to construct supervision signals in DAVSS.
- We first propose a *Segmentation-to-Flow Module* (SFM) to convert segmentation maps to optical flows, and then propose a novel *Segmentation-based Flow Consistency* (SFC) method to impose consistency constraints between SF and optical flow, which implicitly supervises semantic segmentation.
- We experimentally evaluate the effectiveness of our proposed method, and the results on two challenging benchmarks demonstrate the superiority of our method to previous state-of-the-art methods.

Related Work

Domain Adaptive Image Semantic Segmentation

Domain adaptive image semantic segmentation (DAISS) has been widely investigated to address the pixel-level dense annotation challenge and domain shift issues (Vu et al. 2019; Melas-Kyriazi and Manrai 2021). Most existing methods can be mainly divided into adversarial training (AT) and self-training (ST). AT methods focus on learning domain-invariant characteristics by adopting adversarial training (Goodfellow et al. 2014) at the image space (Zhang et al. 2020), intermediate feature space (Wan et al. 2020), and output space (Vu et al. 2019; Kim and Byun 2020). ST methods iteratively train SegNet on source domain and generate pseudo labels on target domain for further training. However, pseudo labels are usually noisy because of domain shift (Zhang et al. 2021; Zheng and Yang 2021). Recently, some works are proposed to improve pseudo labels, *e.g.*, using meta-learning (Guo et al. 2021), filtering out noisy samples based on confidence score (Li et al. 2022; Mei et al. 2020) and uncertainty (Pan et al. 2020). Besides, PCL (Li et al. 2021) proposes a novel probability contrastive loss which greatly simplifies the process of pseudo-labeling and achieves good results. In a word, existing methods mainly construct supervision on target domain by SegNet, which is domain-sensitive and unreliable. Differently, in this work, we propose to exploit domain-robust optical flow for constructing reliable supervision signals.

Domain Adaptive Video Semantic Segmentation

Video semantic segmentation aims to predict pixel-level segmentation for each video frame. Existing works usually exploit inter-frame temporal relations for accurate and efficient segmentation. For example, DFF (Zhu et al. 2017) and DAVSS (Zhuang, Wang, and Wang 2020) propose feature propagation to reuse key frame features under the guidance of estimated optical flows to reduce computational cost. Accel (Jain, Wang, and Gonzalez 2019) presents an adaptive fusion policy to effectively integrate predictions from different frames. However, existing methods still require sufficient pixel-level annotations, which is expensive and time-consuming. To address this issue, DA-VSN (Guan et al. 2021) first proposes the domain adaptive video semantic segmentation (DAVSS) task. Inspired by DAISS methods, DA-VSN extends ADVENT (Vu et al. 2019) to DAVSS with both spatial and temporal adversarial learning. Besides, DA-VSN further proposes a temporal consistency regularization that uses temporal pseudo labels as supervision signals. TPS (Xing et al. 2022) abandons unstable adversarial learning and extends pixmatch (Melas-Kyriazi and Manrai 2021) to DAVSS with cross-frame augmentation and cross-frame pseudo labeling.

DA-VSN and TPS essentially belong to AT and ST methods, and thus they still suffer from the similar problem on inaccurate supervision signals.

