Unified Domain Adaptive Semantic Segmentation

Zhe Zhang, Gaochang Wu, *Member, IEEE*, Jing Zhang, Xiatian Zhu, Dacheng Tao, *Fellow, IEEE*, Tianyou Chai, *Life Fellow, IEEE*

***Abstract*—Unsupervised Domain Adaptive Semantic Segmentation (UDA-SS) aims to transfer the supervision from a labeled source domain to an unlabeled and shifted target domain. The majority of existing UDA-SS works typically consider images whilst recent attempts have extended further to tackle videos by modeling the temporal dimension. Although two lines of research share the major challenges – overcoming the underlying domain distribution shift, their studies are largely independent. It causes several issues: (1) The insights gained from each line of research remain fragmented, leading to a lack of holistic understanding of the problem and potential solutions. (2) Preventing the unification of methods and best practices across two scenarios (images and videos) will lead to redundant efforts and missed opportunities for cross-pollination of ideas. (3) Without a unified approach, the knowledge and advancements made in one scenario may not be effectively transferred to the other, leading to suboptimal performance and slower progress. Under this observation, we advocate unifying the study of UDA-SS across video and image scenarios, enabling a more comprehensive understanding, synergistic advancements, and efficient knowledge sharing. To that end, we explore the unified UDA-SS from a general domain augmentation perspective, serving as a unifying framework, enabling improved generalization, and potential for cross-pollination, ultimately contributing to the practical impact and overall progress. Specifically, we propose a Quad-directional Mixup (QuadMix) method, characterized by tackling intra-domain discontinuity, fragmented gap bridging, and feature inconsistencies through four-directional paths designed for intra- and inter-domain mixing within an explicit feature space. To deal with temporal shifts within videos, we incorporate optical flow-guided feature aggregation across spatial and temporal dimensions for fine-grained domain alignment, which is extendable to image scenarios. Extensive experiments show that QuadMix outperforms the state-of-the-art works by large margins on four challenging UDA-SS benchmarks. Our source code and models will be released at https://github.com/ZHE-SAPI/UDASS.**

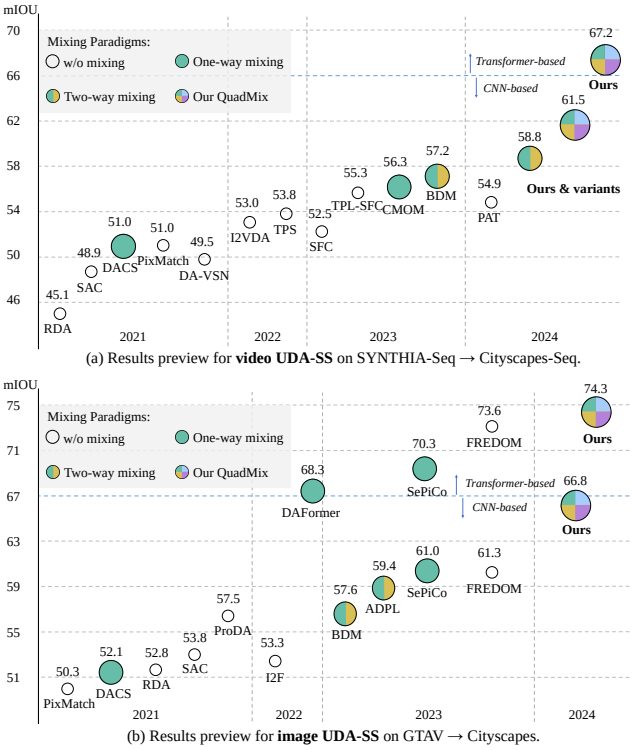
***Index Terms*—Unsupervised domain adaptation, semantic segmentation, unified adaptation, domain mixup.**

I. Introduction

UNSUPERVISED Domain Adaptive Semantic Segmentation (UDA-SS) focuses on transferring pixel-level dense knowledge from a labeled source domain to an unlabeled target domain with distribution shift [1]–[4]. The key challenge is all about how to tackle intrinsic distributional gaps between the two domains without target domain supervision. As typical, the literature is dominated by `image` UDA-SS with the aim to reduce the domain discrepancy through image alignment [5]–[7], metric minimization [7]–[9], self-supervised learning with pseudo-label [10]–[12], or regularization [13], [14]. This inspires the progress of `video` UDA-SS by additionally handling the temporal dimension via video alignment [15], [16], geometric-based cues [17], [18] for domain-invariant learning.

While the two UDA-SS tasks share the congenetic challenge, i.e., feature distributional gap across domains, their investigation is clearly separated. This could cause limitations like fragmented insights over tasks, lack of holistic understanding of the challenge, preventing the unification of techniques, and finally slowing the development progress.

Under the aforementioned insights, in this work, we advocate the unified investigation of image and video UDA-SS. To have a more generic solution, we explore the novel domain augmentation strategy by introducing a novel method, **Quad-directional Mixup** (QuadMix), capable of generating a diversity of intermediate domains within both source and target



(a) Results preview for **video UDA-SS** on SYNTHIA-Seq → Cityscapes-Seq.

(b) Results preview for **image UDA-SS** on GTAV → Cityscapes.

Fig. 1. Instead of studying UDA-SS for images and videos independently, we explore the unified study with a single approach.

Zhe Zhang, Gaochang Wu, and Tianyou Chai are with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, China. Email: zhangzhe17@stumail.neu.edu.cn, wugc@mail.neu.edu.cn, tychai@mail.neu.edu.cn.

Jing Zhang is with the School of Computer Science, Wuhan University, China. E-mail: jingzhang.cv@gmail.com.

Xiatian Zhu is with the Surrey Institute for People-Centred Artificial Intelligence, and Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK. E-mail: Eddy.zhuxt@gmail.com.

Dacheng Tao is with the College of Computing & Data Science, Nanyang Technological University, Singapore. E-mail: dacheng.tao@gmail.com

Corresponding author: Tianyou Chai, Gaochang Wu.

domains and across them within an explicit feature space, facilitated by an adaptive feature aggregation module for fine-grained domain alignment. Specifically, we generate adaptive category-aware sample patch templates from either (source or target) domain, followed by fusing them within the domain to create generalized `intra-mixed` domains. Concurrently, we further fuse those patches across the original domains interactively to construct more enhanced `inter-mixed` domains. In addition to common pixel-level mixing, feature-level mixing across the quad-mixed domains is also proposed to alleviate feature inconsistencies caused by domain-wise mixed-semantic discrepancies. Furthermore, to deal with persistent category-aware domain gaps, we design a unified feature aggregation module guided by spatial cues for image UDA-SS and spatio-temporal cues for video UDA-SS, enabling fine-grained domain alignment. Technically, our unified method goes beyond existing alternatives for image or video UDA-SS with limitations such as intra-domain discontinuity due to over distributed samples [10], [16], [19], [20], less generalizable feature distribution due to limited mixing directions [1], [2], [5], [21], [22], and feature inconsistencies due to diverse mixed-semantic contexts [22], [23].

We make the following contributions: (1) We *for the first time* investigate image and video UDA-SS from a unified perspective due to their shared major challenges. (2) We introduce a unifying UDA-SS framework, Quad-directional Mixup (QuadMix), which involves pixel- and feature-level mixing to facilitate comprehensive domain gap bridging and alleviate feature inconsistencies through four-directional paths for intra- and inter-domain mixing. For temporal shifts in videos, we incorporate optical flow-guided spatio-temporal feature aggregation, which is well-extensible to image scenarios. (3) As shown in Fig 1, extensive experiments on four image and video benchmarks show that our method sets the new state-of-the-art performance in comparison to prior works.

II. Related Works

A. Unsupervised Domain Adaption for Image Segmentation

Image UDA-SS has garnered significant attention in both theoretical and practical research. At the image level, the mainstream is to transfer image styles while preserving original semantics. Various techniques including cycle translation [6], patch adversarial [24], dual path learning [22], and diffusion models [25] have been proposed. At the feature level, influential metrics and the variants [8], [9], [26] are developed to address distribution discrepancies across domains. Additionally, using prototypes to represent feature centroids is proposed for pseudo-label denoising [27], divergence measurement [12], [28], and contrastive learning [23], [29]. At the output level, adversarial adaptation on low-dimensional logit space has been proposed in [30]. Other works, such as progressive confidence [31] and curriculum learning [32], [33], aim to refine target pseudo-labels within a structured output space [34]. Moreover, from the perspective of feature distribution, generating intermediate samples [1], [5], [21]–[23] to mitigate domain gaps has yielded impressive results. However, the intra-domain discontinuity, less generalizable distribution, and feature inconsistencies are significant challenges that often hinder effective knowledge transfer to the target domain.

B. Unsupervised Domain Adaptive Video Segmentation

Extending image segmentation, Video Semantic Segmentation (VSS) entails pixel-level discrimination for adjacent video frames [35], [36]. Mainstream works focus on enhancing segmentation accuracy, proposing methods involving feature warping [37], [38], recurrent networks [39], or auxiliary networks [40]. To improve inference efficiency, methods like attention-based spatial feature reconstruction [41] and weight-sharing subnetworks [42] have been proposed. However, they heavily rely on laborious dense labels, posing challenges in adapting to shifted and unlabeled target videos.

To address this, video UDA-SS has recently gained great interest. Pioneering works [15], [16] employ adversarial learning with DCGAN [43] to extract domain-invariant video features. Guan et al. [16] further improve this by aligning uncertain target predictions with more confident adjacent ones. However, asynchronous issues with generator-discriminator convergence often cause training instability. Additionally, auxiliary cues involving geometric motion [18], [44] or optical flow [17], [19] are introduced to generate pseudo-labels. However, they train the model on samples from each domain alternately but overlook domain interactions, leading to isolated learning that hampers adaptation. Meanwhile, Wu et al. [45] explore transferring spatial knowledge from image to video domains while overlooking temporal feature gaps. Cho et al. [2] perform one-way pixel-level mixing from source to target but overlook the intra-domain discontinuity across domains. In addition, they obtain pseudo-labels offline using models in [16] and only align single-frame spatial features. In contrast, our method is trained end-to-end and holistically mitigates temporal shifts.

From a unified view, we consider the image to be a special case of the video without the temporal dimension. For video UDA-SS, we further leverage **video temporal cues** fully by (1) using optical flow to generate cross-frame consistent pseudo-labels; (2) integrating temporal cues to spatial features for each domain; and (3) aggregating video features of consecutive frames for fine-grained domain alignment. This is the first work applying the vision transformer [46] to video UDA-SS.

C. Intermediate Domains for Adaption

Current works have achieved impressive results by addressing feature gaps with meticulously designed intermediate domains for images [1], [4], [5], [21]–[23] and videos [2]. However, the discontinuous distribution of features within each domain (intra-domain) always results in distinct point attributes [47], which makes current intermediates derived from such distinct domains with poorly generalized distributions [48]. Fig. 3 (a)-(d) present a comparison of existing methods, distinguishing source features in green and target features in purple. Inspired by [49], [50], strategies like one-way mixing [1], [2], [21], [23] and two-way mixing [5], [22] extensively explore the integration of source patches into the target domain, or vice versa, aiming to bridge domain gaps. Nonetheless, t-SNE [51] feature visualization in Fig. 9

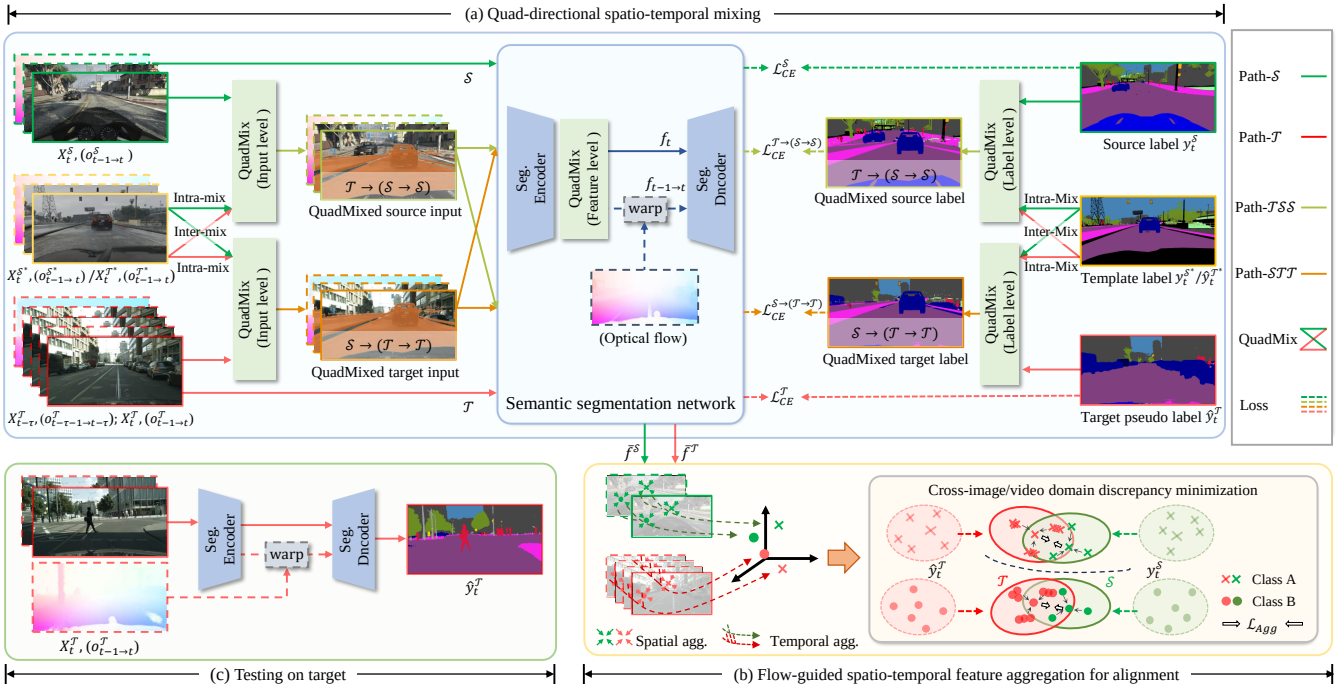


Fig. 2. Overview of the proposed QuadMix for UDA-SS. ***Image UDA-SS*** follows a parallel approach, with the exception of temporal cues, as indicated by the dashed lines. (i) In part (a), QuadMix comprises four comprehensive intra/inter-domain mixing paths to bridge domain gaps at spatio-temporal pixel and feature levels. Pixel-level mixing is performed on adjacent frames, optical flow, and labels/pseudo-labels, aiming to generate two enhanced inter-mixed domains that derived from intra-mixed domains iteratively: $\mathcal{T} \rightarrow (\mathcal{S} \rightarrow \mathcal{S})$ and $\mathcal{S} \rightarrow (\mathcal{T} \rightarrow \mathcal{T})$. These intermediates overcome the *intra-domain discontinuity* within $\mathcal{S}$ and $\mathcal{T}$ and exhibit more *generalizable features* for gap bridging. Additionally, feature-level mixing across quad-mixed domains alleviates *feature inconsistencies* caused by distinct domain-wise video contexts; (ii) In part (b), optical flow-guided spatio-temporal feature aggregation compresses video features across domains into a compacted category-aware space, minimizing intra-category discrepancies and enhancing inter-category discriminability for target domain; (iii) The training process is end-to-end. In part (c), stacked adjacent frames $X_t^{\mathcal{T}}$ and optical flow $o_{t-1\rightarrow t}^{\mathcal{T}}$ are needed for target domain testing.

