Pay Attention to Target: Relation-Aware Temporal Consistency for Domain Adaptive Video Semantic Segmentation

Huayu Mai[1*], Rui Sun[1*], Yuan Wang[1], Tianzhu Zhang[1,2†], Feng Wu[1,2]

[1]Deep Space Exploration Laboratory/School of Information Science and Technology, University of Science and Technology of China

[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

{mai556, issunrui, wy2016}@mail.ustc.edu.cn, {tzzhang, fengwu}@ustc.edu.cn

Abstract

Video semantic segmentation has achieved conspicuous achievements attributed to the development of deep learning, but suffers from labor-intensive annotated training data gathering. To alleviate the data-hunger issue, domain adaptation approaches are developed in the hope of adapting the model trained on the labeled synthetic videos to the real videos in the absence of annotations. By analyzing the dominant paradigm consistency regularization in the domain adaptation task, we find that the bottlenecks exist in previous methods from the perspective of pseudo-labels. To take full advantage of the information contained in the pseudo-labels and empower more effective supervision signals, we propose a coherent PAT network including a target domain focalizer and relation-aware temporal consistency. The proposed PAT network enjoys several merits. First, the target domain focalizer is responsible for paying attention to the target domain, and increasing the accessibility of pseudo-labels in consistency training. Second, the relation-aware temporal consistency aims at modeling the inter-class consistent relationship across frames to equip the model with effective supervision signals. Extensive experimental results on two challenging benchmarks demonstrate that our method performs favorably against state-of-the-art domain adaptive video semantic segmentation methods.

Introduction

Video semantic segmentation, which aims to predict a specific semantic class for each pixel in consecutive video frames, has achieved conspicuous achievements attributed to the recent advances in deep neural network (Long, Shelhamer, and Darrell 2015; Wang, Luo, and Zhang 2023; Pan et al. 2023; Sun et al. 2023b) with widespread applications such as autonomous driving, robotics, augmented reality (Cordts et al. 2016; Couprie et al. 2013; Ngan and Li 2011), *etc*. However, it is labor-intensive and time-consuming to gather massive pixel-level annotations as training data. To alleviate the data-hunger issue, a feasible solution is to resort to synthetic data rendered by video game engines (*e.g.*, GTA5 (Richter et al. 2016)) in a self-generated manner with minimal cost. However, video segmentation models trained on synthetic data (*source domain*) inevitably suffer from performance degradation when applied directly to real-world videos (*target domain*) raised by distribution differences (*domain shift*). How to alleviate this gap to empower the learned model generalization capability is thus extremely challenging.

In this work, we focus on the domain adaptive video semantic segmentation (DAVSS) task, which aims to adapt a model trained on source domain videos equipped with segmentation annotations to target domain videos in the absence of accessible labels. To tackle this issue, existing methods can be roughly categorized as adversarial training methods and consistency regularization methods. In adversarial training formulation (Guan et al. 2021), the model seeks to capture domain-invariant spatial-temporal information, but it cannot guarantee a low empirical error on unlabeled target domain videos (Chen et al. 2019; Kumar et al. 2018), along with training stability (Kodali et al. 2017). Recently, consistency regularization paradigm (Gao et al. 2023; Xing et al. 2022; Gao et al. 2023) dominate this field credited to its simplicity yet competitive performance. The core idea of the consistency regularization methods is to impose temporal consistency constraint to the prediction (pseudo labels) of the current frame as well as the one from the previous frame that is warped to the current frame resorting to optical flow.

After an in-depth analysis of the consistency regularization paradigm, we argue that pseudo-labels matter in DAVSS, which is intuitively sensible from the definition of the task itself; that is, pseudo-labels play a dual role - minimizing the inter-domain discrepancy and maximizing the target domain's temporal consistency. However, we find pseudo-labels become performance bottleneck raised by two key ingredients lacking in previous works. **(1)** Accessibility of pseudo-labels. Considering the absent target annotations, models trained on the source domain with significant domain shift are prone to suffer from limited coverage of underlying features for the same class in the target domain (*e.g.*, the decision boundary crosses the high-density region for target domain class *vegetation* in Figure 1 (a), leading to incomplete segmentation with different fragments of object *vegetation*). In this case, considerable unconfident yet reliable pseudo-labels tend to be easily overwhelmed by noise ones. Therefore, it is highly desirable to suppress noisy pseudo-labels and guarantee the truly reliable ones embrace higher weights, in pursuit of enhancing their accessibility. **(2)** Effectiveness of supervision