Domain Adaptive Video Classification

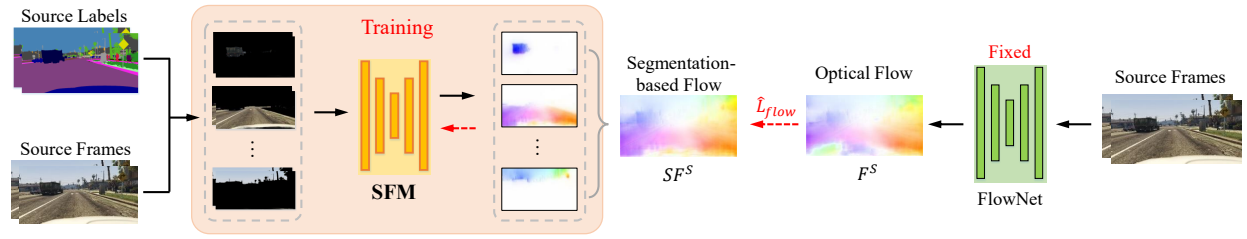
Another related task to DAVSS is domain adaptive video classification, which mainly investigates domain discrepancy in action recognition. Existing works (Jamal et al. 2018; Chen et al. 2019, 2020) mainly leverage the insights from domain adaptive image classification and extend them to video tasks. Recently, Munro *et al*. (Munro and Damen 2020) found that optical flow is environmentally robust through t-SNE visualization. Based on the observation, several works (Han, Xie, and Zisserman 2020; Kim et al. 2021) propose a cross-modal contrastive learning method to mitigate modal discrepancy between RGB and flow features, which can achieve better performance with two-stream feature fusion.

These methods also exploit optical flow, but they focus on improving the discriminativeness of features by integrating optical flow information, which essentially belongs to feature fusion. Differently, this work aims at constructing domain-robust supervision signals to optimize video semantic segmentation model on target domain.

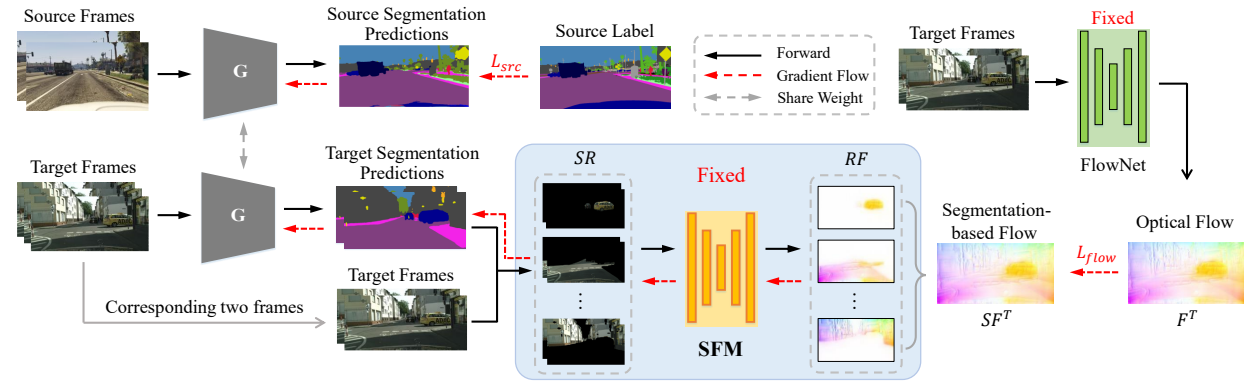
Our Method

In this work, we focus on the domain adaptive video semantic segmentation (DAVSS) task. Formally, given video frames X_S with the corresponding segmentation labels Y_S on source domain, we aim to train a video semantic segmentation (VSS) model G that can produce accurate segmentation predictions P_T on target domain. However, due to domain shift, G cannot generalize well to target domain. To tackle this issue, we propose a novel Segmentation-based Flow Consistency (SFC) method to provide reliable supervision signals on target domain. Similar to previous methods (Zhang et al. 2021; Melas-Kyriazi and Manrai 2021; Guan et al. 2021; Xing et al. 2022), the learning objective can be generally summarized as two parts. The first part is a regular cross-entropy loss on source domain:

$$L_{src} = -\mathbb{E}_{y_s \in Y_S} \sum_{i=1}^{H \times W} \sum_{c=1}^{C} y_s^{(i,c)} \log p_s^{(i,c)}, \qquad (1)$$



(a) Pretraining Segmentation-to-Flow Module (SFM) on Source Domain



(b) The overview of the proposed Segmentation-based Flow Consistency (SFC)

Figure 4: (a) Illustration of Segmentation-to-Flow Module (SFM) pre-training. (b) Overview of our proposed Segmentation-based Flow Consistency (SFC) method. Here we use G to represent the VSS model. SR and RF means segmentation-based region and corresponding predicted region flow, respectively.

where $p_s = G(x_s)$, x_s are consecutive frames from X_S with spatial resolution of $H \times W$ and ${p_s}^{(i,c)}$ represents the softmax probability of the pixel i belonging to the cth class. The second part aims to supervise model training on target domain by exploiting domain-robust optical flows, which is provided by our proposed SFC and is denoted by L_{flow}. Then the overall learning objective can be presented as follows

$$L = L_{src} + \lambda_f L_{flow}, \tag{2}$$

where λ_f is a trade-off parameter.