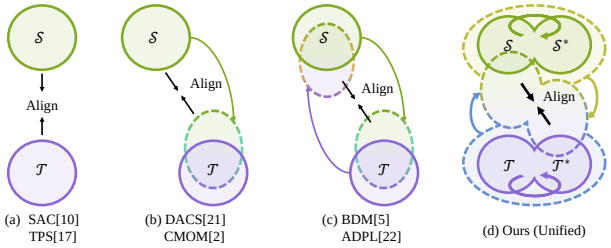


Fig. 3. Comparison of domain mixing paradigms. Going beyond existing ideas with the limitations such as intra-domain discontinuity [10], [17], less generalizable feature distribution [2], [5], [21], [22], and feature inconsistencies [2], [5], [10], [17], [21], [22], the proposed QuadMix generalizes **intra-mixed** domains and enhances **inter-mixed** domains at both spatial (temporal) pixel- and feature-levels. The symbol "*" denotes the sample templates.

(a)(c)(d) reveals that the effect of existing intermediates across domains (inter-domain) is fragmented and insufficient, and persistent category-aware domain gaps continue to hinder effective knowledge generalization. Moreover, pure pixel-level mixing may lead to model confusion because of the feature inconsistencies caused by distinct domain-wise contexts.

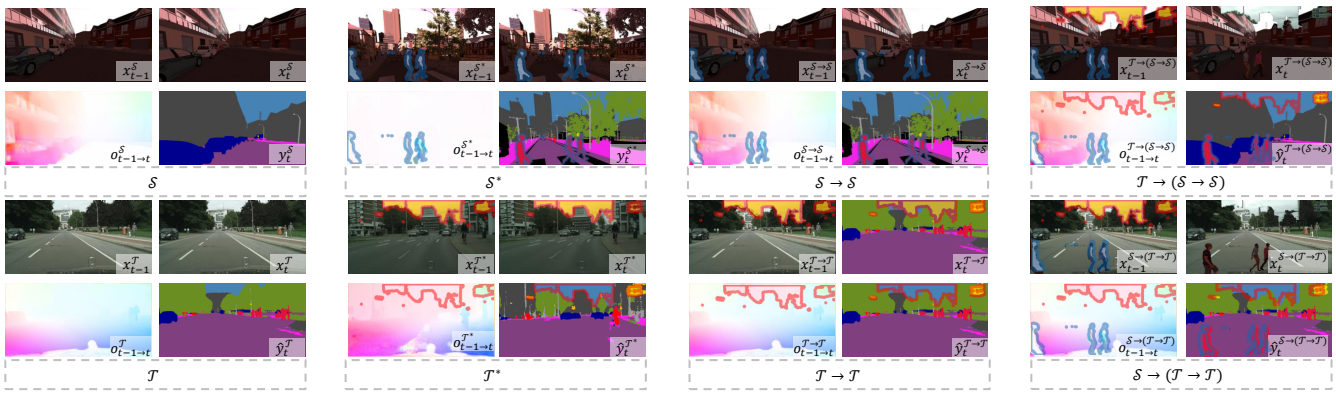
III. Method

In this section, we start with an overview in Section III-A. Then, we elaborate on pixel-level mixing for intra- and inter-domains, as well as feature-level mixing across the overall quad-mixed domains in Section III-B to Section III-D. Next, Section III-E presents the flow-guided spatio-temporal feature aggregation module, and Section III-F covers the training loss.

A. Method Overview

Generally, current studies tend to address image and video UDA-SS tasks independently despite the shared challenges, leading to redundant efforts and missed opportunities for cross-pollination. In this paper, we tackle the more challenging task of unified UDA-SS. As depicted in Fig. 2, we propose a QuadMix method that integrates four-directional paths designed for both intra- and inter-domains, operating at spatial (frames and labels), temporal (optical flow of videos), and feature levels. Technically, the online generated category-aware video/image patch templates from two domains are internally fused with their original domains to create intra-mixed source ($\mathcal{S}\rightarrow\mathcal{S}$) and target ($\mathcal{T}\rightarrow\mathcal{T}$) domains with smoother and generalized feature embeddings. Additionally, both pixel- and feature-level patch templates are interactively fused with the intra-mixed domains to generate more enhanced inter-mixed source ($\mathcal{T}\rightarrow(\mathcal{S}\rightarrow\mathcal{S})$) and target ($\mathcal{S}\rightarrow(\mathcal{T}\rightarrow\mathcal{T})$) domains, serving as comprehensive and efficient intermediates for gap bridging (see Fig. 9 (e)). Our QuadMix is fairly practical and effective. A detailed comparison with previous works is shown in Fig. 3.

Video is uniquely characterized by the temporal dimension as compared to image. To account for this intrinsic information, we propose an adaptive flow-guided spatial-temporal feature aggregation module for fine-grained domain alignment (Fig. 2 (b)). The spatial feature of each frame is enriched with temporal knowledge of previous frames via optical flow propagation. Based on this, spatial-level aggregation is guided by pseudo masks from the source and target domain adaptively.



(a) Before QuadMix (b) Templates (c) Intra-domain Mix (d) After QuadMix

Fig. 4. Examples of various mixing strategies in QuadMix for ***video UDA-SS***: (a) $\mathcal{S}$ and $\mathcal{T}$ (before QuadMix), (b) $\mathcal{S}^*$ and $\mathcal{T}^*$ (source templates: *person* and *rider*, target templates: *sign* and *sky*), (c) $\mathcal{S} \to \mathcal{S}$ and $\mathcal{T} \to \mathcal{T}$ (intra-domain mixing), (d) $\mathcal{S} \to (\mathcal{T} \to \mathcal{T})$ and $\mathcal{T} \to (\mathcal{S} \to \mathcal{S})$ (further with inter-domain mixing, i.e. after QuadMix). The effects of these strategies on video frames, optical flow, and labels/pseudo-labels are illustrated. We present $x_t^{\mathcal{T}\to(\mathcal{S}\to\mathcal{S})}$ and $x_t^{\mathcal{S}\to(\mathcal{T}\to\mathcal{T})}$ without masks for better understanding. Notably, the patch templates required in training iteration n are generated online adaptively from iteration $n-1$. Please zoom in for details.

While at the temporal level, features are further integrated based on logit entropy to enhance overall temporal coherence. The aggregated features across domains are finely aligned in a fine-grained manner. It can be extended to image UDA-SS using spatial feature aggregation, as an image is essentially a special case of a video without the temporal dimension.

The main notations are listed as follows. For video UDA-SS, let $X_t^{\mathcal{S}} = \mathbb{S}\{x_{t-1}^{\mathcal{S}}, x_t^{\mathcal{S}}\}$ with $o_{t-1\to t}^{\mathcal{S}}$ representing source frames alongside optical flow, and $y_t^{\mathcal{S}}$ to be the truth label of $x_t^{\mathcal{S}}$, where $\mathbb{S}$ denotes a stack operation along time dimension. The target frames and optical flow are $X_t^{\mathcal{T}} = \mathbb{S}\{x_{t-1}^{\mathcal{T}}, x_t^{\mathcal{T}}\}$ and $o_{t-1\to t}^{\mathcal{T}}$ but without label $y_t^{\mathcal{T}}$. For image UDA-SS, the source image and label are denoted as $x^{\mathcal{S}}$ and $y^{\mathcal{S}}$ while the target image is $x^{\mathcal{T}}$ without label $y^{\mathcal{T}}$. As for the pseudo-label generation, for the video scenario, $X_{t-\tau}^{\mathcal{T}} = \mathbb{S}\{x_{t-\tau-1}^{\mathcal{T}}, x_{t-\tau}^{\mathcal{T}}\}$ and $o_{t-\tau-1\to t-\tau}^{\mathcal{T}}$ from previous timestamp are input into the model to obtain $\hat{y}_{t-\tau}^{\mathcal{T}}$, which is then warped using $o_{t-\tau\to t}^{\mathcal{T}}$ to generate temporally consistent $\hat{y}_t^{\mathcal{T}}$, where τ denotes the video temporal deviation. For the image scenario, pseudo-labels are generated by a momentum network [52] for each single image.

B. Video/Image Patch Templates Generation Across Domains

The domain gap becomes more evident when we observe the motion of diverse objects at the category-aware level. To mitigate the intrinsic distributional gaps caused by *intra-domain discontinuity* and *less generalizable feature distribution* within both video and image scenarios, we propose quad-directional mixing, which generates intermediate intra- and inter-domain representations within a unified feature space, thereby generalizing features and bridging domain gaps. This design necessitates the adaptive generation of diverse and reliable video/image patch sequences as templates, as denoted in Fig. 4. In this paper, we adaptively generate cross-domain patch templates online. Specifically, for training iteration n, the source domain patch templates $Z_{t,n}^{\mathcal{S}^*,k}$, $k \in K_{\mathcal{S}}$ and target domain patch templates $Z_{t,n}^{\mathcal{T}^*,k}$, $k \in K_{\mathcal{T}}$ required for quad-directional mixing are derived from the previous iteration $n-1$, where the symbol "*" denotes the templates. These patch templates contain pixel regions belonging to specific categories k, which are randomly selected from segmentation label spaces $K_{\mathcal{S}}$ and $K_{\mathcal{T}}$, respectively.

For **video UDA-SS**, patch templates form a unified set of category-aware adjacent frames, labels/pseudo-labels, and optical flow. The target video patch template is defined as:

$$Z_{t,n}^{\mathcal{T}^*,k} = \left\{ X_{t,n-1}^{\mathcal{T}^*,k}, \hat{y}_{t,n-1}^{\mathcal{T}^*,k}, o_{t-1\to t,n-1}^{\mathcal{T}^*,k} \right\}, \tag{1}$$

specifically,

$$X_{t,n-1}^{\mathcal{T}^*,k} = X_{t,n-1}^{\mathcal{T}^*} \odot \mathbb{1}(\mathcal{F}(\hat{Y}_{t,n-1}^{\mathcal{T}^*}) == k), \tag{2}$$

$$\hat{y}_{t,n-1}^{\mathcal{T}^*,k} = \hat{y}_{t,n-1}^{\mathcal{T}^*} \odot \mathbb{1}(\mathcal{F}(\hat{y}_{t,n-1}^{\mathcal{T}^*}) == k), \tag{3}$$

$$o_{t-1\to t,n-1}^{\mathcal{T}^*,k} = o_{t-1\to t,n-1}^{\mathcal{T}^*} \odot \mathbb{1}(\mathcal{F}(\hat{y}_{t-1,n-1}^{\mathcal{T}^*}) == k), \tag{4}$$

where $k \in K_{\mathcal{T}^*}$ and corresponding to two selected categories, $\mathbb{1}(\cdot)$ is the indicator function, $\odot$ refers to the pixel mapping operation using Hadamard product, and $\mathcal{F}$ stands for a target filtering operation based on the confidence level of segmentation logits. Detailed depictions are shown in Fig. 5(a)(c).

Notably, the model takes $X_t = \mathbb{S}\{x_{t-1}, x_t\}$ with $o_{t-1\to t}$ (indices omitted) as inputs and generates a segmentation map for x_t. Therefore, the bottom-right time subscripts of pseudo-labels are t in Eq. (3) and $t-1$ in Eq. (4) for corresponding category-aware masks. Additionally, $\hat{Y}_t^{\mathcal{T}} = \mathbb{S}\{\hat{y}_{t-1}^{\mathcal{T}}, \hat{y}_t^{\mathcal{T}}\}$, the generation of which will be shown in Eq. (17). In this paper, masks for pixel mapping evolve with pseudo labels rather than being pre-generated offline as in [2], [5], avoiding the accuracy limitation imposed by a frozen warm-up model.

The source video patch template $Z_{t,n}^{\mathcal{S}^*,k}$, $k \in K_{\mathcal{S}}$ follows a pattern parallel to Eq. (1), just needing to replace $\mathcal{T}$ with $\mathcal{S}$ and employ source labels instead of target pseudo-labels in Eq. (1) to Eq. (4). To promote feature diversity within quad-mixed domains, we ensure that the categories k of video patch templates differ across domains at each training iteration.

For **image UDA-SS**, the image patch template also forms a set but excludes temporal cues, which is denoted as $Z_{t,n}^{\mathcal{T}^*,k} = \left\{ x_{t,n-1}^{\mathcal{T}^*,k}, \hat{y}_{t,n-1}^{\mathcal{T}^*,k} \right\}$ for the target domain, and $Z_{t,n}^{\mathcal{S}^*,k}$ similarly for the source domain. For convenience, we retain the subscript t for images, considering them as a special case of videos.