*Equal contribution

†Corresponding author

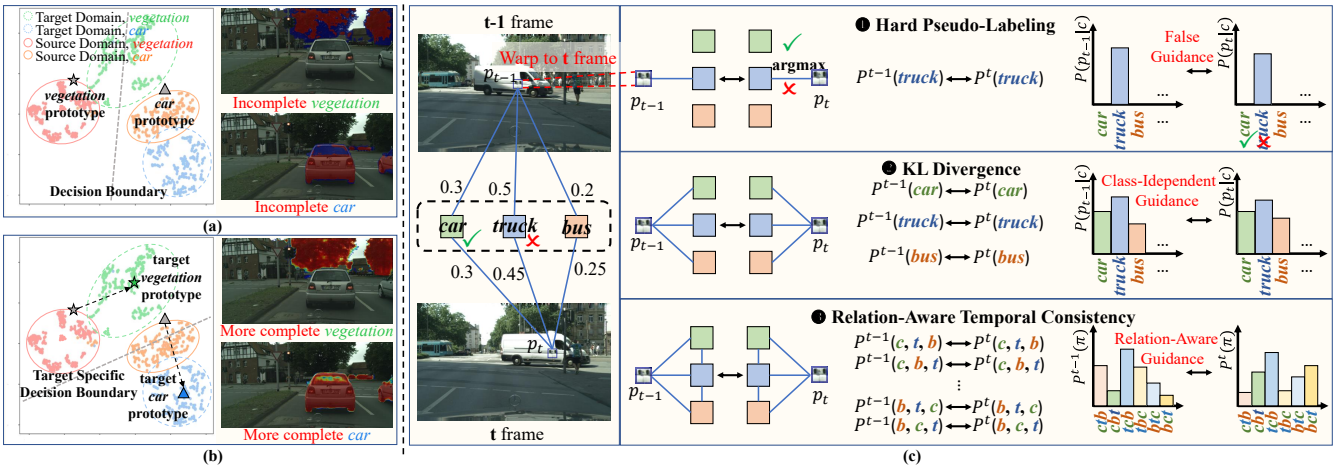

Figure 1: Illustration of our motivation. (a) (b) show the decision boundary with original prototypes and target prototypes. Models trained on the source domain with significant domain shift are prone to suffer from limited coverage of underlying features for the same class in the target domain. (c) shows different ways of modeling temporal consistency. Hard Pseudo-Labeling may suffer from being guided by false pseudo-label, while the KL divergence take each class separately into consideration. Our relation-aware temporal consistency provides more effective guidance.

signals from pseudo-labels. To fully probe the intrinsic temporal information in videos, previous methods often impose temporal consistency by aligning the pseudo-labels of the current frame with those of the previous frame warped to the current frame as supervision signals. The construction strategies for supervision signals involve hard pseudo-labeling and soft pseudo-label. On the one hand, the naive hard pseudo-labeling strategy only selectively recruits the classes with the highest confidence for training but neglects that these labels may be inaccurate (Figure 1 (c) ❶), which can undesirably manifest in the feature learning process and leads to confirmation bias (Guo et al. 2017). On the other hand, retaining scores for all classes as soft pseudo-labels and imposing temporal constraint with KL divergence, as proven by (Ke et al. 2020), can alleviate confirmation bias, but is trapped in taking each class independently and heavily relying on strong i.i.d. assumption (Figure 1 (c) ❷), hindering the learning process. Then, the question naturally arises: how to leverage the inter-class relationship to model structure information to empower more effective supervision signals?

In this paper, we analyze the bottlenecks that exist in previous methods from the perspective of pseudo-labels, and shed light on the possibility of closer collaboration between the pseudo-labels themselves and the supervision signals they constitute. Specifically, we design a coherent PAT network, including a target domain focalizer and relation-aware temporal consistency regularization to ***P**ay **A**ttention to the **T**arget domain and to model the inter-class consistent relationship across frames*. **In the target domain focalizer**. In order to alleviate the inconsistency of class-level underlying features between the two domains raised by domain shift to enhance pseudo-labels accessibility, we draw inspiration from Gestalt law (Koffka 2013) that pixels belonging to the same class within a domain are more similar than those belonging to different classes or domains. We devise the target domain focalizer to pay attention to the target domain by effectively capturing more complete underlying characteristics that fittingly match the features of the target domain, enabling the decision boundary to lie in the low-density region (Figure 1 (b)). The main idea is, enabling prototypes *vegetation* trained on the source domain directly to the target domain undoubtedly leads to incomplete segmentation caused by domain shift, resulting in segmentation fragments (Figure 1 (a)). However, if we retrieve other unrevealed parts supported by segmentation fragments, more complete segmentation results can be achieved (Figure 1 (b)), which generate prototypes closer to the true class *vegetation* centroid of target domain, since these fragments are from the same class within the same domain. In specific, we generate the initial segmentation by directly applying prototypes trained on the source domain to the target domain. Then, based on the resultant segmentation, we collect confident features to generate target-aware prototypes, which are further employed for segmenting this frame. In this way, noisy pseudo-labels will be suppressed while the reliable ones will be highlighted, thus increasing their involvement in consistency regularization-based training.

In the relation-aware temporal consistency regularization. To harness the inter-class relationship modeling structure information for more effective supervision signals, instead of taking each class independently, we carefully design the relation-aware temporal consistency regularization to impose the pixel-class relations of the current frame to be consistent with their counterparts from the previous frame. The core idea is that we take the class ranking as a random event rather than a deterministic permutation. For example, in ranking, given pixel p_t in Figure 1 (c) ❸, its scores vary for different classes, which can be regarded as probabilities. The probability of being ranked first is 0.25 of the class *bus* and 0.3 of the class *car*. The ranking permutation reflects the relevance of classes w.r.t. the pixel p_t. In specific, we transform

the scores of pixel-class relations into class-ranking probability distributions and associate the probability with every rank permutation between the pixel in the current frame and the counterpart from the previous frame. Finally, by constraining the pixel-class relation ranking permutation to be consistent on adjacent frames, the model can be equipped with more effective supervision signals.

In this work, our contributions can be concluded as follows: (1) We analyze the bottlenecks that exist in previous methods from the perspective of pseudo-labels, and shed light on the possibility of closer collaboration between the pseudo-labels themselves and the supervision signals they constitute. (2) We propose a coherent PAT network. Specifically, we design the target domain focalizer to pay the prototypes' attention to the target domain, and the relation-aware temporal consistency to model the inter-class consistent relationship across frames. (3) Extensive experimental results on two challenging benchmarks demonstrate that our method performs favorably against state-of-the-art DAVSS methods.

Related Work

In this section, we briefly overview methods that are related to domain adaptive image semantic segmentation and domain adaptive video semantic segmentation, respectively.