Our proposed SFC aims to provide reliable supervision signals on target domain by exploiting domain-robust optical flows. As illustrated in Figure 4, we first introduce the core component of SFC, *i.e.*, Segmentation-to-Flow Module (SFM), and then elaborate on the design of SFC on target domain. Finally, we will explain the details of training and inference procedures.

Segmentation-to-Flow Module

Design of SFM In order to exploit optical flow to supervise the segmentation on target domain, we need to convert segmentation map to optical flow. However, we cannot only use segmentation maps to predict optical flow because the segmentation maps do not contain appearance information (*e.g.*, illumination, texture), which is necessary for flow estimation. In this work, we design a novel Segmentation-to-Flow Module (SFM), which takes the same network architecture as FlowNet but uses both segmentation and video frames as input. As shown in the blue box in Figure 4(b), SFM takes a series of segmentation-based region pairs (SR) that are modulated from video frames according to segmentation results (*i.e.*, split regions by category):

$$SR_t^c = X_t P_t^c, \tag{3}$$

where SR_t^c represents a segmentation-based region belonging to class $c \in [1, C]$ at timestamp t and C is the number of classes, $X_t \in R^{H \times W}$ represents the frame image with spatial resolution of $H \times W$, and $P_t^c \in R^{H \times W}$ represents the corresponding cth class softmax segmentation prediction from the VSS model G. It can be seen that more accurate segmentation results can get the more correct SR for each class. Then SFM utilizes SR to predict the corresponding region flow (RF) for each class as

$$RF_t^c = SFM(SR_{t-1}^c, SR_t^c). \tag{4}$$

Obviously, only correct SR can result in high-quality RF. After calculating region flows of all classes, SFM generates a complete segmentation-based flow (SF) with a simple fusion:

$$SF_t = \sum_{c=1}^{C} RF_t^c. \tag{5}$$

Through this way, SFM can convert segmentation maps to optical flows.

Pretraining of SFM SFM utilizes segmentation predictions to predict segmentation-based flow (SF). Thus we can

use the domain-robust optical flow from FlowNet to supervise SF, which would indirectly supervise the segmentation predictions since the quality of SF highly depends on the accuracy of input semantic segmentation. To this end, we propose to pretrain SFM on source domain by exploiting the provided segmentation labels. Specifically, as shown in Figure 4(a), we feed a pair of video frames and corresponding segmentation labels to SFM, and impose a consistency constraint between predicted SF^S and optical flow F^S from FlowNet. Particularly, we adopt the endpoint error loss (EPE) (Dosovitskiy et al. 2015), which is the Euclidean distance averaged over all pixels and commonly used in optical flow estimation (Dosovitskiy et al. 2015; Ilg et al. 2017; Sun et al. 2018),i.e.,

$$\hat{L}_{flow} = \frac{1}{N}\sum_{i}^{N}\sqrt{(SF_u^S(i) - F_u^S(i))^2 + (SF_v^S(i) - F_v^S(i))^2}, \tag{6}$$

where i represents a pixel location and N is the total pixel number. u and v represent the horizontal and vertical components of optical flow, respectively.