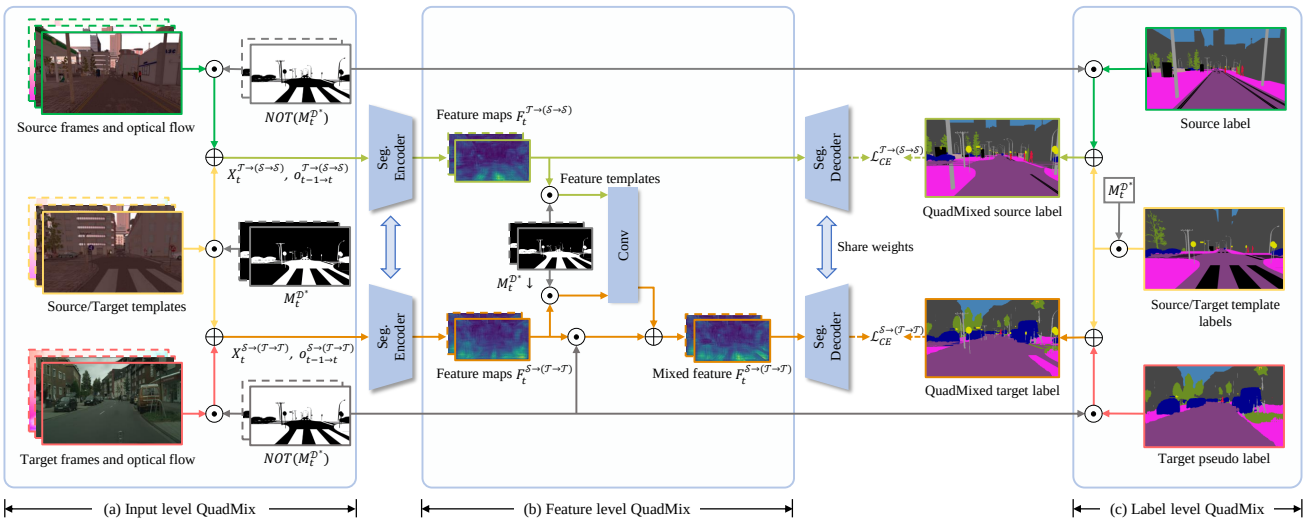


Fig. 5. Details of the quad-directional mixing for UDA-SS. To alleviate intra-domain discontinuity for comprehensive domain gap bridging, we conduct QuadMix at the spatial level (video adjacent frames and label), temporal level (optical flow), and spatio-temporal feature level, constructing more enhanced inter-mixed source ($\mathcal{T}\to(\mathcal{S}\to\mathcal{S})$) and inter-mixed target ($\mathcal{S}\to(\mathcal{T}\to\mathcal{T})$) domains that derived from intra-mixed domains ($\mathcal{S}\to\mathcal{S}$) and ($\mathcal{T}\to\mathcal{T}$) with generalized feature spaces. The mask $M_t^{\mathcal{D}^*}$ denotes the union of source patch template mask $M_t^{\mathcal{S}^*}$ and target patch template mask $M_t^{\mathcal{T}^*}$.

In this way, we generate category-aware patch templates online as templates, which are then utilized for quad-directional mixing to bridge domain gaps in video or image scenarios.

C. Intra-Mixed Video/Image Domain Adaptation

According to Fig. 3(a) and feature visualization in Fig. 9(a), despite the semantic similarity between source and target domains, a notable feature gap often persists in spatial structure and temporal coherence. Fig. 4 presents a depiction of four mixed paths of QuadMix. Rather than purely alternative training with cross-domain data [10], [17], or one way [1], [2], [21], [23]/bidirectional [5], [22] data pasting guided by offline-generated pseudo-labels via warm-up models stage-wisely [2], [5], we propose a comprehensive QuadMix trained end-to-end. A more detailed process is shown in Fig. 5. In this section, we present the intra-mixed domain adaptation.

1) Target to Target Mixing ($\mathcal{T}\to\mathcal{T}$): From the view of feature distribution, we consider source and target video/image domains as two distinct points featured with discontinuous intra-domain feature distribution, and significant inter-domain feature gap. To address the feature discontinuity within each domain for better gap bridging, we propose smoothing and generalizing intra-domains within an explicit feature space. In this part, we focus on the target domain.

For **video UDA-SS**, adaptive target video patch templates are iteratively integrated into target domain samples, guided by pseudo mask sequences. The process unfolds as follows:

$$X_{t,n}^{\mathcal{T}\to\mathcal{T}} = X_{t,n-1}^{\mathcal{T}^*,k} + X_{t,n}^{\mathcal{T}} \odot (1 - M_{t,n-1}^{\mathcal{T}^*,k}), \tag{5}$$

$$\hat{y}_{t,n}^{\mathcal{T}\to\mathcal{T}} = \hat{y}_{t,n-1}^{\mathcal{T}^*,k} + \hat{y}_{t,n}^{\mathcal{T}} \odot (1 - m_{t,n-1}^{\mathcal{T}^*,k}), \tag{6}$$

$$o_{t-1\to t,n}^{\mathcal{T}\to\mathcal{T}} = o_{t-1\to t,n-1}^{\mathcal{T}^*,k} + o_{t-1\to t,n}^{\mathcal{T}} \odot (1 - m_{t-1,n-1}^{\mathcal{T}^*,k}), \tag{7}$$

for simplicity, Eq. (5) to Eq. (7) can be unified as follows:

$$Z_{t,n}^{\mathcal{T}\to\mathcal{T}} = Z_{t,n}^{\mathcal{T}^*,k} + Z_{t,n}^{\mathcal{T}} \odot (1 - M_{t,n-1}^{\mathcal{T}^*,k}), \tag{8}$$

where $M_{t,n-1}^{\mathcal{T}^*,k} = \mathbb{S}\{m_{t-1,n-1}^{\mathcal{T}^*,k}, m_{t,n-1}^{\mathcal{T}^*,k}\}$ is the category-aware binary pseudo mask sequence corresponding to video patch template $Z_{t,n}^{\mathcal{T}^*,k}$, $k \in K_{\mathcal{T}}$, with a resolution of $B \times H \times W$, where B is the batch size. It can be denoted as follows:

$$M_{t,n-1}^{\mathcal{T}^*,k} = \mathbb{1}(\mathcal{F}(\hat{Y}_{t,n-1}^{\mathcal{T}^*}) == k), k \in K_{\mathcal{T}}. \tag{9}$$

In this way, we iteratively integrate category-aware target video templates alongside target video samples mapped by $1 - M$ to generate data for the intra-mixed target domain.

For **image UDA-SS**, the generation of *target to target mixing* samples $Z_{t,n}^{\mathcal{T}\to\mathcal{T}}$ can be performed following Eq. (8) using $M_{t,n-1}^{\mathcal{T}^*,k} = \mathbb{1}(\mathcal{F}(\hat{y}_{t,n-1}^{\mathcal{T}^*}) == k)$, without the optical flow-level mixing as in Eq. (7).

Generally, spatial-level intra-domain mixing is detailed in Eq. (5) and Eq. (6), while temporal-level mixing is specified in Eq. (7), where n denotes the training iteration, and t represents the timestamp of adjacent frames within each iteration.

2) Source to Source Mixing ($\mathcal{S}\to\mathcal{S}$): As illustrated in Fig. 4, the generation of the intra-mixed source domain follows a similar pattern to that of the target domain. To achieve this, we substitute $\mathcal{T}$ with $\mathcal{S}$ in Eq. (8) and replace $\mathcal{F}(\hat{Y}_{t,n-1}^{\mathcal{T}^*})$ with the source domain labels $Y_{t,n-1}^{\mathcal{S}^*}$ in Eq. (9), which leverage rich semantic segmentation annotations available in the source domain. It can be denoted as follows:

$$Z_{t,n}^{\mathcal{S}\to\mathcal{S}} = Z_{t,n}^{\mathcal{S}^*,k} + Z_{t,n}^{\mathcal{S}} \odot (1 - M_{t,n-1}^{\mathcal{S}^*,k}), \tag{10}$$

where $Z_{t,n}$ denotes *a unified concept for both video and image scenarios*, and similarly hereinafter.

Discussion. Through intra-domain mixing, we adaptively generate intra-mixed source ($\mathcal{S}\to\mathcal{S}$) and target ($\mathcal{T}\to\mathcal{T}$) domains. As feature visualization in Fig. 9(b) for video examples, the feature distribution of the latter one is closer to the source domain, and both intra-mixed domains exhibit more generalized feature embeddings. Ablation studies in Table VI demonstrate that our intra-domain mixing (Exp. ID 6) outperforms purely alternating training, one-way or two-way mixing across domains. We attribute this to the generalized

and continuous distribution of intra-mixed domains, which diminishes the feature discontinuity within naive domains, thus effectively facilitating knowledge transfer.

D. Inter-Mixed Video/Image Domain Adaptation

The feature gap bridging facilitated by intra-mixed domains prompts us to delve deeper into domain-wise interaction. In addition, existing methods [1]–[3], [5], [21]–[23] often focus purely on image-level mixing, neglecting the resulting semantic feature inconsistencies, which will impact precise dense prediction on the target domain. Given our design of video or image patch templates at the pixel level, could we further enhance feature consistency across quad-mixed domains at the feature level? Driven by this, building upon our intra-mixed domains, we propose more enhanced inter-mixed domains and feature-level mixing, thereby bridging domain gaps comprehensively and alleviating feature inconsistencies.

Formally, we term the outcome of *target to intra-mixed source mixing* as inter-mixed source domain, and vice versa.

1) Target to Intra-Mixed Source Mixing ($\mathcal{T}\rightarrow(\mathcal{S}\rightarrow\mathcal{S})$)*:* Following our design, the mixed domains should undergo feature-level interaction rather than being trained in isolation. To achieve this, we also utilize source and target patch templates to create more enhanced inter-mixed domains derived from our intra-domains, enabling the model to learn rich and invariant features across domains. The formula for the target to intra-mixed source domain mixing is as follows:

$$Z_{t,n}^{\mathcal{T}\rightarrow(\mathcal{S}\rightarrow\mathcal{S})} = Z_{t,n}^{\mathcal{T}^*,k} + Z_{t,n}^{\mathcal{S}\rightarrow\mathcal{S}} \odot (1 - M_{t,n-1}^{\mathcal{T}^*,k}). \tag{11}$$

For **image UDA-SS**, the generated inter-mixed source domain sample is a set that integrates spatial elements. While for **video UDA-SS**, it further involves temporal elements, akin to Eq. (1), which encompasses video frames, labels, and inter-frame optical flow. Notably, the inter-mixed source domain integrates semantics from both source and target domains, featured with a more diverse and enhanced feature embedding space, which can facilitate the domain gap bridging.

2) Source to Intra-Mixed Target Mixing ($\mathcal{S}\rightarrow(\mathcal{T}\rightarrow\mathcal{T})$) *:* The generation of the inter-mixed target domain follows a similar approach to Eq. (11):

$$Z_{t,n}^{\mathcal{S}\rightarrow(\mathcal{T}\rightarrow\mathcal{T})} = Z_{t,n}^{\mathcal{S}^*,k} + Z_{t,n}^{\mathcal{T}\rightarrow\mathcal{T}} \odot (1 - M_{t,n-1}^{\mathcal{S}^*,k}). \tag{12}$$

Compared to the large feature gap between naive source and target domains, the gap between the inter-mixed domains is narrowed effectively, as depicted in Fig. 9(a)(e). This facilitates the transfer of rich knowledge learned from the source domain to the target domain, promoting effective domain adaptation. The partial labels from source patch templates within the inter-mixed target domain could aid in the learning of global features. Notably, to enhance the diversity of patch templates, we ensure that categories of cross-domain templates are exclusive.

In this paper, we collectively refer to the two *inter-mixed* domains as *quad-mixed* domains because they are derived from intra-mixed domains and undergo quad-directional mixing.

3) Feature-Level Template Mixing: Based on the design that the generated quad-mixed domains encompass shared category-aware sample templates from both the source and target domains, we encourage the model to alleviate *feature inconsistencies* arising from domain-wise mixed-semantic contexts and prioritize learning domain-invariant features of them.

For **video UDA-SS**, to enhance the consistency learning of diverse spatio-temporal features, we perform feature-level template mixing across the generated quad-mixed domains adaptively, as depicted in Fig. 5(b). Specifically, feature-level mask sequence $M_{t,n-1}^{\mathcal{D}^*}\downarrow$ is generated to map spatio-temporal features from inter-mixed source domain samples, the mapping results are then fused with inter-mixed target domain features. The process is as follows:

$$\begin{aligned} F_{t,n}^{\mathcal{S}\rightarrow(\mathcal{T}\rightarrow\mathcal{T})} =& \psi(F_{t,n}^{\mathcal{T}\rightarrow(\mathcal{S}\rightarrow\mathcal{S})} \odot M_{t,n-1}^{\mathcal{D}^*}\downarrow, F_{t,n}^{\mathcal{S}\rightarrow(\mathcal{T}\rightarrow\mathcal{T})} \\ & \odot M_{t,n-1}^{\mathcal{D}^*}\downarrow) + F_{t,n}^{\mathcal{S}\rightarrow(\mathcal{T}\rightarrow\mathcal{T})} \odot (1 - M_{t,n-1}^{\mathcal{D}^*}\downarrow), \end{aligned} \tag{13}$$

where $F_{t,n}$ denotes the feature stack of adjacent frames, ψ denotes a 1×1 convolution layer, and the category-aware feature binary mask $M_{t,n-1}^{\mathcal{D}^*}\downarrow$ is downsampled from the integrated $M_{t,n-1}^{\mathcal{D}^*}$ using bilinear interpolation. The resolution of $M_{t,n-1}^{\mathcal{D}^*}\downarrow$ is same as F_t, and $M_{t,n-1}^{\mathcal{D}^*}$ is obtained as follows:

$$M_{t,n-1}^{\mathcal{D}^*} = M_{t,n-1}^{\mathcal{S}^*} \cup M_{t,n-1}^{\mathcal{T}^*}, \tag{14}$$

where $\mathcal{D}$ denotes the union of source $\mathcal{S}$ and target $\mathcal{T}$ domain.

For **image UDA-SS**, $F_{t,n}$ corresponding to spatial image features, following a parallel manner.

Discussion. From a unified view, we propose a fairly simple, efficient, and practical solution to bridge video/image feature gaps and alleviate feature inconsistencies, facilitated by quad-directional mixing at pixel and feature levels. The diverse and rich mixed-domain contexts aid in the domain gap mitigation, as illustrated in Fig. 3(d) and feature visualization in Fig. 9(e).