Domain Adaptive Image Semantic Segmentation

With the recent advances in deep neural network (Sun et al. 2021, 2023a,c,d; Wang et al. 2022; Wang, Sun, and Zhang 2023; Luo et al. 2023; Mai et al. 2023), domain adaptive image semantic segmentation (DAISS) has gained significant attention as a solution to address the challenges posed by dense pixel-level annotations and domain shift issues (Melas-Kyriazi and Manrai 2021). Most existing methods can be generally divided into two groups, including adversarial learning based methods and self-training based methods. Adversarial learning based methods endeavor to learn domain-invariant representations by adopting adversarial training at image-level (Choi, Kim, and Kim 2019; Huang et al. 2021b; Kim and Byun 2020), feature-level (Chen, Li, and Van Gool 2018; Huang et al. 2021a; Luo et al. 2019a) or output-level (Luo et al. 2019b; Vu et al. 2019; Lv et al. 2020). The self-training based methods (Li, Yuan, and Vasconcelos 2019; Yang and Soatto 2020) attempt to obtain pseudo labels for target domain data and then utilize the predicted pseudo labels to fine-tune the segmentation model. However, pseudo labels for the target domain data are usually unreliable due to the domain shift (Zheng and Yang 2021). Recently, several methods are proposed to improve the reliability of pseudo labels by domain-aware meta learning (Guo et al. 2021), filtering out noisy samples (Mei et al. 2020) and uncertainty estimation (Zheng and Yang 2021).

Domain Adaptive Video Semantic Segmentation

Video semantic segmentation is a task to predict pixel-level segmentation for individual frames within a video sequence. Current works usually exploit inter-frame temporal relations to achieve accurate and efficient segmentation. For instance, DFF (Zhu et al. 2017) and DAVSS (Zhuang, Wang, and Wang 2020) introduce feature propagation to reuse keyframe features under the guidance of estimated optical flows to reduce computational cost. Accel (Jain, Wang, and Gonzalez 2019) presents an adaptive fusion policy to integrate predictions derived from different frames effectively. However, these methods still need dense pixel-level annotations for training, 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, DA-VSN extends the ADVENT framework (Vu et al. 2019) to target the DAVSS task, encompassing both spatial and temporal adversarial learning. 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. SFC (Gao et al. 2023) converts segmentation maps to optical flows and then imposes consistency between segmentation-based flow and optical flow to implicitly supervise the training of the segmentation model.

Method

In this section, we first formulate the domain adaptive video semantic segmentation (DAVSS) task and present the overview of our method. Then we describe the details of the target domain focalizer and the relation-aware temporal consistency customized for DAVSS. Finally, the training and inference procedure are discussed.

Overview

The DAVSS task aims at learning an accurate video segmentation model in the target domain based on the labeled source domain video sequence $V^{\mathcal{S}} = \{\mathbf{X}_t^{\mathcal{S}}, \mathbf{Y}_t^{\mathcal{S}}\}_{t=1}^{N^{\mathcal{S}}}$ and the unlabeled target domain video sequence $V^{\mathcal{T}} = \{\mathbf{X}_t^{\mathcal{T}}\}_{t=1}^{N^{\mathcal{T}}}$, where $N^{\mathcal{S}}$ and $N^{\mathcal{T}}$ denote the number of frames in each domain. As shown in Figure 2, given a video frame $\mathbf{X} \in \mathbb{R}^{H\times W}$ (omit the superscript $\mathcal{S}/\mathcal{T}$ and the subscript t for convenience) from either domain, the feature extractor first extracts feature map $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$ with the same spatial resolution as the original input, where C denotes the channel number of the feature map. Then a set of learnable prototypes $\mathbf{P} = \{\mathbf{p_k}\}_{k=1}^{K} \in \mathbb{R}^{K\times C}$ is applied to classify the feature $\mathbf{F}_{i,j}$ of each pixel $p_{i,j}$, where K denotes the number of classes. Besides, the target domain focalizer and the relation-aware temporal consistency regularization are proposed to pay the prototypes' attention to the target domain and to model the inter-class consistent relationship across frames. The details are as follows.

Target Domain Focalizer

Considering the absent target annotations, models trained on the source domain with significant domain shift are prone to suffer from limited coverage of underlying features for the same class in the target domain, leading considerable unconfident yet reliable pseudo-labels easily overwhelmed by noise ones. To pay attention to the target domain, we design the target domain focalizer, effectively capturing more complete underlying characteristics of the target domain features. As shown in Figure 3, the target domain focalizer contains two