Segmentation-based Flow Consistency

SFC supervision signals The pretrained SFM builds a relationship between segmentation maps and segmentation-based flow (SF). Then we can exploit the domain-robust optical flow to supervise the SF from SFM on target domain, which would indirectly supervise the segmentation predictions. Figure 4(b) shows the framework of SFC. For the target domain, we feed SFM the segmentation predictions from VSS model G instead of segmentation labels in pretraining on source domain. Here we adjust the softmax operation in G by multiplying a scale factor λ_s, which can make the predictions more consistent with the one-hot labels used in SFM pretraining. Apparently, due to suffering from domain shift, the VSS model G cannot provide accurate segmentation predictions (*i.e.*, noisy SR input for SFM) and further results in inaccurate SF^T, which would lead to the inconsistency of SF^T and F^T from FlowNet. Then we can impose a novel Segmentation-based Flow Consistency (SFC) constraint between SF^T and F^T to optimize segmentation predictions on target domain.

$$L_{flow} = \frac{1}{N}\sum_{i}^{N}\sqrt{(SF_u^T(i) - F_u^T(i))^2 + (SF_v^T(i) - F_v^T(i))^2}. \tag{7}$$

We keep the pretrained SFM fixed during training, and the gradient flow would pass through SFM and directly act on G. Intuitively, the VSS model G would improve its segmentation predictions on target domain as the consistency constraint is gradually achieved.

FlowNet pretraining In both the SFM pretraining and SFC supervision procedures, we need domain-robust optical flows from FlowNet on source and target domains, *i.e.*, F^S and F^T, to construct EPE loss. Here we explain the pretraining of FlowNet. As a common practice in VSS methods (Jain, Wang, and Gonzalez 2019; Zhuang, Wang, and Wang 2020), SegNet and FlowNet are jointly trained with a feature propagation paradigm on a labeled video semantic segmentation dataset. Therefore, we perform the FlowNet pretraining on source domain because of its rich annotations. Benefiting from the domain-robustness of optical flow, the source pretrained FlowNet can also predict reliable optical flow on target domain.

Training and Inference

Here we elaborate on the details of the training and inference procedures. Firstly, following common practices of VSS methods (Jain, Wang, and Gonzalez 2019; Zhu et al. 2017), we train a VSS model on source domain and obtain a pretrained FlowNet. Secondly, we pretrain SFM on source domain. Finally, as shown in Figure 4(b), we construct SFC supervision signals to train a VSS model from scratch with the pretrained FlowNet and SFM. During inference, only the well-trained VSS model is used for video segmentation prediction, and SFM would be discarded, which thus would not introduce extra computational cost.

Experiments

Experimental Setup

Datasets Following DA-VSN (Guan et al. 2021) and TPS (Xing et al. 2022), our experiments involve two challenging synthetic-to-real benchmarks: VIPER $\rightarrow$ Cityscapes-Seq and SYNTHIA-Seq $\rightarrow$ Cityscapes-Seq. **Cityscapes-Seq** (Cordts et al. 2016) is a representative dataset in semantic segmentation and autonomous driving domain. We use it as the target domain dataset without using any annotations during training. The training and validation subsets contain $2,975$ and 500 videos, respectively, and each video contains 30 frames at a resolution of 1024×2048. **VIPER** (Richter, Hayder, and Koltun 2017) is a synthetic dataset consisting of $133,670$ synthesized video frames with segmentation labels generated by game engines, which is used as a source domain dataset. The frame resolution of VIPER is 1080×1920. **SYNTHIA-Seq** (Ros et al. 2016) is also a synthetic dataset consisting of $8,000$ synthesized video frames with automatically generated segmentation annotations, which is used as another source domain dataset. The frame resolution is 760×1280. For the efficiency of training and inference, we apply bicubic interpolation to resize every video frame in Cityscapes-Seq and VIPER to 512×1024 and 720×1280, respectively.