Notably, we only utilize quad-mixed data ($\mathcal{T}\rightarrow(\mathcal{S}\rightarrow\mathcal{S})$) and ($\mathcal{S}\rightarrow(\mathcal{T}\rightarrow\mathcal{T})$) for training. Compared to incrementally using ($\mathcal{S}\rightarrow\mathcal{S}+\mathcal{T}\rightarrow\mathcal{T}+\mathcal{S}\rightarrow\mathcal{T}+\mathcal{T}\rightarrow\mathcal{S}$), our method yields superior results while reducing computational costs. Detailed ablation analysis is provided in Table VI and Table XI.

E. Flow-Guided Feature Aggregation for Domain Alignment

To further mitigate domain gaps for **video UDA-SS**, we propose compressing video features into compact distribution for fine-grained alignment. Unlike only aligning spatial features of single frames [2], we fully leverage temporal cues to generate cross-frame consistent target pseudo-labels, maintain temporal consistency of spatial features for each domain, and aggregate spatio-temporal features guided by optical flow for precise domain alignment, thus enhancing segmentation for the target domain. Note that it is extendable to **image UDA-SS**.

1) Flow-Guided Spatio-Temporal Feature Extraction: In this paper, we employ a variant of ACCEL [37] as the video segmentation network, which involves two weight-shared CNN/transformer-based segmentation branches and an optical flow branch. The segmentation branches capture the spatial feature, taking stacked adjacent frames $X_t = \mathbb{S}\{x_{t-1}, x_t\}$ as the input. The optical flow $o_{t-1\rightarrow t}$ is used to propagate

temporal information across frames. We conduct feature-level template mixing proposed in Section III-D3 during forward training across quad-mixed domain samples.

For clarity, we illustrate with pure source and target domain, showing that the inter-mixed source and target domain also adhere to the same paradigm. During model training, the deep feature $F_t = \mathbb{S}\{f_{t-1}, f_t\}$ of stacked adjacent frames are fused to generate the holistic feature of frame x_t, leveraging the video temporal continuity. Specifically, we feed feature f_t and optical flow-warped feature $f_{t-1\to t}$ into a convolutional fusion layer ϕ_f to generate fused feature vector $\bar{f}_{t-1\to t}$. The process is described as follows:

$$\bar{f}_{t-1\to t} = \phi_f(f_{t-1\to t},\ f_t), \tag{15}$$

where the optical flow-warped feature $f_{t-1\to t}$ is obtained by:

$$f_{t-1\to t} = \phi_{\mathcal{W}}(f_{t-1}, o_{t-1\to t}), \tag{16}$$

where $\phi_{\mathcal{W}}$ is a warp operation using bilinear interpolation.

For target domain self-supervised training, we fully exploit inherent temporal continuity to generate cross-frame coherent pseudo-labels. As shown in the left part of Fig. 2(a), consecutive sets $Z^{\mathcal{T}}_{t-\tau}$ are fed into the segmentation network to generate the feature $\bar{f}^{\mathcal{T}}_{t-\tau-1\to t-\tau}$. Then, we yield pseudo-label $\hat{y}^{\mathcal{T}}_t$ with optical flow $o^{\mathcal{T}}_{t-\tau\to t}$ warpping, as described below:

$$\hat{y}^{\mathcal{T}}_t = \arg\max_k(\phi_{\mathcal{W}}(\phi_{\mathcal{D}}(\bar{f}^{\mathcal{T},(k)}_{t-\tau-1\to t-\tau})), o^{\mathcal{T},(k)}_{t-\tau\to t}), \tag{17}$$

where $\phi_{\mathcal{D}}$ denotes the segmentation decoder layer. For target domain testing, as shown in Fig. 2(c), only stacked sets $Z^{\mathcal{T}}_t$ are input to the model to obtain segmentation results.

2) Spatio-Temporal Feature Aggregation for Domain Alignment: To further mitigate the video domain gap, we propose adaptive aggregation along spatial and temporal dimensions to reduce the distribution discrepancy of scattered video features. Unlike previous video UDA-SS methods [2], [45] only emphasizing image feature discrepancy across domains that may overlook holistic temporal learning, we comprehensively consider spatial features of consecutive frames and fully leverage temporal information. In the spatial dimension, fused features $p^{\mathcal{T}}_{t'\to t}$ that integrate flow-guided features from previous timestamps t' to frame t are adaptively aggregated. For the target domain, we set $t' \in \mathbb{T} = \{t-1, t-\tau, t-\tau-1\}$, while for the source domain, $t' \in \mathbb{T} = \{t-1\}$, benefiting from source label supervision to reduce computational cost.

Specifically, as shown in Fig. 6, we warp frame features for two domains from the previous time step t' to t following Eq. (16) to get the warped feature $f_{t'\to t}$, respectively. Then, upon merging $f_{t'\to t}$ with feature f_t through the fusion layer ϕ_f, we generate temporal consistent spatial feature $\bar{f}^{\mathcal{T}}_{t'\to t}$ for frame t. Based on this, we adaptively distribute the spatial feature to category-aware subspaces by pixel mapping operation in a fine-grained manner. For the target domain, we obtain spatial-aggregated feature $\tilde{f}^{\mathcal{T},(k)}_{t'\to t}$ defined as follows:

$$\tilde{f}^{\mathcal{T},(k)}_{t'\to t} = \frac{\sum_{i=1}^{H\times W}(\bar{f}^{\mathcal{T},(i,k)}_{t'\to t} \odot \mathbb{1}(\hat{y}^{\mathcal{T},(i,k)}_{t'\to t} == k))}{\sum_{i=1}^{H\times W}\mathbb{1}(\hat{y}^{\mathcal{T},(i,k)}_{t'\to t} == k)}, \tag{18}$$

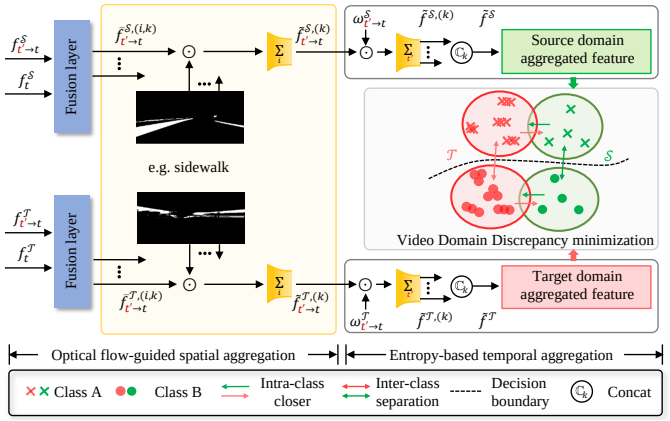


Fig. 6. The illustration of the optical flow-guided spatio-temporal feature aggregation for domain alignment, where t' signifies previous time step, and $\omega^{\mathcal{T}}_{t'\to t}$ denotes temporal aggregation weights of target frames. The symbol $\bar{f}_{t'\to t}$ denotes the optical flow-guided frame feature warped from the previous time step t', where "$\to$" denotes the warp direction along the time dimension.

where k belongs to the whole category space. And source spatial-aggregated feature $\tilde{f}^{\mathcal{S},(k)}_{t'\to t}$ follows a similar manner. In this way, adaptive aggregation along the spatial dimension across two domains is conducted in a fine-grained way.

Based on this, we performed temporal aggregation by combining spatially aggregated features from consecutive frames to generate spatial-temporal features $\tilde{f}^{\mathcal{T},(k)}$ using entropy confidence [53] $\omega_{t'\to t}$, for the target domain:

$$\tilde{f}^{\mathcal{T},(k)} \leftarrow \sum_{t'\in\mathbb{T}} \tilde{f}^{\mathcal{T},(k)}_{t'\to t} \cdot \omega^{\mathcal{T}}_{t'\to t}, \tag{19}$$

where $\omega^{\mathcal{T}}_{t'\to t}$ is obtained by softmax normalization of logits entropy among the fused spatial features, formulated as:

$$\omega^{\mathcal{T}}_{t'\to t} = \underset{t'}{\text{softmax}}(-\frac{1}{H\times W}\sum_{i}^{H\times W}(-\bar{f}^{\mathcal{T},(i)}_{t'\to t}\cdot\log(\bar{f}^{\mathcal{T},(i)}_{t'\to t}))), \tag{20}$$

where the segmentation confidence and temporal weight $\omega^{\mathcal{T}}_{t'\to t}$ are promoted to be positively correlated.

Subsequently, we concatenate the aggregated features of each category to generate fine-grained features $\tilde{f}^{(k)}$ for each domain, where the concatenation is denoted as $\mathbb{C}_k$. We can mitigate the fine-grained video feature gap using MMD (Maximum Mean Discrepancy) [8] minimization, thanks to the optical flow-guided temporal information propagation and distribution alignment.

As for **image UDA-SS**, spatial feature aggregation can be used for fine-grained alignment without relying on temporal cues, i.e., optical flow guidance.

F. Model Training

The model is trained jointly using cross-domain samples, with the objective function defined as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{QuadMix} + \mathcal{L}_{Agg} + \mathcal{L}_{SSL}. \tag{21}$$

Each loss is discussed first for **video UDA-SS** as below.

TABLE I

VIDEO UDA-SS RESULTS ON SYNTHIA-SEQ → CITYSCAPES-SEQ. THE BEST RESULTS ARE IN **BOLD**, AND THE SUBOPTIMAL RESULTS ARE IN UNDERLINED. "ADV.": ADVERSARIAL LEARNING, "SSL": SELF-SUPERVISED LEARNING, "WARM": WARM-UP, "VIT": VISION TRANSFORMER-BASED NETWORK. † DENOTES RETRAINED RESULTS OF IMAGE DOMAIN METHODS USING A VIDEO SEGMENTATION BACKBONE [37] FOR FAIR COMPARISON.

Methods	Adv.	SSL	Warm	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† [54]	✓			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† [55]		✓		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† [56]	✓			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† [13]		✓		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† [57]				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† [58]	✓			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† [59]		✓		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† [60]		✓		84.1	32.8	67.6	28.1	5.5	20.3	61.1	64.8	43.1	19.0	70.6	45.2
SAC† [10]		✓		87.0	41.1	64.0	20.4	12.1	32.8	38.2	47.6	53.1	19.3	81.1	45.2
DACS† [21]		✓		86.4	40.0	74.0	27.8	9.5	28.2	71.6	72.0	55.6	20.0	76.4	51.0
PixMatch† [61]		✓		90.2	49.9	75.1	23.1	17.4	34.2	67.1	49.9	55.8	14.0	84.3	51.0
DA-VSN [16]	✓			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 [45]		✓		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 [17]		✓		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 [19]		✓		90.9	32.5	76.8	28.6	6.0	36.7	76.0	78.9	51.7	13.8	85.6	52.5
TPL-SFC [19]		✓		90.0	32.8	80.4	28.9	14.9	35.3	80.8	81.1	57.5	19.6	86.7	55.3
PAT [20]		✓		91.5	41.3	76.1	29.6	20.9	33.8	72.4	75.9	51.3	24.7	86.2	54.9
CMOM [2]		✓	✓	90.4	39.2	82.3	30.2	16.3	29.6	83.2	84.9	59.3	19.7	84.3	56.3
QuadMix		✓		90.5	41.2	81.1	29.3	23.1	47.5	82.7	83.8	61.1	28.4	87.0	59.6
QuadMix		✓	✓	90.8	39.9	**83.2**	33.2	30.1	50.7	84.8	82.3	61.2	32.7	87.4	61.5
QuadMix (ViT)		✓		**94.1**	**61.9**	82.9	**36.9**	**41.0**	**59.1**	**85.2**	**85.6**	**64.3**	**37.8**	**90.3**	**67.2**

QuadMix loss. As shown in Fig. 5, the loss for quad-directional mixing used for gap bridging is as follows:

$$\begin{aligned}\mathcal{L}_{QuadMix} = &\mathcal{L}_{CE}^{\mathcal{S}'}(\mathcal{G}(X_t^{\mathcal{S}'}, o_{t-1\to t}^{\mathcal{S}'}), y_t^{\mathcal{S}'}) \\ &+ \lambda_{\mathcal{T}}\mathcal{L}_{CE}^{\mathcal{T}'}(\mathcal{G}(\mathcal{A}(X_t^{\mathcal{T}'}), o_{t-1\to t}^{\mathcal{T}'}), \hat{y}_t^{\mathcal{T}'})\end{aligned}, \tag{22}$$

where $\mathcal{S}'$ and $\mathcal{T}'$ indicate two quad-mixed domains for simplicity, $\mathcal{G}$ denotes the semantic segmentation model, $\mathcal{A}$ means image enhancement, and $\mathcal{L}_{CE}$ denotes cross-entropy loss.

Aggregated feature alignment loss. We define the following loss to minimize the discrepancy of aggregated spatio-temporal feature distribution across domains:

$$\mathcal{L}_{Agg} = \lambda_f \left\| \mathbb{E}\left(\mathbb{C}_k\left(f^{\mathcal{S},(k)}\right)\right) - \mathbb{E}\left(\mathbb{C}_k\left(f^{\mathcal{T},(k)}\right)\right) \right\|_{\mathcal{H}}^2 \tag{23}$$

where $\mathbb{E}$ denotes the mathematical expectation, and $\mathcal{H}$ represents the reproducing Hilbert space [62].

Self-supervised learning loss. We also employ a cross-domain self-supervised learning loss for two video domains:

$$\begin{aligned}\mathcal{L}_{SSL} = &\mathcal{L}_{CE}^{\mathcal{S}}(\mathcal{G}(X_t^{\mathcal{S}}, o_{t-1\to t}^{\mathcal{S}}), y_t^{\mathcal{S}}) \\ &+ \lambda_{\mathcal{T}}\mathcal{L}_{CE}^{\mathcal{T}}(\mathcal{G}(\mathcal{A}(X_t^{\mathcal{T}}), o_{t-1\to t}^{\mathcal{T}}), \hat{y}_t^{\mathcal{T}})\end{aligned}. \tag{24}$$

For **image UDA-SS**, the input of the model $\mathcal{G}$ in Eq. (22) and Eq. (24) will be replaced by a single image x_t. The sensitivity analysis for hyperparameters $\lambda_{\mathcal{T}}$ and λ_f will be discussed in the experiments. The training algorithm is summarized in supplementary materials.