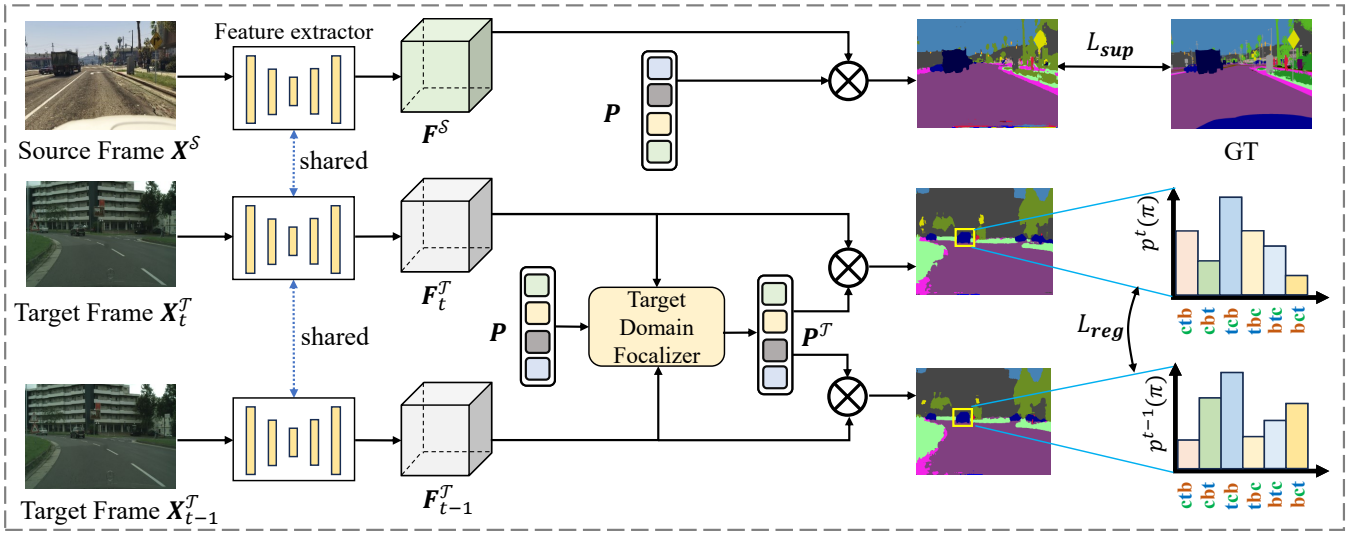

Figure 2: The framework of our method. The source clip and the target clip are fed into the network at the same time. A shared Feature Extractor extracts the feature map for each frame, and a set of prototypes $\mathbf{P}$ is applied to classify the feature of each pixel. The prediction of the source domain will be supervised by the provided ground truth. Considering the domain gap, a target domain focalizer is designed to make the original prototypes focusing on target domain data. A relation-aware temporal Consistency regularization is proposed to model the inter-class consistent relationship across frames.

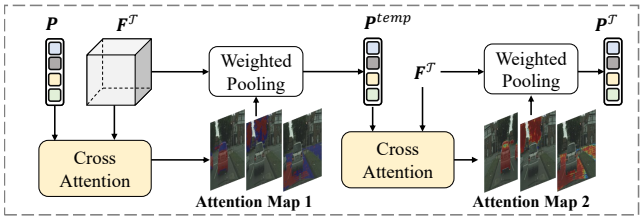

Figure 3: Illustration of the target domain focalizer. The target domain focalizer takes the original prototypes $\mathbf{P}$ as input and outputs the target-aware prototype $\mathbf{P}^{\mathcal{T}}$.

cross-attention stages and produces a set of target-aware prototypes $\mathbf{P}^{\mathcal{T}} = \{\mathbf{p}_k^{\mathcal{T}}\}_{k=1}^K$. Specifically, regarding the original prototypes $\{\mathbf{p}_k\}_{k=1}^K$ as query and the feature map $\mathbf{F}$ as keys, then the k^{th} correlation map $\mathbf{c}_k$ is given as

$$\mathbf{c}_k = \mathbf{p}_k \mathbf{F}^{\mathsf{T}}, \tag{1}$$

where the T refers to the matrix transpose operation. We apply $softmax(\cdot)$ operation along the k dimension and get the initial attention maps $\{\mathbf{a}_k\}_{k=1}^K$:

$$[\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_K] = softmax([\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K]), \tag{2}$$

where $[\cdot, \cdot]$ denotes concatenate operation. With the K attention maps, we can derive another set of prototypes $\mathbf{P}^{temp}$ by weighted pooling the feature map $\mathbf{F}$:

$$\mathbf{p}_k^{temp} = \frac{\sum_{i=1,j=1}^{H,W} \mathbf{a}_{k,i,j} \mathbf{F}_{i,j}}{\sum_{i=1,j=1}^{H,W} \mathbf{a}_{k,i,j}}. \tag{3}$$

Note that due to the domain gap, the attention maps carry lots of noise, which is detrimental to the prototypes. Therefore, before the weighted pooling operation, we filter out those unconfident weights with a predefined threshold τ.

As shown in Figure 3, the **Attention Map 1** produced by $\mathbf{P}$ cannot completely cover the area of the corresponding class, hurting the representative ability of prototypes. Inspired by the Gestalt law (Koffka 2013), *i.e.*, pixels belonging to the same class within a domain are more similar than those belonging to different classes or domains, we repeat the above process again except thresholding and get the target-aware prototypes $\mathbf{P}^{\mathcal{T}}$. It can be seen that the $\mathbf{P}^{temp}$ activates the area of the corresponding class more precisely and completely in Figure 3 (**Attention Map 2**), ensuring $\mathbf{P}^{\mathcal{T}}$ is more target-focused.

In order to expand the coverage of the underlying target domain feature distribution, we maintain a target prototype bank updated by exponential momentum averaging with the momentum θ at each time stamp t:

$$\mathbf{P}^{\mathcal{T}} = \theta \mathbf{P}^{\mathcal{T}} + (1-\theta)\mathbf{P}_t^{\mathcal{T}} \tag{4}$$

Relation-aware Temporal Consistency

In the typical consistency regularization-based methods, either adopt a hard pseudo-labeling or KL divergence strategies, taking each class independently. In fact, there is a certain relation between classes. For example, a pixel with the ground-truth *bus* should be more similar to *truck* rather than *road*, while such a relation is ignored in the previous method. We argue that the inter-class relationship should be considered for more effective supervision signals. We first derive the scores of pixel-class relation $\mathbf{s}_{i,j} \in \mathbb{R}^{1\times K}$ between a pixel $p_{i,j}$ and the target-aware prototypes $\mathbf{P}^{\mathcal{T}}$ by:

$$\mathbf{s}_{i,j} = softmax(\mathbf{F}_{i,j}(\mathbf{P}^{\mathcal{T}})^{\mathsf{T}}). \tag{5}$$

The core idea is that we take the class ranking as a random event rather than a deterministic permutation. That is to say,