Implementation details As in DA-VSN (Guan et al. 2021) and TPS (Xing et al. 2022), we adopt Accel (Jain, Wang, and Gonzalez 2019) throughout experiments. It consists of two segmentation branches, an optical flow network, and a score fusion layer. Two segmentation branches are used to generate semantic predictions on consecutive frames using Deeplab (Chen et al. 2017), whose backbones are both ResNet-101 (He et al. 2016) pretrained on ImageNet (Deng et al. 2009). FlowNet (Dosovitskiy et al. 2015) is adopted as an optical flow network to propagate prediction from the previous frame, which is pretrained on Flying Chairs dataset (Dosovitskiy et al. 2015). The score fusion layer adaptively integrates predictions from consecutive frames using a 1×1 convolutional layer. Following the official implementation, Accel (Jain, Wang, and Gonzalez 2019) uses a two-stage training procedure. In stage one, two segmentation branches (*i.e.*, SegNet) are pretrained on a segmentation

VIPER → Cityscapes-Seq																	
Methods	Venue	road	side.	buil.	fence	light	sign	vege.	terr.	sky	pers.	car	truck	bus	mot.	bike	mIoU
Source only	–	60.4	19.9	79.2	9.7	22.4	20.4	79.0	12.6	82.2	54.4	67.3	5.4	18.6	17.0	12.3	37.4
AdvEnt	CVPR 2019	78.2	32.8	80.3	19.0	25.6	22.3	80.1	17.7	83.4	56.1	66.6	9.2	36.2	6.9	6.3	41.4
PixMatch	CVPR 2021	87.5	30.7	84.7	5.7	22.5	**29.7**	85.5	37.4	83.3	58.9	79.2	29.5	47.3	20.1	8.6	47.4
DA-VSN*(TPL)	ICCV 2021	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
DA-VSN (TPL)	ICCV 2021	88.1	38.1	**85.8**	13.2	33.5	29.5	**85.9**	25.8	82.1	59.2	82.4	17.6	50.5	18.5	10.9	48.1
TPS*(TPL)	ECCV 2022	82.4	36.9	79.5	9.0	26.3	29.4	78.5	28.2	81.8	61.2	80.2	**39.8**	40.3	**28.5**	31.7	48.9
TPS (TPL)	ECCV 2022	88.0	28.5	84.6	3.0	33.5	27.0	85.8	**39.1**	85.3	60.4	81.3	33.9	**52.5**	22.5	10.1	49.0
Ours	–	**89.9**	40.8	83.8	6.8	34.4	25.0	85.1	34.3	84.1	**62.6**	82.1	35.3	47.1	23.2	31.3	51.1
Ours (TPL)	–	**89.9**	**41.5**	84.0	7.0	**36.5**	27.1	85.6	33.7	**86.6**	62.4	**82.6**	36.3	47.6	23.2	**31.9**	**51.7**

Table 1: Results on VIPER → Cityscapes-Seq benchmark. Here * means the result from the official paper, and TPL means using temporal pseudo label technique. Our SFC outperforms other domain adaptation semantic segmentation methods by a large margin. Moreover, as our SFC supervises the VSS model by domain-robust optical flow, the VSS model can provide more accurate pseudo labels for extra supervision on target domain, which can further improve the performance.

SYNTHIA-Seq → Cityscapes-Seq													
Methods	Venue	road	side.	buil.	pole	light	sign	vege.	sky	pers.	rider	car	mIoU
Source only	–	60.9	29.9	74.7	24.8	6.1	21.6	69.5	52.2	39.4	13.7	34.3	38.8
AdvEnt	CVPR 2019	80.5	22.9	68.6	20.9	7.8	18.8	67.0	65.9	43.2	13.4	62.7	42.9
PixMatch	CVPR 2021	88.1	17.1	**80.7**	24.6	9.7	32.0	80.1	**81.2**	52.5	14.2	83.8	51.3
DA-VSN*(TPL)	ICCV 2021	89.4	31.0	77.4	26.1	9.1	20.4	75.4	74.6	42.9	16.1	82.4	49.5
DA-VSN (TPL)	ICCV 2021	87.5	9.7	80.6	21.7	8.7	32.1	79.7	81.0	51.8	14.4	82.5	50.0
TPS*(TPL)	ECCV 2022	**91.2**	**53.7**	74.9	24.6	**17.9**	**39.3**	68.1	59.7	57.2	**20.3**	84.5	53.8
TPS (TPL)	ECCV 2022	89.8	36.0	79.9	27.8	12.9	31.9	80.3	80.4	54.9	17.2	83.1	54.0
Ours	–	90.9	32.5	76.8	28.6	6.0	36.7	76.0	78.9	51.7	13.8	85.6	52.5
Ours (TPL)	–	90.0	32.8	80.4	**28.9**	14.9	35.3	**80.8**	81.1	**57.5**	19.6	**86.7**	**55.3**