IV. EXPERIMENTS

A. Experimental Setups

1) Video UDA-SS Datasets: We conduct experiments on two challenging benchmarks: Synthia-Seq [63] → Cityscapes-Seq [64] and VIPER [65] → Cityscapes-Seq. **Cityscapes-Seq** (target domain) is a standard dataset for semantic urban scene understanding, featuring real-world videos from 50 cities in Germany and neighboring countries. It comprises 2,975 training video clips and 500 validation video clips, and each clip contains continuous 30 frames. **Synthia-Seq** (source domain) contains 8,000 photo-realistic frames with dense segmentation labels. We employ video frames in the *RGB* folder and choose 11 categories common with Cityscapes-Seq for adaptation. **VIPER** (source domain) is another standard synthetic video dataset, consisting of 133,670 frames of virtual videos rendered from urban landscape scenes in the virtual computer game "Grand Theft Auto V". We use 13,367 frames with segmentation labels as another source domain, choosing 15 common categories with Cityscapes-Seq for adaptation. Notably, our dataset setting is consistent with previous works [2], [16]–[20], [45], etc.

2) Image UDA-SS Datasets: We conduct experiments on two popular benchmarks GTAV [66] → Cityscapes and SYNTHIA [63] → Cityscapes. **Cityscapes** (target domain) is a subset of Cityscapes-Seq, with each image being the 20-th frame of each corresponding video clip. **GTAV** (source domain) is a synthetic image dataset with labels for semantic segmentation, where images are also collected from "Grand Theft Auto V." It comprises 19 categories common with Cityscapes and 24,966 labeled images. **SYNTHIA** (source domain) is another synthetic image dataset, we use images from the *SYNTHIARAND-CITYSCAPES* folder which contains 9,400 images and 16 common categories with Cityscapes, akin to previous works [3], [5], [21]–[23], [67], [68], etc.

3) Implementation Details: Our model is implemented in PyTorch and trained on a single NVIDIA Tesla V100 GPU.

TABLE II
VIDEO UDA-SS RESULTS ON VIPER → CITYSCAPES-SEQ. THE BEST RESULTS ARE IN **BOLD**, AND THE SUBOPTIMAL RESULTS ARE IN UNDERLINED. "ADV.": ADVERSARIAL LEARNING, "SSL": SELF-SUPERVISED LEARNING, "WARM": WARM-UP, "VIT": VISION TRANSFORMER-BASED NETWORK. † DENOTES RETRAINED RESULTS OF IMAGE DOMAIN METHODS USING A VIDEO SEGMENTATION BACKBONE [37] FOR FAIR COMPARISON.

Methods	Adv.	SSL	Warm	Road	Side.	Buil.	Fenc.	Light	Sign	Vege.	Terr.	Sky	Pers.	Car	Truc.	Bus	Mot.	Bike	mIOU
Source-Only	✓			56.7	18.7	78.7	6.0	22.0	15.6	81.6	18.3	80.4	55.9	66.3	4.5	16.8	20.4	10.3	37.1
AdvEnt† [54]	✓			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† [55]		✓		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† [56]	✓			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† [13]		✓		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† [57]				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† [58]	✓			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† [59]		✓		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† [60]		✓		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
DACS† [21]		✓		69.6	24.1	76.9	9.1	16.1	15.3	74.1	20.3	76.5	59.4	74.8	38.6	43.1	7.7	1.9	40.5
SAC† [10]		✓		52.2	19.6	73.4	3.7	23.1	25.2	73.9	17.3	78.1	56.9	80.3	38.3	48.2	17.8	14.1	41.5
PixMatch† [61]		✓		79.4	26.1	84.6	16.6	28.7	23.0	85.0	30.1	83.7	58.6	75.8	34.2	45.7	16.6	12.4	46.7
DA-VSN [16]	✓			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
TPS [17]		✓		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
MoDA [18]		✓		72.2	25.9	80.9	18.3	24.6	21.1	79.1	23.2	78.3	**68.7**	84.1	43.2	49.5	28.8	38.6	49.1
I2VDA [45]		✓		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
SFC [19]		✓		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
TPL-SFC [19]		✓		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
PAT [20]		✓		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
CMOM [2]		✓	✓	89.0	**53.8**	86.8	**31.0**	32.5	**47.3**	85.6	25.1	80.4	65.1	79.3	21.6	43.4	25.7	40.6	53.8
QuadMix		✓		91.4	43.5	85.5	21.3	26.6	40.2	84.7	34.2	84.6	62.1	84.3	39.9	46.6	33.5	45.6	54.9
QuadMix		✓	✓	**91.6**	51.4	87.0	24.1	32.3	37.2	84.1	28.4	84.8	64.4	85.7	41.4	46.5	34.0	49.6	56.2
QuadMix (ViT)		✓		87.3	43.8	**87.3**	25.2	**40.0**	36.9	**86.7**	20.8	**90.3**	65.8	**86.8**	**48.6**	**65.6**	37.6	**49.7**	**58.2**

The details are as follows: (i) As for the network, we employ a variant of ACCEL [37] for video UDA-SS, using DeepLab-V2 [69] or MiT-B5 in DAFormer [1] as the segmentation branch, and FlowNet [70] as the optical flow branch. For image UDA-SS, we only employ the segmentation branch. Notably, DeepLab-V2 uses ResNet-101 [71] as its backbone, and both DeepLab-V2 and MiT-B5 are pre-trained on ImageNet [72]; (ii) As for input resolution, for video tasks, we resize the resolution of Cityscapes-Seq, Synthia-Seq, and VIPER to 1024×512, 760×1280, and 720×1280, respectively. While for image tasks, images undergo resizing to dimensions of 1024×2048, 1440×2560, and 1520×2560 for Cityscapes, GTAV, and SYNTHIA; (iii) We set batch size B to 1 for video and 2 for image scenarios; (iv) For DeepLab-V2, we use the SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The initial learning rate is 2.5×10^{-4} for Synthia-Seq to Cityscapes-Seq and two image benchmarks, and 1×10^{-4} for VIPER to Cityscapes-Seq. For transformer networks, we use an AdamW optimizer with a weight decay of 0.01 and betas set to (0.9, 0.999). The learning rate is initially set to 6×10^{-5} for the encoder and 6×10^{-4} for the decoder, with a linear warmup followed by linear decay; (v) Our models are trained for a maximum of 40k iterations for four benchmarks; (vi) We apply random Gaussian blur and color jitter to the target domain for image/video frame enhancement; (vii) Aligned with [2], [16], [17], we set video temporal deviation τ to 1; (viii) Additionally, the confidence threshold for target filtering operation $\mathcal{F}$ is 0.9; (ix) During quad-directional mixing, source videos undergo random cropping to match the resolution of target videos; (x) For image tasks, we set $\lambda_{\mathcal{T}}$ and λ_f to 1 for simplicity; (xi) In this paper, we use per-category IOU (Intersection over Union) and mIOU (mean IOU) as evaluation metrics. Experiment settings are consistent with prior works for fair comparison.

B. Experimental Results

1) Comparison With State-of-the-Arts of Video UDA-SS: In Table I and Table II, we compare our method with the state-of-the-art methods on Synthia-Seq → Cityscapes-Seq and VIPER → Cityscapes-Seq benchmark. To the best of our knowledge, recent methods [2], [16]–[20], [45] are the most superior works in video UDA-SS, all employing CNN-based architectures. Notably, for the first time, we explore **transformer-based architectures** for video UDA-SS. Moreover, we also compare several variants to video UDA-SS that are initially designed for image UDA-SS, including works [10], [13], [21], [54]–[61]. Technically, following the previous video UDA-SS works [2], [16], [17], [45], we replace their image segmentation backbones with a video segmentation network (denoted by †). Additionally, "Adv." signifies methods that employ adversarial learning, and "SSL" denotes self-supervised learning. "Warm" denotes initializing parameters of the DeepLab-V2 model with a warm-up model, whereas others use ResNet-101 pre-trained on ImageNet. "ViT" refers to the transformer backbone.

Synthia-Seq → Cityscapes-Seq. In Table I, we present experimental results comparing our method with state-of-the-art techniques on the Synthia-Seq → Cityscapes-Seq. For clarity, the best per-category IOU and mIOU results are in bold, while suboptimal results are underlined. We observed that: (i) Our method is highly compatible with both CNN and transformer-based architectures. The proposed CNN-based Quadmix significantly outperforms the previous best CMOM [2] by 5.2%

TABLE III
IMAGE UDA-SS RESULTS ON GTAV → CITYSCAPES USING DEEPLAB-V2 (DL-V2) AND VISION TRANSFORMER (VIT). THE BEST RESULTS USING VIT ARE IN **BOLD**, AND THE BEST RESULTS USING DL-V2 ARE IN UNDERLINED.

Methods	Network	Road	Side.	Buil.	Wall	Fence	Pole	Light	Sign	Vege.	Terr.	Sky	Pers.	Rider	Car	Truck	Bus	Train	Moto.	Bike	mIOU
Source-Only	DL-V2	63.3	15.7	59.4	8.6	15.2	18.3	26.9	15.0	80.5	15.3	73.0	51.0	17.7	59.7	28.2	33.1	3.5	23.2	16.7	32.9
PixMatch [61]	DL-V2	91.6	51.2	84.7	37.3	29.1	24.6	31.3	37.2	86.5	44.3	85.3	62.8	22.6	87.6	38.9	52.3	0.7	37.2	50.0	50.3
RDA [59]	DL-V2	93.1	55.0	84.7	33.1	29.5	38.7	49.3	44.9	84.8	41.6	80.2	62.3	33.2	85.6	37.3	51.3	18.5	34.6	45.3	52.8
DACS [21]	DL-V2	89.9	39.7	87.9	39.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
SAC [10]	DL-V2	90.4	53.9	86.6	42.4	27.3	45.1	48.5	42.7	87.4	40.1	86.1	67.5	29.7	88.5	49.1	54.6	9.8	26.6	45.3	53.8
DSP [3]	DL-V2	92.4	48.0	87.4	33.4	35.1	36.4	41.6	46.0	87.7	43.2	89.8	66.6	32.1	89.9	57.0	56.1	0.0	44.1	57.8	55.0
BDM [5]	DL-V2	91.3	51.8	86.7	49.9	49.2	53.3	43.1	43.3	85.5	47.9	85.7	62.3	45.9	87.8	55.5	54.4	4.4	46.3	50.4	57.6
I2F [67]	DL-V2	90.8	48.7	85.2	30.6	28.0	33.3	46.4	40.0	85.6	39.1	88.1	61.8	35.0	86.7	46.3	55.6	11.6	44.7	54.3	53.3
ADPL [22]	DL-V2	93.4	60.6	87.5	45.3	32.6	37.3	43.3	55.5	87.2	44.8	88.0	64.5	34.2	88.3	52.6	61.8	49.8	41.8	59.4	59.4
SePiCo [23]	DL-V2	95.2	67.8	88.7	41.4	38.4	43.4	55.5	63.2	88.6	46.4	88.3	73.1	49.0	91.4	63.2	60.4	0.0	45.2	60.0	61.0
FREDOM [68]	DL-V2	90.9	54.1	87.8	44.1	32.6	45.2	51.4	57.1	88.6	42.6	89.5	68.8	40.0	89.7	58.4	62.6	55.3	47.7	40.0	61.3
QuadMix	DL-V2	97.1	78.0	90.4	49.7	40.3	53.4	61.7	70.9	90.7	49.7	92.9	77.9	53.9	93.5	72.4	65.9	0.7	60.5	68.5	66.8
DAFormer [1]	ViT	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
SePiCo [23]	ViT	96.7	76.7	89.7	55.5	49.5	53.2	60.0	64.5	90.2	50.3	90.8	74.5	44.2	93.3	77.0	79.5	63.6	61.0	65.3	70.3
FREDOM [68]	ViT	96.7	74.8	90.9	58.1	49.0	57.5	63.4	**71.4**	91.6	52.1	**94.4**	78.4	53.1	94.1	83.9	85.2	72.5	62.8	**68.9**	73.6
QuadMix	ViT	**97.5**	**80.9**	**91.6**	**62.3**	**57.6**	**58.2**	**64.5**	71.2	**91.7**	**52.3**	94.3	**80.0**	**55.9**	**94.6**	**86.3**	**90.5**	**82.3**	**65.1**	68.1	**76.1**

TABLE IV
IMAGE UDA-SS RESULTS ON SYNTHIA → CITYSCAPES USING DEEPLAB-V2 (DL-V2) AND VISION TRANSFORMER (VIT). THE BEST RESULTS USING VIT ARE IN **BOLD**, AND THE BEST RESULTS USING DL-V2 ARE IN UNDERLINED. MIOU* DENOTES 13-CATEGORY ADAPTATION WITHOUT THREE CLASSES MARKED BY *, WHILE MIOU MEANS 16-CATEGORY ADAPTATION.