VIPER → Cityscapes-Seq																
Methods	road	side.	buil.	fence	light	sign	vege.	terr.	sky	pers.	car	truck	bus	mot.	bike	mIoU
Src.-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'19]	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
FDA [CVPR'20]	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
RDA [ICCV'21]	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
PixMatch [CVPR'21]	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 [ICCV'21]	86.8	36.7	83.5	22.9	30.2	27.7	83.6	26.7	80.3	60.0	79.1	20.3	47.2	21.2	11.4	47.8
I2VDA [ECCV'22]	84.8	36.1	84.0	**28.0**	**36.5**	36.0	85.9	32.5	74.0	**63.2**	81.9	33.0	**51.8**	**39.9**	0.1	51.2
TPS [ECCV'22]	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
SFC [AAAI'23]	**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
PAT (Ours)	85.3	**42.3**	82.5	25.5	33.7	**36.1**	**86.6**	32.8	**84.9**	61.5	**83.3**	34.9	46.9	29.3	29.9	**53.0**
$\Delta\uparrow$	+24.9	+22.4	+3.3	+15.8	+11.3	+15.7	+7.6	+20.2	+2.7	+7.1	+16.0	+29.5	+28.3	+12.3	+17.6	+15.6

Table 1: Quantitative results of different domain adaptive methods on VIPER → Cityscapes-Seq benchmark. We report mIoU (%) and show the improvements over *Src.-only* baseline. The best is highlighted in bold.

every permutation of the classes exists with some probability rather than only the permutation from largest to smallest exists. The probability of one permutation $\pi \in \mathcal{P}$ ($|\mathcal{P}| = K!$) given $\mathbf{s}$ (omit the subscript i, j for convenience) can be calculate as:

$$P(\pi|\mathbf{s}) = \prod_{k=1}^{K} \frac{\mathbf{s}_{\pi(k)}}{\sum_{k'=k}^{K} \mathbf{s}_{\pi(k')}}, \tag{6}$$

where $\pi(k)$ denotes the k^{th} class index of this permutation.

For example, suppose we have three classes: *car*, *truck* and *bus*. One permutation of these three classes is π =(*truck*, *car*, *bus*). Based on the scores of pixel-class relation $\mathbf{s}$, we can derive the probability of π:

$$P(\pi|\mathbf{s}) = \frac{\mathbf{s}(truck)}{\mathbf{s}(car) + \mathbf{s}(truck) + \mathbf{s}(bus)} \cdot \frac{\mathbf{s}(car)}{\mathbf{s}(car) + \mathbf{s}(bus)}. \tag{7}$$

By calculating the probabilities of all $|\mathcal{P}|$ permutations, we transform the scores of pixel-class relation $\mathbf{s}$ into class ranking probability distributions $P(\pi \in \mathcal{P}|\mathbf{s}) \in \mathbb{R}^{1\times|\mathcal{P}|}$, which has modeled the inter-class relationship. In fact, if we calculate full permutations for all K classes, the computational overhead is indeed unacceptable. For computational efficiency, we focus on the permutations of the top-4 classes in each prediction, based on our observation that in every prediction, the top-4 classes have occupied almost all probabilities.

To constraint the inter-class consistent relationship along the temporal dimension, we employ the optical flow network embedded in the VSS model to estimate the optical $\mathcal{O}_{t\to t-1}$, and warp the feature map $\mathbf{F}_{t-1}$ of the $t-1$ frame to spatially aligned with the t frame. Then the relation-aware temporal consistency regularization can be obtained by:

$$L_{reg} = \sum_{i=1}^{H}\sum_{j=1}^{W} \mathcal{L}_{KL}[P(\pi \in \mathcal{P}|\mathbf{s}_{t,i,j}), P(\pi \in \mathcal{P}|\mathbf{s}_{t-1,i,j})], \tag{8}$$

where $\mathcal{L}_{KL}$ denotes the Kullback-Leibler (KL) Divergence loss. Note that KL divergence is applied to measure the class ranking probability distributions between $t-1$ and t frames here, no longer just considering each class independently.

Training and Inference

During training, the prediction of the source domain will be supervised by the provided ground truth $\mathbf{Y}^{\mathcal{S}}$:

$$L_{sup} = \sum_{i=1}^{H}\sum_{j=1}^{W} \mathcal{L}_{ce}(softmax(\mathbf{s}^{\mathcal{S}}_{i,j}), \mathbf{Y}^{\mathcal{S}}_{i,j}), \tag{9}$$

where the $\mathcal{L}_{ce}$ denotes the standard cross entropy loss. As a result, the overall objective of our method is as follows:

$$L = L_{sup} + \lambda L_{reg}, \tag{10}$$

where the λ is the trade-off weight.

During validation, the maintained target prototypes are utilized in place of the original prototypes to be the classifier. Therefore our approach introduces no extra computational cost at validation time.

Experiments

Datasets and Evaluation Metrics

Cityscapes-Seq (Cordts et al. 2016), a real urban dataset, containing $2,975/500$ clips for training/validation at a resolution of $1,024 \times 2,048$, is adopted as the target domain dataset. There are 30 consecutive frames per clip, with the 20^{th} frame carefully annotated.

VIPER (Richter, Hayder, and Koltun 2017) is a synthetic dataset, consisting of $254,064$ frames with labels rendered by the game engine at a resolution of $1,920 \times 1,080$. We use $13,367$ clips of them as one of the source domain datasets.

SYNTHIA-Seq (Ros et al. 2016) is another source domain dataset, comprising $8,000$ labeled photo-realistic frames at a resolution of $1,280 \times 720$, and 850 clips are used for training.

We measure the mean Intersection over Union (mIoU) on the validation set of Cityscapes-Seq over the common classes. Specifically, 15 common classes exit for VIPER → Cityscapes-Seq, and 11 common classes exit for SYNTHIA-Seq → Cityscapes-Seq.