Table 2: Results on SYNTHIA → Cityscapes-Seq benchmark. Our method outperforms other domain adaptation semantic segmentation methods by a large margin.

dataset, which uses an SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The learning rate is set at 2.5×10^{-4} for backbone parameters and 2.5×10^{-3} for others, which is annealed following the poly learning rate policy. In stage two, the backbones of two SegNets are fixed while other components (*i.e.*, two classifiers, FlowNet and the score fusion layer) are jointly trained, which uses an SGD optimizer with a momentum of 0.9, a weight decay of 5×10^{-4} and a learning rate of 5×10^{-4}.

We introduce domain adaptive techniques in both training stages. For our method, we train SFM with the same training hyperparameters as FlowNet (Dosovitskiy et al. 2015). Besides, there is a trade-off hyperparameter λ_f in Eq. 2 and a scale factor λ_s in constructing SFC. We set λ_f as 0.005 and 0.001 in two training stages respectively, while set $\lambda_s = 100$ in both stages.

Performance Comparison

To demonstrate the superiority of our method, we make a comparison with current state-of-the-art methods, including the latest DAVSS methods, *i.e.*, DA-VSN (Guan et al. 2021) and TPS (Xing et al. 2022), and several DAISS methods. Here we particularly select a representative AT method, *i.e.*, ADVENT (Vu et al. 2019), and a representative ST method, *i.e.*, PixMatch (Melas-Kyriazi and Manrai 2021). For fair comparison, we not only show the results of DA-VSN and TPS taken from their papers but also show our reproduced results in our training manner. Besides, since DAISS methods are not designed for VSS models, we follow their official implementations and re-implement them in both training stages of Accel. The results on two synthetic-to-real benchmarks are shown in Table 1 and Table 2. On VIPER → Cityscapes-Seq, our SFC surpasses DA-VSN (3% in mIoU) and TPS (2.1% in mIoU) by a large margin. Considering that both DA-VSN and TPS use temporal pseudo labeling

Stage One				
Methods	Source-only	DA-VSN	TPS	Ours (SFC)
mIoU	37.62	45.12	46.71	**48.87**

Stage Two				
Methods	Source-only	DA-VSN	TPS	Ours (SFC)
mIoU	37.39	44.93	46.58	**48.02**

Table 3: Ablation study on improvement in two-stage. In stage two, we use the same SegNet model from source-only (37.62% mIoU) baseline for fair comparison.

for domain adaptation, our SFC can also be combined with this technique. Unlike DA-VSN and TPS whose pseudo labels suffer from domain-sensitive SegNet of VSS model, our method can provide more accurate pseudo labels since our SFC optimizes SegNet by exploiting the domain-robust optical flow. Thus our method can further improve the performance with extra pseudo label supervision, *i.e.*, when combined with temporal pseudo label, our method outperforms TPS 2.7% in mIoU. On SYNTHIA-Seq $\rightarrow$ Cityscapes-Seq, our SFC surpasses DA-VSN (2.5% in mIoU) by a large margin. When combined with temporal pseudo label, our method outperforms TPS 1.3% in mIoU.

Ablation Study

In this section, we conduct experiments to reveal the effectiveness of our proposed method. All experiments are conducted on the VIPER $\rightarrow$ Cityscapes-Seq benchmark.

Improvement in two stages Accel is trained in a two-stage manner, and we further investigate the effectiveness of our SFC in different training stages. As shown in Table 3, our SFC achieves better performance than other DAVSS methods in both stages.