Methods	Network	Road	Side.	Buil.	Wall*	Fence*	Pole*	Light	Sign	Vege.	Sky	Pers.	Rider	Car	Bus	Moto.	Bike	mIOU	mIOU*
Source-Only	DL-V2	36.3	14.6	68.8	9.2	0.2	24.4	5.6	9.1	69.0	79.4	52.5	11.3	49.8	9.5	11.0	20.7	33.7	29.5
PixMatch [61]	DL-V2	92.5	54.6	79.8	4.8	0.1	24.1	22.8	17.8	79.4	76.5	60.8	24.7	85.7	33.5	26.4	54.4	46.1	54.5
RDA [59]	DL-V2	89.4	39.0	79.8	8.1	1.2	33.0	22.6	28.1	81.8	82.0	60.9	30.1	82.7	41.4	37.5	48.9	47.9	55.7
DACS [21]	DL-V2	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	90.8	67.6	38.3	82.9	38.9	28.5	47.6	48.3	54.8
SAC [10]	DL-V2	89.3	47.2	85.5	26.5	1.3	43.0	45.5	32.0	87.1	89.3	63.6	25.4	86.9	35.6	30.4	53.0	52.6	59.3
DSP [3]	DL-V2	86.4	42.0	82.0	2.1	1.8	34.0	31.6	33.2	87.2	88.5	64.1	31.9	83.8	64.4	28.8	54.0	51.0	59.9
BDM [5]	DL-V2	91.0	55.8	86.9	-	-	-	58.3	44.7	85.8	85.7	84.1	40.3	86.0	55.2	45.0	50.6	-	66.8
I2F [67]	DL-V2	84.9	44.7	82.2	9.1	1.9	36.2	42.1	40.2	83.8	84.2	68.9	35.3	83.0	49.8	30.1	52.4	51.8	60.1
ADPL [22]	DL-V2	86.1	38.6	85.9	29.7	1.3	36.6	41.3	47.2	85.0	90.4	67.5	44.3	87.4	57.1	43.9	51.4	55.9	63.6
SePiCo [23]	DL-V2	77.0	35.3	85.1	23.9	3.4	38.0	51.0	55.1	85.6	80.5	73.5	46.3	87.6	69.7	50.9	66.5	58.1	66.5
FREDOM [68]	DL-V2	86.0	46.3	87.0	33.3	5.3	48.7	38.1	46.8	87.1	89.1	71.2	38.1	87.1	54.6	51.3	59.9	59.1	66.0
QuadMix	DL-V2	88.5	52.9	87.1	3.7	1.4	56.2	62.7	59.2	87.2	89.0	79.1	55.8	87.9	61.7	58.1	71.2	62.6	72.3
DAFormer [1]	ViT	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	89.8	73.2	48.2	87.2	53.2	53.9	61.7	60.9	67.4
SePiCo [23]	ViT	87.0	**52.6**	88.5	40.6	**10.6**	49.8	57.0	55.4	56.8	86.2	75.4	52.7	**92.4**	78.9	53.0	62.6	64.3	71.4
FREDOM [68]	ViT	**89.4**	50.8	**89.3**	**48.8**	9.3	57.3	**65.1**	60.1	**89.9**	93.7	79.4	51.6	90.5	66.0	62.3	**68.1**	67.0	73.6
QuadMix	ViT	88.1	51.2	88.9	46.7	7.9	**58.6**	64.7	**63.7**	88.1	**93.9**	**81.3**	**56.6**	90.3	**66.9**	**66.8**	66.0	**67.5**	**74.3**

mIOU. When equipped with a transformer, we further boost mIOU by 5.7%, achieving the highest result of 67.2%; (ii) For CNN-based QuadMix, without a warm-up model, we achieved 59.6% mIOU, which is 4.3% higher than TPL-SFC [19] by a large margin. When using a warm-up model, we outperform the CMOM, which relies heavily on the warm-up model in [16] for both model initialization and pseudo-label generation; (iii) Adversarial learning methods can facilitate domain adaptation but are less improving compared to self-supervised methods; (iv) Our method boosts the IOU for most categories. It can be attributed to our comprehensive quad-directional mixing, which bridges the feature gap effectively and addresses the intra-domain discontinuity and feature inconsistencies, as well as flow-guided video feature aggregation for cross-domain alignment.

Viper → Cityscapes-Seq. The domain gap between Viper and Cityscapes-Seq is much larger than that of Synthia-Seq and Cityscapes-Seq. As shown in Table II, our proposed QuadMix also achieves state-of-the-art results. We can conclude that: (i) For CNN-based QuadMix, we achieve 56.2% mIOU with a warm-up model and 54.9% mIoU without it; furthermore, we obtain 58.2% mIOU when we employ transformer-based network. All our results significantly outperform previous methods; (ii) Recent work I2VDA [45] focuses exclusively on transferring source spatial information to target domain videos, resulting in a target domain mIOU of 51.2%. This is inferior to our model, which comprehensively investigates both spatial and temporal level domain shifts. (iii) Variants of image domain adaptation methods generally underperform on target videos compared to video UDA-SS methods, likely due

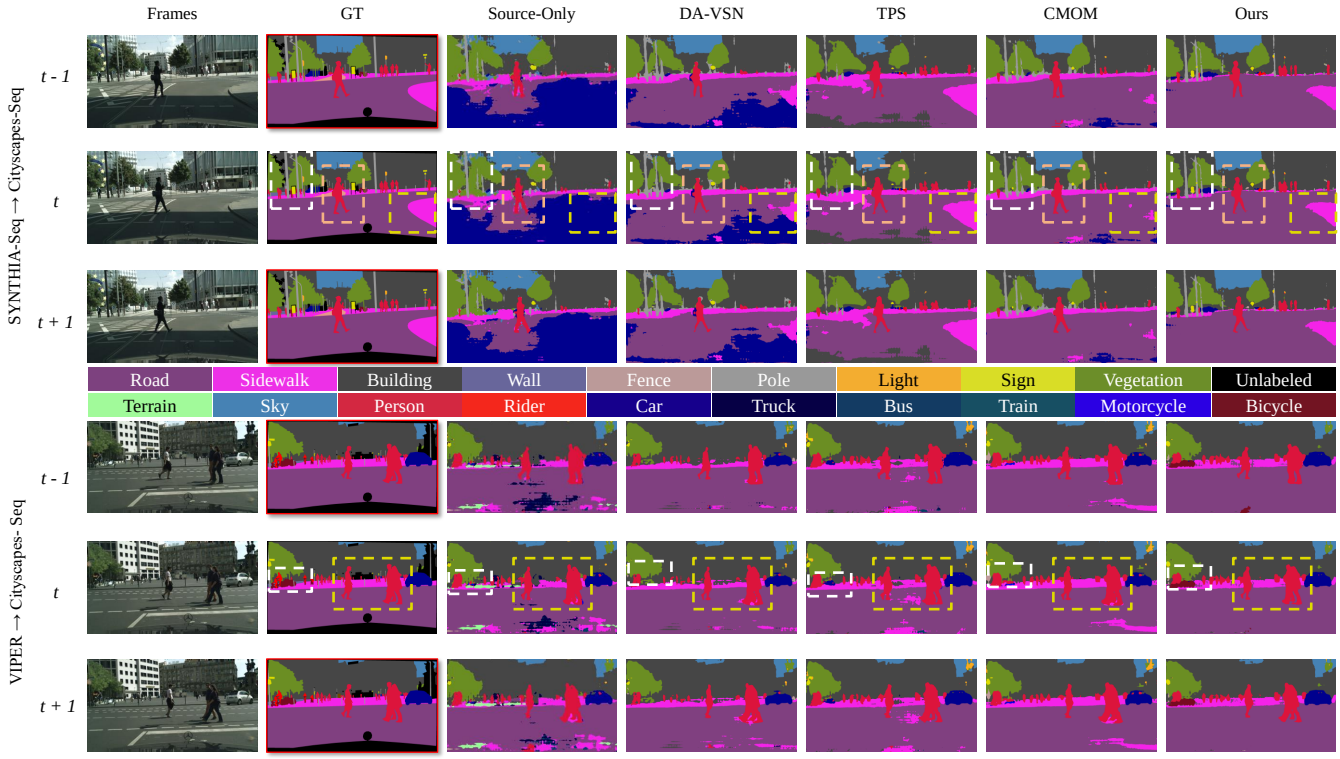


Fig. 7. Qualitative results for three consecutive frames from the Cityscapes-Seq validation set. The results are presented in the following order: target image, ground truth, semantic segmentation results from the source-only model, DA-VSN [16], TPS [17], CMOM [2], and the proposed method. We achieve the highest segmentation accuracy, exhibiting smoother edges and refined details. Please zoom in for details.

to inadequate utilization of crucial temporal information.

Qualitative results comparison on two benchmarks. In Fig. 7, we present adjacent three frames and corresponding qualitative segmentation results of our method along with the source-only model, DA-VSN [16], TPS [17], and CMOM [2] on Cityscape-Seq validation dataset. Notably, it only offers a ground-truth label for the 19-th frame of each 30-frame video clip, so we replicate the label from frame t for frames $t-1$ and $t+1$, marked with red borders for clarity. Our results exhibit smoother edges and refined details. For example, on Synthia-Seq $\rightarrow$ Cityscapes-Seq benchmark, we obtain more accurate segmentation for categories like *person*, *sign*, and *sidewalk*. Similarly, in the case of Viper $\rightarrow$ Cityscapes-Seq, our model predicts contours that closely resemble actual *bicycle* and *person*, indicating a significant improvement. Please refer to the supplementary video for more visual comparisons.

2) Comparison With State-of-the-Arts of Image UDA-SS: For image UDA-SS, we replace spatio-temporal QuadMix with spatial QuadMix, and substitute flow-guided spatio-temporal feature aggregation with spatial feature aggregation. Technically, we shift elements of $Z_{t,n}^{\mathcal{S}^*,k}$ and $Z_{t,n}^{\mathcal{T}^*,k}$ for images defined in Section III-B, and remove the reliance on temporal cues that are not present in image scenarios. Additionally, target pseudo-labels are generated using a momentum network [52].

Table III and Table IV compare our method with existing image UDA-SS approaches. Impressively, we maintain superiority on two challenging benchmarks: GTAV $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes. We compare with recent advanced CNN-based methods [3], [5], [10], [21]–[23], [59], [61], [67], [68] and transformer-based methods [1], [23], [68]. Notably, DACS [21] proposes one-way mixing, which is followed by subsequent works [1], [23], [68] combing contrastive learning or long-tail learning. Other works [5], [22] employ two-way mixing, with the latter integrating style transfer. However, as analyzed above, they overlooked the intra-domain discontinuity within each domain, leading to fragmented and insufficient domain gap bridging. Additionally, their focus on pixel-level mixing alone can result in feature inconsistencies, hindering effective knowledge transfer. In contrast, we address these challenges comprehensively and obtain remarkable results.

GTAV $\rightarrow$ Cityscapes. As shown in Table III, our method achieves mIOUs of 66.8% with DeepLab-V2 and 76.1% with transformer networks. Notably, QuadMix (ViT) more effectively extracts spatial features from long-tail categories. Specifically, for the *train* class, our method achieves an IOU of 82.3%, surpassing that of FREDOM [68] (72.5%), which is tailored for long-tail learning. Notably, QuadMix (DL-V2) underperforms for the *train* class due to the limited feature extraction capability of CNNs. We believe that more frequent resamples for models to learn such categories could improve the results. Generally, our method boosts the performance of most categories, exhibiting robust improvement.

SYNTHIA $\rightarrow$ Cityscapes. In Table IV, we present results for 13-category (mIOU*) and 16-category (mIOU) adaptation, following [22], [23], etc. BDM [5] only studies 13-category adaption. Note that QuadMix (DL-V2) has already outperformed DAFormer (ViT) by 1.7% in 13-category adaptation and 4.9% in 16-category adaptation, with further improvements of 4.9% and 2.0%, respectively, when equipped with

TABLE V
ABLATION STUDIES ON SYNTHIA-SEQ → CITYSCAPES-SEQ. ALL MODELS ARE TRAINED END-TO-END FOR A TOTAL OF 40K ITERATIONS. "V-TEMPLATE" REPRESENTS OUR VIDEO PATCH TEMPLATE AS DEFINED IN EQ. (1). "F-TEMPLATE" DENOTES FEATURE-LEVEL TEMPLATE MIXING ACROSS THE GENERATED QUAD-MIXED DOMAINS, "AGG" REFERS TO THE OPTICAL-GUIDED VIDEO FEATURE AGGREGATION, "WARM" MEANS INITIALIZING MODEL PARAMETERS USING A WARM-UP MODEL.

ID	V-template	F-template	Agg.	Warm	mIOU (Gain)
0*					52.01
1	✓				55.71 (+3.70)
2	✓	✓			57.25 (+5.24)
3			✓		54.01 (+2.00)
4	✓	✓	✓		59.62 (+7.61)
5	✓	✓		✓	57.95 (+5.94)
6			✓	✓	55.97 (+3.96)
7	✓		✓	✓	60.26 (+8.25)
8	✓	✓	✓	✓	**61.48 (+9.47)**

TABLE VI
ABLATION STUDIES OF THE QUAD-DIRECTIONAL MIXING PATHS ON SYNTHIA-SEQ → CITYSCAPES-SEQ.

ID	$\mathcal{S}\to\mathcal{T}$	$\mathcal{T}\to\mathcal{S}$	$\mathcal{S}\to\mathcal{S}$	$\mathcal{T}\to\mathcal{T}$	$\mathcal{S}\to(\mathcal{T}\to\mathcal{T})$	$\mathcal{T}\to(\mathcal{S}\to\mathcal{S})$	mIOU (Gain)
CMOM	✓						56.31
0*							55.97
1	✓						58.11 (+2.14)
2		✓					57.84 (+1.87)
3	✓	✓					58.77 (+2.80)
4			✓				58.98 (+3.01)
5				✓			56.67 (+0.70)
6			✓	✓			59.24 (+3.27)
7	✓			✓			58.39 (+2.42)
8		✓	✓				59.55 (+3.58)
9	✓		✓				60.20 (+4.23)
10		✓		✓			59.39 (+3.42)
11	✓	✓	✓	✓			60.39 (+4.42)
12					✓	✓	**61.48 (+5.51)**

ViT. Overall, our method consistently demonstrates superior performance and is well-adapted for transformers and CNNs.

For a qualitative comparison with previous works, please refer to the supplementary material.

C. Ablation Study

Given that an image is essentially a special case of a video without the temporal dimension, we opted for the more general video UDA-SS for ablation studies, using the DeepLab-V2 backbone by default.

1) Ablation Study on Each Component: In Table V, we conduct ablation studies to verify the effectiveness of three key components: video patch template (V-template) mixing, feature-level template (F-template) mixing, and optical-guided video feature aggregation (Agg.). Additionally, we also study the impact of the warm-up mentioned in Table I and Table II. For a fair comparison, we take the result of TPS [17] without random scale augmentation as the baseline (Exp. ID 0*).

It can be concluded that: (i) Video patch template (V-template) $Z^{*,k}_{t,n}$ result in a 3.70% mIOU gain in Exp. ID 1; (ii) As for feature-level template mixing (F-template) across two enhanced quad-mixed domains, Exp. ID 2 further incorporates the F-template and demonstrates a 1.54% mIOU gain. In addition, Exp. 8 outperforms Exp. 7 by 1.22%, also validating the efficacy of "F-template"; (iii) In another line, Exp. 3 introduces the optical-guided video feature aggregation (Agg.), resulting in growth to 54.01%; (iv) Furthermore, Exp. ID 4 combines "V-template", "F-template", and "Agg.", achieving an mIOU of 59.62%, surpassing the state-of-the-art TPL-SFC [19] mIOU of 55.3% by 4.32%; (v) Moreover, we also employ the warm-up setup in Exp. ID 5 to Exp. ID 8 for model parameter initialization, exceeding the state-of-the-art CMOM of 56.31% with a warm-up stage by 5.17%, resulting in the highest mIOU of 61.48%, a remarkable 9.47% increase compared to the baseline Exp. ID 0*.