Implementation Details

For a fair comparison, we adopt Accel (Jain, Wang, and Gonzalez 2019) as the video semantic segmentation model

SYNTHIA-Seq → Cityscapes-Seq												
Methods	road	side.	buil.	pole	light	sign	vege.	sky	pers.	rider	car	mIoU
Src.-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 [CVPR'19]	80.5	22.9	68.6	20.9	7.8	18.8	67.0	65.9	43.2	13.4	62.7	42.9
FDA [CVPR'20]	84.1	32.8	67.6	28.1	5.5	20.3	61.1	64.8	43.1	19.0	70.6	45.2
RDA [ICCV'21]	84.7	26.4	73.9	23.8	7.1	18.6	66.7	68.0	48.6	9.3	68.8	45.1
PixMatch [CVPR'21]	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 [ICCV'21]	89.4	31.0	77.4	26.1	9.1	20.4	75.4	74.6	42.9	16.1	82.4	49.5
I2VDA [ECCV'22]	89.9	40.5	77.6	27.3	18.7	23.6	76.1	76.3	48.5	22.4	82.1	53.0
TPS [ECCV'22]	91.2	**53.7**	74.9	24.6	17.9	**39.3**	68.1	59.7	**57.2**	20.3	84.5	53.8
SFC [AAAI'23]	90.9	32.5	76.8	28.6	6.0	36.7	76.0	78.9	51.7	13.8	85.6	52.5
PAT (Ours)	**91.**5	41.3	76.1	**29.6**	**20.9**	33.8	72.4	75.9	51.3	**24.7**	**86.2**	**54.9**
$\Delta\uparrow$	+35.2	+14.7	+0.5	+4.1	+15.2	+18.2	+1.4	+17.4	+9.6	+7.6	+58.3	+16.6

Table 2: Quantitative results of different domain adaptive methods on SYNTHIA-Seq → Cityscapes-Seq benchmark. We report mIoU (%) and show the improvements over *Src.-only* baseline. The best is highlighted in bold.

B	*TDF*	*RTC*	mIoU
✓			48.9
	✓		51.4
		✓	51.7
	✓	✓	**53.0**

Table 3: Ablation study on different components.

TDF	mIoU	FPS
0	51.7	1.15
1	53.0	1.04
2	53.2	0.95
3	53.1	0.87

Table 4: Ablation study on the number of *TDF*.

TC	mIoU
PL	51.3
L2	51.6
KL	51.9
RTC	**53.0**

Table 5: Ablation on ways modeling TC.

following the DAVSS methods (Guan et al. 2021; Xing et al. 2022; Wu et al. 2022; Gao et al. 2023). In specific, it consists of an optical flow network (FlowNet (Dosovitskiy et al. 2015) is adopted), two image segmentation networks (Deeplabv2 (Chen et al. 2017) with ResNet-101 (He et al. 2016) pretrained on ImageNet (Deng et al. 2009) is employed) and a 1×1 convolution fusion layer. During the training and validation phase, we resize the frame of VIPER and Cityscapes-Seq to $720 \times 1,280$ and $512 \times 1,024$. We adopt the SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4} for the network, where the learning rate of the backbone is set to 2.5×10^{-4} and the other is set to 5×10^{-3}. The hyperparameters threshold τ, trade-off weight λ and momentum θ are eventually set to 0.8, 1.0 and 0.999, respectively. All the experiments are implemented on NVIDIA GeForce RTX 3090 with 24 GB memory.

Comparison with State-of-the-art Methods

We conduct experiments on two popular synthetic-to-real benchmarks including VIPER → Cityscapes-Seq and SYNTHIA-Seq → Cityscapes-Seq and make a comparison with SOTA DAVSS methods including DA-VSN (Guan et al. 2021), TPS (Xing et al. 2022), I2VDA (Wu et al. 2022) and SFC (Gao et al. 2023), and several representative DAISS methods including AdvEnt (Vu et al. 2019), FDA (Yang and Soatto 2020), RDA (Huang et al. 2021c) and PixMatch (Melas-Kyriazi and Manrai 2021).

Results on VIPER → Cityscapes-Seq. From Table 1 we can observe that our method outperforms the source-only (*Src.-only*) model by 15.6%. Our method also significantly outperforms the existing SOTA DAVSS methods. Taking the recently proposed method I2VDA (Wu et al. 2022) as an example, the performance gain of our method reaches to +1.8% mIoU.

Results on SYNTHIA-Seq → Cityscapes-Seq. Table 2 compares results of PAT and some SOTA methods on SYNTHIA-Seq → Cityscapes-Seq. The proposed PAT achieves consistent performance gains over the source-only baseline, obtaining improvements of 16.6% mIoU. Our work also significantly outperforms SOTA methods TPS (Xing et al. 2022) by 1.1% mIoU in this scenario, demonstrating the effectiveness of our approach.

Qualitative Results To further analyze and understand the proposed method, we present qualitative comparisons with the state-of-the-art methods TPS (Xing et al. 2022) and SFC (Gao et al. 2023) on VIPER → Cityscapes-Seq in Figure 4. From Figure 4 we can observe that the proposed PAT can achieve more accurate prediction compared to the TPS and SFC (denoted in yellow boxes).

Ablation Study and Analysis

To look deeper into our method, we perform a series of ablation studies on VIPER→Cityscapes-Seq to analyze each component of our PAT, including the **T**arget **D**omain **F**ocalizer (*TDF*) and the **R**elation-aware **T**emporal **C**onsistency (*RTC*). Note that we remove *TDF* and *RTC*, and adopt hard pseudo-labeling as our baseline (*B*).

Effectiveness of Components. As shown in Table 3, both target domain focalizer and relation-aware temporal consistency bring a certain performance lift compared with the baseline.