Effect of SFC supervision Our SFC imposes a consistency constraint between SF and F with EPE loss. Figure 5 shows the relationship between segmentation predictions and segmentation-based flow. It can be seen that the

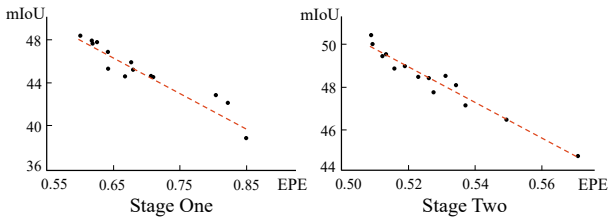


Figure 5: Analysis of relationship between segmentation predictions and segmentation-based flow (SF). Here we provide mIoU score of segmentation predictions and EPE loss between SF and the optical flow (F) from FlowNet at different training periods. Notably, a lower EPE score means the better consistency SF and F.

Methods	DFF	DFF + Ours
mIoU	37.04	$48.51_{\uparrow 11.47}$

Methods	DA-VSS	DA-VSS + Ours
mIoU	38.70	$48.97_{\uparrow 10.27}$

Table 4: Performance on different VSS methods

Network \ Methods	Source-only	SFC
FlowNet	37.39	$51.05_{\uparrow 13.66}$
RAFT	38.64	$51.57_{\uparrow 12.93}$

Table 5: Performance with different optical flow networks.

VSS model G gradually improves the segmentation predictions (*i.e.*, the higher mIoU score) on target domain as the SFC constraint is gradually achieved (*i.e.*, the lower EPE loss). That is, the supervision on SF can correctly act on semantic segmentation, which verifies the rationality of our proposed SFC.

Performance with different VSS methods To study the generalization ability of our method, we further apply it to different video semantic segmentation methods. Particularly, we adopt two popular video segmentation methods, *i.e.*, DFF (Zhu et al. 2017) and DA-VSS (Zhuang, Wang, and Wang 2020). As shown in Table 4, our proposed method can bring significant performance improvement consistently.

Performance with different FlowNet To verify the generalization ability of our SFC (*i.e.*, exploiting domain-robust optical flow for domain adaptation), we further try other network for flow estimation, *i.e.*, using RAFT (Teed and Deng 2020) to replace the FlowNet (Dosovitskiy et al. 2015) in VSS model and the network architecture of SFM. As shown in Table 5, our SFC can achieve consistent improvement with RAFT, verifying it generalization ability on FlowNet.

Conclusion

In this paper, we focus on the domain adaptive video semantic segmentation task. Different from existing works, we propose to exploit a domain-robust clue for domain adaptation and further verify the domain-robustness of optical flow. Inspired by this, we propose a *Segmentation-to-Flow Module* (SFM) that converts segmentation maps to optical flows and further propose a novel *Segmentation-based Flow Consistency* (SFC) method to supervise semantic segmentation on target domain by exploiting optical flows, which provides more robust and reliable supervision signals. Extensive experiments on two challenging benchmarks validate the effectiveness of our method, which outperforms previous state-of-the-art methods. We believe our proposed method provides a new route for domain adaptive video semantic segmentation.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62176246 and Grant 61836008. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*.

Chen, M.-H.; Kira, Z.; AlRegib, G.; Yoo, J.; Chen, R.; and Zheng, J. 2019. Temporal attentive alignment for large-scale video domain adaptation. In *ICCV*.

Chen, M.-H.; Li, B.; Bao, Y.; AlRegib, G.; and Kira, Z. 2020. Action segmentation with joint self-supervised temporal domain adaptation. In *CVPR*.

Cheng, Y.; Wei, F.; Bao, J.; Chen, D.; Wen, F.; and Zhang, W. 2021. Dual path learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9082–9091.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

Cui, S.; Wang, S.; Zhuo, J.; Su, C.; Huang, Q.; and Tian, Q. 2020. Gradually vanishing bridge for adversarial domain adaptation. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; and Brox, T. 2015. Flownet: Learning optical flow with convolutional networks. In *ICCV*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *IJCV*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *NeurIPS*.