Further comparison. To valid the video patch templates (V-template) $Z^{*,k}_{t,n}$ in Exp. ID 1, we perform a variant using image patch templates (I-template) akin to [5], [21]. Thus, the model input is also substituted with $x_t^{\mathcal{S}'}$ and $x_t^{\mathcal{T}'}$ in Eq. (22). It is observed that with "I-template", the mIOU dropped by 1.37% compared with Exp. ID 1, which suggests that "V-template" enhances the video adaptation more effectively.

Qualitative results of each component. In Fig. 8, we present segmentation results of three adjacent frames on Synthia-Seq → Cityscapes-Seq, following experiment settings corresponding to Exp. ID 0*, ID 2, ID 3, ID 5, ID 8 in Table V. These results demonstrate the incremental integration of each component. The segmentation results progressively improve from left to right, particularly for categories like *sign* and *car*. Notably, the challenging issue with a *person* on the far left within the frame is effectively addressed by integrating the proposed key components (Exp. ID 8).

2) Ablation Study on Quad-Directional Mixing: We evaluate the effects of QuadMix paths through quantitative experiments and detailed feature visualization.

The effectiveness of quad-directional mixing paths. In this paper, we propose four comprehensive mixing paths designed for sufficient domain gap bridging. It naturally prompts us to explore the outcomes when only the intra/inter-mixed source or target domains are constructed or when templates are exclusively from the source or target domain. Table VI provides thorough ablation results for these scenarios. For fairness, we employ the flow-guided feature aggregation module (i.e., Exp. ID 6 in Table V) as the baseline (Exp. ID 0*).

It can be observed: (i) In Exp. ID 1, our $\mathcal{S}\to\mathcal{T}$ setting outperformed the CMOM [2] by 1.80% in mIoU, which also follows a source to target pasting paradigm. We attribute this improvement to the proposed flow-guided feature aggregation for domain alignment; (ii) The construction of intra/inter-mixed video domains always yields better results than solely one-way mixed domains. It is because the model can learn consistent spatio-temporal features with diverse domain contexts, with the feature-level mixing across quad-mixed domains alleviating feature inconsistencies; (iii) Utilizing source domain templates commonly outperforms using target domain templates with higher predictive confidence. The overall mIOU

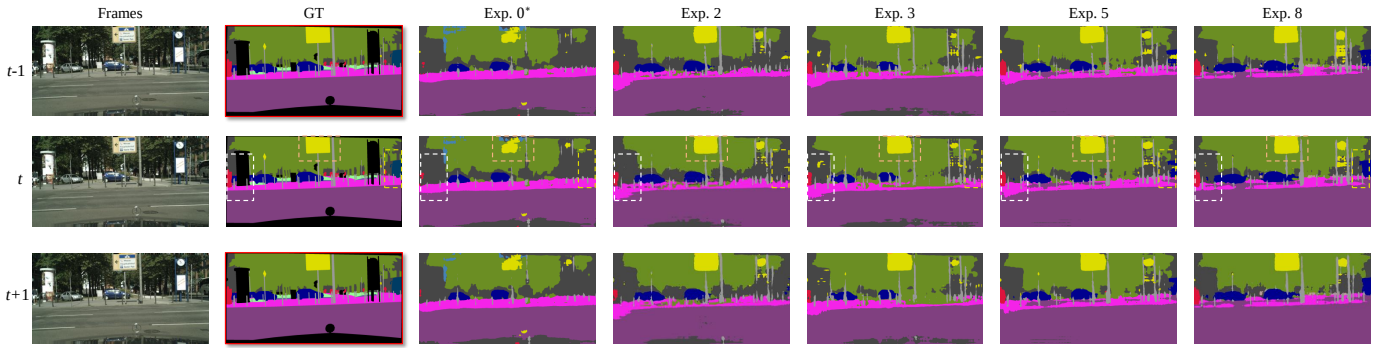


Fig. 8. Qualitative results of ablation studies in Table V for three consecutive frames on Synthia-Seq → Cityscapes-Seq. The best results are obtained by integrating all of the proposed components (Exp. ID 8), especially for categories of *person*, *rider*, and *car*. Please zoom in for details.

TABLE VII
ABLATION STUDIES ON CATEGORY SPACES FOR VIDEO PATCH TEMPLATE GENERATION ON SYNTHIA-SEQ → CITYSCAPES-SEQ.

Mixed categories	Things	Stuffs	Movable	Stationary	All
mIOU(%)	58.53	57.50	58.29	55.55	**61.48**

TABLE VIII
ABLATION STUDIES ON FLOW-GUIDED FEATURE AGGREGATION FOR ALIGNMENT, AND IMAGE ENHANCEMENT $\mathcal{A}$ ON THE TARGET DOMAIN.

Method	SYN-Seq → City-Seq	VIPER → Citys-Seq
w/o align	57.95	52.80
global align	58.24	53.62
flow-guided aggregation	**61.48**	**56.16**
w/o $\mathcal{A}$	60.23	55.25
w/ $\mathcal{A}$	**61.48**	**56.16**

exhibits further improvement when incorporating templates from both domains; (iv) More intriguingly, compared with incrementally combining various directional mixing as in Exp. ID 11, we not only reduce computational load significantly as shown in Table XI, but also achieve a mIoU gain of 0.92% to 61.48% by fully utilizing the enhanced quad-mixed domains for comprehensive domain bridging (Exp. ID 12).

It's noteworthy that regardless of the intra/inter-mixing strategy adopted, the results of ablation studies in Table VI `all surpass` previous state-of-the-art methods CMOM [2] and TPL [19], which highlights the effectiveness of our novel and unified perspective that has been largely overlooked by prior works [2], [3], [5], [17], [21], etc.

Feature visualization for quad-directional mixing paths. In Fig. 9, following experimental setings in Table VI, we present t-SNE [51] feature visualization for pure two domains (Exp. ID 0*), our `intra-mixed` domains (Exp. ID 6), one-way mixing (Exp. ID 1), two-way mixing (Exp. ID 3), and the proposed `QuadMix` (Exp. ID 12) from left to right. Notably, the number of points in Fig. 9(b)(d)(e) are equal. It can be found that quad-directional mixing effectively generalizes intra-domain features and enhances inter-domain features, addressing the insufficient feature gap bridging across distinct domains. By bridging domain gaps sufficiently, it promotes dense knowledge transfer from the source to the target domain.

The effect of category selection on video patch template generation. The video patch templates used for quad-mixing involve patch regions corresponding to the randomly selected two categories from the source or target domain. We investigated the impact of category selection on generating video patch templates. Drawing on the concept from panoramic segmentation [73], objects are often categorized into "things" (e.g., *sign*, *car*) and "stuff" (e.g., *sky*, *road*). Therefore, in Table VII, we examined the results when the category selection comprised only "things" or "stuff". Moreover, we also explore the movable (e.g., *person*, *car*) or stationary (e.g., *sign*, *light*) objects. The experimental findings indicate that "things" and movable categories yield better results than "stuff" and stationary categories, and the model performs best when the category selection encompasses all categories, which highlights the versatility of our method across different categories.

3) Ablation Study on Flow-Guided Spatio-Temporal Feature Aggregation Module: In this paper, to further address spatio-temporal gaps across video domains, we propose the optical flow-guided video feature aggregation module for fine-grained alignment. When exclusively substituting our fine-grained aggregation with treating video features as a whole [3], the mIOUs on two benchmarks decreased by 3.24% and 2.54%, respectively, as shown in Table VIII. It demonstrates the effectiveness of our method in compressing video features across domains into a more aligned distribution, thereby reducing the domain discrepancy with precise alignment. In addition, we also investigate the effect of image enhancement $\mathcal{A}$ in Eq. (22) and Eq. (24) for target domain samples.

Note that our contribution focuses on enhancing spatio-temporal coherence of aggregated features for domain alignment, which alignment loss is optimal to minimize discrepancies of aggregated features across domains is another research branch and goes beyond the main consideration of this work.

D. Further Analysis

We conduct further analysis on the general task of video UDA-SS, using the DeepLab-V2 backbone by default.

1) Hyperparameter Sensitivity: For model optimization, we utilize two hyperparameters, $\lambda_{\mathcal{T}}$ in Eq. (22) and Eq. (24) and λ_f in Eq. (23). In Table IX, we progressively increase the values of $\lambda_{\mathcal{T}}$ and λ_f using Viper → Cityscapes-Seq as an example to explore the hyperparameter sensitivity. The model achieves an mIOU of 56.16% when these parameters are set

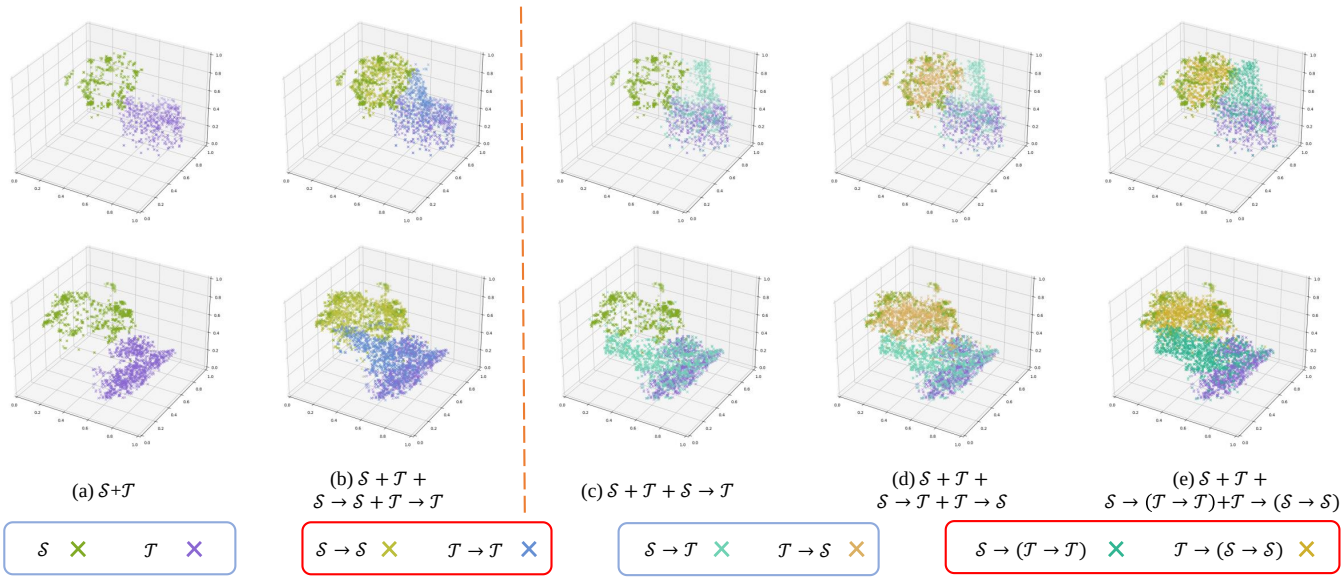


Fig. 9. T-SNE visualization of different mixing paradigms using *sign* (line one) and *pole* (line two) as examples. The experimental settings for subfigures (a) to (e) correspond to ablation studies Exp. ID 0*, ID 6, ID 1, ID 3, and ID 12 in Table VI. Note that the number of points in (b), (d), and (e) is equal. Subfigure (b) shows intra-mixed domains with continuous feature embeddings, while (e) depicts quad-mixed domains with a more generalized feature distribution, rather than a clustered one, which better addresses intra-domain discontinuity and fragmented gap bridging for knowledge transfer. Please zoom in for details.

TABLE IX
LOSS WEIGHTS SENSITIVITY ANALYSIS ON VIPER → CITYSCAPES-SEQ.

Weights	mIoU(%)
$\lambda_{\mathcal{T}}$, λ_f = 0.10, 0.01	54.56
$\lambda_{\mathcal{T}}$, λ_f = 0.20, 0.01	**56.16**
$\lambda_{\mathcal{T}}$, λ_f = 0.50, 0.01	54.49
$\lambda_{\mathcal{T}}$, λ_f = 0.50, 0.10	54.56
$\lambda_{\mathcal{T}}$, λ_f = 0.70, 0.10	54.58
$\lambda_{\mathcal{T}}$, λ_f = 1.00, 0.10	<u>55.21</u>

TABLE X
EFFECTIVENESS COMPARISON ON MODEL INITIALIZATION. THE "ORIGINAL" COLUMN DENOTES THE RESULTS OF ORIGINAL WORKS, WHILE THE "QUADMIX" COLUMN MEANS THE RESULTS OF OUR MODELS INITIALIZED WITH WARM-UP MODELS FROM THE "ORIGINAL" COLUMN.

	SYN-Seq → City-Seq		VIPER → Citys-Seq	
Method	Original	QuadMix	Original	QuadMix
ResNet-101 [71]	32.01	59.62	24.60	54.94
DA-VSN [16]	49.53	59.07	47.84	55.31
CMOM [2]	56.31	60.33	53.81	56.10
TPS [17]	53.76	**61.48**	48.91	**56.16**

to 0.20 and 0.01 and a suboptimal mIOU of 55.21% when the two parameters are set to 1.00 and 0.10. The results display slight fluctuations with parameter setting changes, showing our method's robustness. For convenience, we set $\lambda_{\mathcal{T}}$ and λ_f to 1 for another Synthea-Seq → Cityscapes-Seq benchmark.

2) Results with Model Initialization: As shown in Table X, we achieve 59.62% and 54.94% mIOU on two benchmarks by end-to-end training, surpassing previous methods trained both end-to-end and stage-wise. Initializing our models with warm-up parameters from previous works in the "Original" column can further enhance adaptation. Notably, CMOM [2] initializes model parameters and generates initial pseudo-labels offline using a warm-up model trained in [16]. In contrast, we solely undergo model initialization. The above analysis underscores the remarkable adaptation of our method on video UDA-SS.

TABLE XI
THROUGHPUT IS MEASURED WITH A BATCH SIZE OF 1. RESULTS ARE OBTAINED USING A V100-32G GPU. "QUADMIX*" CORRESPONDS TO THE ABLATION SETTING ID 11 IN TABLE VI.