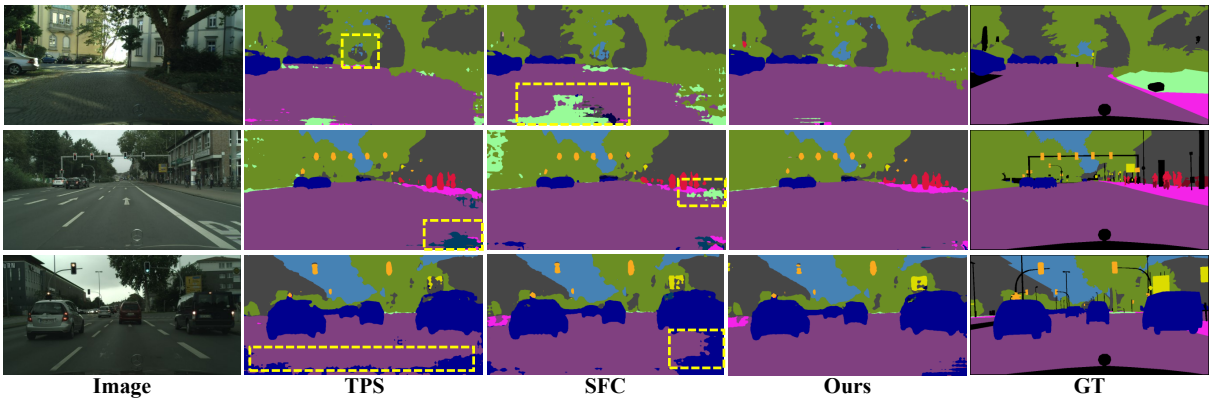

Figure 4: Qualitative comparison of PAT with the state-of-the-art domain adaptive video semantic segmentation benchmark VIPER → Cityscapes-Seq. Compared with existing methods, our method generates better segmentation results on the target video (highlighted by yellow boxes).

Figure 5: Visualization of the attention map in the target domain focalizer.

τ	mIoU
0	51.7
0.6	52.6
0.7	52.7
0.8	**53.0**
0.9	52.9

(a) Threshold τ.

λ	mIoU
0.5	52.5
1.0	**53.0**
1.5	52.7

(b) Trade-off λ.

θ	mIoU
0.9	52.4
0.99	52.8
0.999	**53.0**

(c) Momentum θ.

Table 6: Hyperparameter evaluations.

(1) With the utilization of TDF, a 2.5% improvement of mIoU can be observed, indicating that paying the prototypes' attention to the target domain can benefit the domain adaptation task. In Figure 5, we visualize the attention maps inside the *TDF* for different classes. We can see that in the attention map1 produced by the original prototypes, each class is incompletely activated due to the significant domain shift. With the help of *TDF*, the target-aware prototypes activate the area of the corresponding class more precisely and completely, increasing the accessibility of pseudo-labels in consistency regularization-based training. (2) The introduction of *RTC* achieves further accuracy gains (4.1% in mIoU), mainly ascribed to the leverage of the inter-class relationship to model structure information, providing more effective supervision signals.

Effectiveness of the Target Domain Focalizer. We attempt to to use multiple *TDF*s serially in our experiments, and the results are shown in Table 4. From 0 to 1, we observe a clear performance comparison (51.7% *vs.* 53.0%), demonstrating the efficacy of our *TDF*. From 1 to 3, negligible performance gains are observed but with the cost of reduction of feedforward speed during training. Therefore, we use just one *TDF* in the main experiment.

Effectiveness of the Relation-aware Temporal Consistency. As shown in Tab 5, we compared different ways of modeling **T**emporal **C**onsistency (**TC**), including hard **P**seudo-**L**abeling (*PL*), *L2* and *KL* with soft pseudo-label and our *RTC*. First, by comparing rows 2,3 with row 1 (51.6%, 51.9% *vs.* 51.3%), we can learn that the soft pseudo-label is better than the hard one in the domain adaptive task, attributed to its ability to alleviate confirmation bias. Furthermore, our relation-aware temporal consistency leverages the inter-class to empower more effective supervision signals, achieving better performance.

Hyperparameter Evaluations. As shown in Table 6 (a), (b) and (c), we report the performance of different hyperparameters, including the threshold τ in *TDF*, the trade-off weight λ in training objective and the momentum θ for exponential momentum average updating. Table 6 (a) shows that, without thresholding, (*i.e.*, $\tau = 0$), the performance drop by 1.3% in mIoU, indicating the importance of thresholding operation for constructing representative target-aware prototypes. Except that, our method is insensitive to hyperparameters.

Conclusion

In this paper, we analyze the bottlenecks that exist in previous methods from the perspective of pseudo-labels, and shed light on the possibility of closer collaboration between the pseudo-labels themselves and the supervision signals they constitute. We propose a coherent PAT network. Specifically, we design the target domain focalizer to pay the prototypes attention to the target domain, and the relation-aware temporal consistency to model the inter-class consistent relationship across frames. Extensive experimental results on two challenging benchmarks demonstrate the effectiveness of our method.

Acknowledgments

This work was partially supported by the National Defense Basic Scientific Research Program (Grant JCKY2021130B016).

References

Chen, C.; Xie, W.; Huang, W.; Rong, Y.; Ding, X.; Huang, Y.; Xu, T.; and Huang, J. 2019. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 627–636.

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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.

Chen, Y.; Li, W.; and Van Gool, L. 2018. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7892–7901.

Choi, J.; Kim, T.; and Kim, C. 2019. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6830–6840.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Couprie, C.; Farabet, C.; Najman, L.; and LeCun, Y. 2013. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

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 *Proceedings of the IEEE international conference on computer vision*, 2758–2766.