Guan, D.; Huang, J.; Xiao, A.; and Lu, S. 2021. Domain adaptive video segmentation via temporal consistency regularization. In *ICCV*.

Guo, X.; Yang, C.; Li, B.; and Yuan, Y. 2021. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *CVPR*.

Han, T.; Xie, W.; and Zisserman, A. 2020. Self-supervised co-training for video representation learning. In *NeurIPS*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; and Brox, T. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*.

Jain, S.; Wang, X.; and Gonzalez, J. E. 2019. Accel: A corrective fusion network for efficient semantic segmentation on video. In *CVPR*.

Jamal, A.; Namboodiri, V. P.; Deodhare, D.; and Venkatesh, K. 2018. Deep Domain Adaptation in Action Space. In *BMVC*.

Kim, D.; Tsai, Y.-H.; Zhuang, B.; Yu, X.; Sclaroff, S.; Saenko, K.; and Chandraker, M. 2021. Learning cross-modal contrastive features for video domain adaptation. In *ICCV*.

Kim, M.; and Byun, H. 2020. Learning texture invariant representation for domain adaptation of semantic segmentation. In *CVPR*.

Li, J.; Wang, Z.; Gao, Y.; and Hu, X. 2022. Exploring High-quality Target Domain Information for Unsupervised Domain Adaptive Semantic Segmentation. In *ACM MM*.

Li, J.; Zhang, Y.; Wang, Z.; and Tu, K. 2021. Semantic-aware Representation Learning Via Probability Contrastive Loss. *arXiv preprint arXiv:2111.06021*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, Q.; and Wang, Z. 2022. Collaborating Domain-Shared and Target-Specific Feature Clustering for Cross-domain 3D Action Recognition. In *ECCV*.

Mei, K.; Zhu, C.; Zou, J.; and Zhang, S. 2020. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*.

Melas-Kyriazi, L.; and Manrai, A. K. 2021. PixMatch: Unsupervised domain adaptation via pixelwise consistency training. In *CVPR*.

Milioto, A.; and Stachniss, C. 2019. Bonnet: An open-source training and deployment framework for semantic segmentation in robotics using cnns. In *ICRA*.

Munro, J.; and Damen, D. 2020. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*.

Pan, F.; Shin, I.; Rameau, F.; Lee, S.; and Kweon, I. S. 2020. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *CVPR*.

Richter, S. R.; Hayder, Z.; and Koltun, V. 2017. Playing for benchmarks. In *ICCV*.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *ECCV*.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*.

Siam, M.; Gamal, M.; Abdel-Razek, M.; Yogamani, S.; Jagersand, M.; and Zhang, H. 2018. A comparative study of real-time semantic segmentation for autonomous driving. In *CVPR workshops*.

Sumithra, R.; Suhil, M.; and Guru, D. 2015. Segmentation and classification of skin lesions for disease diagnosis. *Procedia Computer Science*.

Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*.

Wan, Z.; Zhang, B.; Chen, D.; Zhang, P.; Chen, D.; Liao, J.; and Wen, F. 2020. Bringing old photos back to life. In *CVPR*.

Xing, Y.; Guan, D.; Huang, J.; and Lu, S. 2022. Domain Adaptive Video Segmentation via Temporal Pseudo Supervision. In *ECCV*.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*.

Zhang, P.; Zhang, B.; Chen, D.; Yuan, L.; and Wen, F. 2020. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*.

Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; and Wen, F. 2021. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*.

Zhang, Y.; and Wang, Z. 2020. Joint adversarial learning for domain adaptation in semantic segmentation. In *AAAI*.

Zheng, Z.; and Yang, Y. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*.

Zhu, X.; Xiong, Y.; Dai, J.; Yuan, L.; and Wei, Y. 2017. Deep feature flow for video recognition. In *CVPR*.

Zhuang, J.; Wang, Z.; and Wang, B. 2020. Video semantic segmentation with distortion-aware feature correction. *TCSVT*.