Gao, Y.; Wang, Z.; Zhuang, J.; Zhang, Y.; and Li, J. 2023. Exploit domain-robust optical flow in domain adaptive video semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 641–649.

Guan, D.; Huang, J.; Xiao, A.; and Lu, S. 2021. Domain adaptive video segmentation via temporal consistency regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8053–8064.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Guo, X.; Yang, C.; Li, B.; and Yuan, Y. 2021. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3927–3936.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, J.; Guan, D.; Lu, S.; and Xiao, A. 2021a. Mlan: Multi-level adversarial network for domain adaptive semantic segmentation. *arXiv preprint arXiv:2103.12991*.

Huang, J.; Guan, D.; Xiao, A.; and Lu, S. 2021b. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6891–6902.

Huang, J.; Guan, D.; Xiao, A.; and Lu, S. 2021c. Rda: Robust domain adaptation via fourier adversarial attacking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8988–8999.

Jain, S.; Wang, X.; and Gonzalez, J. E. 2019. Accel: A corrective fusion network for efficient semantic segmentation on video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8866–8875.

Ke, Z.; Qiu, D.; Li, K.; Yan, Q.; and Lau, R. W. 2020. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, 429–445. Springer.

Kim, M.; and Byun, H. 2020. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12975–12984.

Kodali, N.; Abernethy, J.; Hays, J.; and Kira, Z. 2017. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*.

Koffka, K. 2013. *Principles of Gestalt psychology*, volume 44. Routledge.

Kumar, A.; Sattigeri, P.; Wadhawan, K.; Karlinsky, L.; Feris, R.; Freeman, B.; and Wornell, G. 2018. Co-regularized alignment for unsupervised domain adaptation. *Advances in neural information processing systems*, 31.

Li, Y.; Yuan, L.; and Vasconcelos, N. 2019. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6936–6945.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Luo, N.; Pan, Y.; Sun, R.; Zhang, T.; Xiong, Z.; and Wu, F. 2023. Camouflaged Instance Segmentation via Explicit De-Camouflaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17918–17927.

Luo, Y.; Liu, P.; Guan, T.; Yu, J.; and Yang, Y. 2019a. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6778–6787.

Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; and Yang, Y. 2019b. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, 2507–2516.

Lv, F.; Liang, T.; Chen, X.; and Lin, G. 2020. Cross-domain semantic segmentation via domain-invariant interactive relation transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4334–4343.

Mai, H.; Sun, R.; Zhang, T.; Xiong, Z.; and Wu, F. 2023. DualRel: Semi-Supervised Mitochondria Segmentation From a Prototype Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19617–19626.

Mei, K.; Zhu, C.; Zou, J.; and Zhang, S. 2020. 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*, 415–430. Springer.

Melas-Kyriazi, L.; and Manrai, A. K. 2021. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12435–12445.

Ngan, K. N.; and Li, H. 2011. *Video segmentation and its applications*. Springer Science & Business Media.

Pan, Y.; Luo, N.; Sun, R.; Meng, M.; Zhang, T.; Xiong, Z.; and Zhang, Y. 2023. Adaptive template transformer for mitochondria segmentation in electron microscopy images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21474–21484.

Richter, S. R.; Hayder, Z.; and Koltun, V. 2017. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2213–2222.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 102–118. Springer.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3234–3243.

Sun, R.; Li, Y.; Zhang, T.; Mao, Z.; Wu, F.; and Zhang, Y. 2021. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10938–10947.

Sun, R.; Luo, N.; Pan, Y.; Mai, H.; Zhang, T.; Xiong, Z.; and Wu, F. 2023a. Appearance Prompt Vision Transformer for Connectome Reconstruction. In Elkind, E., ed., *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 1423–1431. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Sun, R.; Mai, H.; Luo, N.; Zhang, T.; Xiong, Z.; and Wu, F. 2023b. Structure-decoupled adaptive part alignment network for domain adaptive mitochondria segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 523–533. Springer.

Sun, R.; Mai, H.; Zhang, T.; and Wu, F. 2023c. DAW: Exploring the Better Weighting Function for Semi-supervised Semantic Segmentation. In *Advances in Neural Information Processing Systems*.

Sun, R.; Wang, Y.; Mai, H.; Zhang, T.; and Wu, F. 2023d. Alignment before aggregation: trajectory memory retrieval network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1218–1228.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2517–2526.

Wang, Y.; Luo, N.; and Zhang, T. 2023. Focus on Query: Adversarial Mining Transformer for Few-Shot Segmentation. In *Advances in Neural Information Processing Systems*.

Wang, Y.; Sun, R.; and Zhang, T. 2023. Rethinking the Correlation in Few-Shot Segmentation: A Buoys View. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7183–7192.

Wang, Y.; Sun, R.; Zhang, Z.; and Zhang, T. 2022. Adaptive agent transformer for few-shot segmentation. In *European Conference on Computer Vision*, 36–52. Springer.

Wu, X.; Wu, Z.; Wan, J.; Ju, L.; and Wang, S. 2022. Is It Necessary to Transfer Temporal Knowledge for Domain Adaptive Video Semantic Segmentation? In *European Conference on Computer Vision*, 357–373. Springer.

Xing, Y.; Guan, D.; Huang, J.; and Lu, S. 2022. Domain adaptive video segmentation via temporal pseudo supervision. In *European Conference on Computer Vision*, 621–639. Springer.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4085–4095.

Zheng, Z.; and Yang, Y. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4): 1106–1120.

Zhu, X.; Xiong, Y.; Dai, J.; Yuan, L.; and Wei, Y. 2017. Deep feature flow for video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2349–2358.

Zhuang, J.; Wang, Z.; and Wang, B. 2020. Video semantic segmentation with distortion-aware feature correction. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8): 3128–3139.