Modules			SYN-Seq→City-Seq		VIPER→City-Seq	
Src.	QuadMix	QuadMix* Agg.	Time (s/iter)	GPU memory	Time (s/iter)	GPU memory
✓			1.40	8.78 G	1.38	8.78 G
			2.31	12.52 G	2.34	13.54 G
	✓		3.78	21.24 G	3.99	21.90 G
	✓	✓	4.26	24.92 G	4.37	25.90 G
		✓ ✓	5.64	29.11 G	5.89	29.85 G

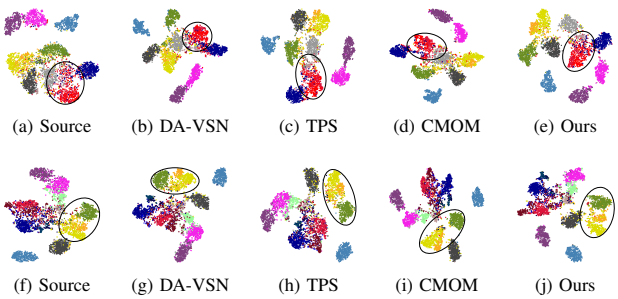


Fig. 10. T-SNE analysis of existing methods and our method on Synthia-Seq → Cityscapes-Seq (line one) and VIPER → Cityscapes-Seq (line two). Our model excels in learning more discriminative features of target videos within embedding space. Please zoom in for details.

3) The Throughput Measured for Different Modules: In Table XI, we present the training time and GPU memory footprint associated with the integration of different components. "Src." refers to training only with the source domain data. "QuadMix*" corresponds to the ablation setting ID 11

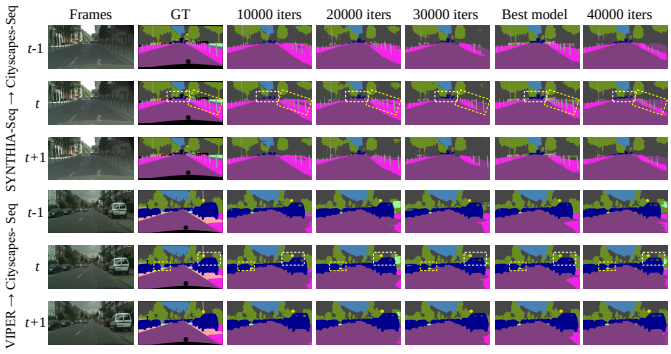


Fig. 11. Qualitative semantic segmentation results of our models at different training iterations on video UDA-SS benchmarks. Please zoom in for details.

in Table VI. Compared to "Src.", the proposed QuadMix $(\mathcal{T}\rightarrow(\mathcal{S}\rightarrow\mathcal{S}))$ and $(\mathcal{S}\rightarrow(\mathcal{T}\rightarrow\mathcal{T}))$ leads to a moderate increase in training time and GPU memory, and "Agg" exhibits relatively lighter overhead.

We also investigate the computational cost of Exp. ID 11 in Table VI (incrementally using $\mathcal{S}\rightarrow\mathcal{S}+\mathcal{T}\rightarrow\mathcal{T}+\mathcal{S}\rightarrow\mathcal{T}+\mathcal{T}\rightarrow\mathcal{S}$), which is higher in cost than our QuadMix. Notably, only adjacent frames and optical flow are required as inputs during model testing. In line with [2], [16], [17], [19], our method maintains similar inference speed and memory footprint.

4) T-SNE Analysis on Two Benchmarks.: For better visualization, we present t-SNE [51] visualizations of video feature embeddings from three recent works and the source-only model, compared with our QuadMix in Fig. 10. The dot colors for categories are consistent with Fig. 7. We observe that category-wise clusters in recent methods tend to be more scattered and overlapped, whereas our method generates more compact clusters. For instance, in our method, categories like *perdon*, *rider*, and *fence* in Synthia-Seq → Cityscapes-Seq, and categories like *light*, *sign*, and *vegetation* in VIPER → Cityscapes-Seq, exhibit more compact intra-category distribution and more distinct inter-category differences, revealing better discriminative capability for the target domain.

5) Results of Different Training Iterations: In Fig. 11, we provide a more comprehensive presentation of the semantic segmentation results of our models trained at different iterations. Notably, the best models for both two benchmarks emerge between the 35k and 40k iterations, providing a clearer insight into the evolution and stability of the training process.

V. Conclusion

This paper tackles the more challenging task of unified UDA-SS across both video and image scenarios by presenting an efficient and comprehensive approach. Specifically, we propose the QuadMix that involves quad-directional intra- and inter-domain mixing at both pixel and feature levels, aiming to generalize intra-domain features, enhance inter-domain gap bridging, and alleviate feature inconsistencies for domain invariant learning. Furthermore, we propose flow-guided video feature aggregation for fine-grained domain alignment, applicable to image scenarios as well. Our method is compatible with both CNN and transformer-based architectures. Extensive experiments on four challenging benchmarks demonstrate the superiority of our method.

Limitation: One limitation is that, while the quad-mixing enhances domain adaptation effectively, the theoretical interpretability of underlying mechanisms remains to be explored.

Future work: An interesting aspect of future work is to explore unified techniques across video and image scenarios, particularly in multi-modal domain adaptation, open-set domain adaptation, and extensions to instance or panoramic segmentation, as well as pixel-level anomaly detection.

Acknowledgment

This work is supported in part by the Major Program of the National Natural Science Foundation of China (NSFC) under Grant No. 61991404, in part by the Research Program of the Liaoning Liaohe Laboratory under Grant No. LLL23ZZ-05-01, in part by the Key Research and Development Program of Liaoning Province under Grant No. 2023JH26/10200011, in part by the Natural Science Foundation of Liaoning Province of China under Grant No. 2024-MSBA-42, and in part by the Program of China Scholarship Council 202306080142. Dr. Tao's research is partially supported by NTU RSR and Start Up Grants. The authors would like to thank Dr. Chunhua Shen and Lan Deng for their valuable suggestions.

References

[1] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9924–9935.

[2] K. Cho, S. Lee, H. Seong, and E. Kim, "Domain adaptive video semantic segmentation via cross-domain moving object mixing," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 489–498.

[3] L. Gao, J. Zhang, L. Zhang, and D. Tao, "Dsp: Dual soft-paste for unsupervised domain adaptive semantic segmentation," in *Proc. ACM Int. Conf. Multimed.*, 2021, pp. 2825–2833.

[4] H. Zhao, J. Zhang, Z. Chen, S. Zhao, and D. Tao, "Unimix: Towards domain adaptive and generalizable lidar semantic segmentation in adverse weather," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 14 781–14 791.

[5] D. Kim, M. Seo, K. Park, I. Shin, S. Woo, I. S. Kweon, and D.-G. Choi, "Bidirectional domain mixup for domain adaptive semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, 2023, pp. 1114–1123.

[6] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn*, 2018, pp. 1989–1998.

[7] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in *Proc. Int. Conf. Mach. Learn*, 2019, pp. 7404–7413.

[8] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, 2012.

[9] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–450.

[10] N. Araslanov and S. Roth, "Self-supervised augmentation consistency for adapting semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15 384–15 394.

[11] K. Mei, C. Zhu, J. Zou, and S. Zhang, "Instance adaptive self-training for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 415–430.

[12] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12 414–12 424.

[13] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5982–5991.

[14] H. Tang, X. Zhu, K. Chen, K. Jia, and C. P. Chen, "Towards uncovering the intrinsic data structures for unsupervised domain adaptation using structurally regularized deep clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6517–6533, 2021.

[15] I. Shin, K. Park, S. Woo, and I. S. Kweon, "Unsupervised domain adaptation for video semantic segmentation," *arXiv preprint arXiv:2107.11052*, 2021.

[16] D. Guan, J. Huang, A. Xiao, and S. Lu, "Domain adaptive video segmentation via temporal consistency regularization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 8053–8064.

[17] Y. Xing, D. Guan, J. Huang, and S. Lu, "Domain adaptive video segmentation via temporal pseudo supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 621–639.

[18] F. Pan, X. Yin, S. Lee, A. Niu, S. Yoon, and I. S. Kweon, "Moda: Leveraging motion priors from videos for advancing unsupervised domain adaptation in semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2024, pp. 2649–2658.

[19] Y. Gao, Z. Wang, J. Zhuang, Y. Zhang, and J. Li, "Exploit domain-robust optical flow in domain adaptive video semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 641–649.

[20] H. Mai, R. Sun, Y. Wang, T. Zhang, and F. Wu, "Pay attention to target: Relation-aware temporal consistency for domain adaptive video semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4162–4170.

[21] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "Dacs: Domain adaptation via cross-domain mixed sampling," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1379–1389.

[22] Y. Cheng, F. Wei, J. Bao, D. Chen, and W. Zhang, "Adpl: Adaptive dual path learning for domain adaptation of semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9339–9356, 2023.

[23] B. Xie, S. Li, M. Li, C. H. Liu, G. Huang, and G. Wang, "Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9004–9021, 2023.

[24] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6502–6509.

[25] G. Kim and S. Y. Chun, "Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2023, pp. 14 203–14 213.

[26] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 10 285–10 295.

[27] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 433–443.

[28] Z. Wang, M. Yu, Y. Wei, R. Feris, J. Xiong, W.-m. Hwu, T. S. Huang, and H. Shi, "Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12 635–12 644.

[29] J. Li, Z. Wang, Y. Gao, and X. Hu, "Exploring high-quality target domain information for unsupervised domain adaptive semantic segmentation," in *Proc. ACM Int. Conf. Multimed.*, 2022, pp. 5237–5245.

[30] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472–7481.

[31] K. Mei, C. Zhu, J. Zou, and S. Zhang, "Instance adaptive self-training for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 415–430.

[32] N. Karim, N. C. Mithun, A. Rajvanshi, H. Chiu, S. Samarasekera, and N. Rahnavard, "C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 24 120–24 131.

[33] X. Li, P. Ye, J. Li, Z. Liu, L. Cao, and F. Wang, "From features engineering to scenarios engineering for trustworthy ai: I&i, c&c, and v&v," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 18–26, 2022.

[34] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472–7481.

[35] S. K. Panda, Y. Lee, and M. K. Jawed, "Agronav: Autonomous navigation framework for agricultural robots and vehicles using semantic segmentation and semantic line detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6271–6280.

[36] I. Alonso, L. Riazuelo, and A. C. Murillo, "Mininet: An efficient semantic segmentation convnet for real-time robotic applications," *IEEE Trans. Robot.*, vol. 36, no. 4, pp. 1340–1347, 2020.

[37] S. Jain, X. Wang, and J. E. Gonzalez, "Accel: A corrective fusion network for efficient semantic segmentation on video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8866–8875.

[38] A. Ganeshan, A. Vallet, Y. Kudo, S.-i. Maeda, T. Kerola, R. Ambrus, D. Park, and A. Gaidon, "Warp-refine propagation: Semi-supervised auto-labeling via cycle-consistency," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 15 499–15 509.

[39] Y. Zhang, S. Borse, H. Cai, and F. Porikli, "Auxadapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2339–2348.

[40] F. Yuan, L. Zhang, X. Xia, Q. Huang, and X. Li, "A gated recurrent network with dual classification assistance for smoke semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 4409–4422, 2021.

[41] M. Guo, T. Xu, J. Liu, Z. Liu, P. Jiang, T. Mu, S. Zhang, R. R. Martin, M. Cheng, and S. Hu, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.

[42] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi, "Temporally distributed networks for fast video semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8818–8827.

[43] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–16.

[44] G. Wu, Y. Zhang, L. Deng, J. Zhang, and T. Chai, "Cross-modal learning for anomaly detection in complex industrial process: Methodology and benchmark," *IEEE Trans. Circuits Syst. Video Technol.*, 2024.

[45] X. Wu, Z. Wu, J. Wan, L. Ju, and S. Wang, "Is it necessary to transfer temporal knowledge for domain adaptive video semantic segmentation?" in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 357–373.

[46] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12 077–12 090.

[47] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann, "Domain adaptation on the statistical manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2481–2488.

[48] R. Gopalan, R. Li, and R. Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2288–2302, 2013.

[49] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1369–1378.

[50] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6023–6032.

[51] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[52] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[53] J. Moon, J. Kim, Y. Shin, and S. Hwang, "Confidence-aware learning for deep neural networks," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2020, pp. 7034–7044.

[54] T. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2517–2526.

[55] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 289–305.

[56] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3764–3773.

[57] D. Guan, J. Huang, S. Lu, and A. Xiao, "Scale variance minimization for unsupervised domain adaptation in image segmentation," *Pattern Recognit.*, vol. 112, p. 107764, 2021.

[58] J. Huang, S. Lu, D. Guan, and X. Zhang, "Contextual-relation consistent domain adaptation for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 705–722.

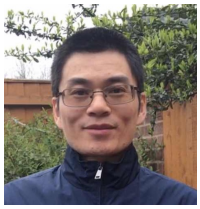
[59] J. Huang, D. Guan, A. Xiao, and S. Lu, "Rda: Robust domain adaptation via fourier adversarial attacking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 8988–8999.

[60] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4085–4095.
[61] L. Melas-Kyriazi and A. K. Manrai, "Pixmatch: Unsupervised domain adaptation via pixelwise consistency training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12 435–12 445.
[62] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.
[63] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3234–3243.
[64] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[65] S. R. Richter, Z. Hayder, and V. Koltun, "Playing for benchmarks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2213–2222.
[66] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 102–118.
[67] H. Ma, X. Lin, and Y. Yu, "I2f: A unified image-to-feature approach for domain adaptive semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1695–1710, 2024.
[68] T.-D. Truong, N. Le, B. Raj, J. Cothren, and K. Luu, "Fredom: Fairness domain adaptation approach to semantic scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19 988–19 997.
[69] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.
[70] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2758–2766.
[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[72] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Feifei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[73] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, "Panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9404–9413.

Zhe Zhang received the BS degree in the College of Information Science and Engineering, Northeastern University, China, in 2021. He is currently working toward a Ph.D. degree in the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, China. His current research interests include computer vision, deep learning, video representation learning, multi-modal learning, and their applications in industrial fields.

Gaochang Wu received the BE and MS degrees in mechanical engineering in Northeastern University, Shenyang, China, in 2013 and 2015, respectively, and Ph.D. degree in control theory and control engineering in Northeastern University, Shenyang, China in 2020. He is currently an associate professor in the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University. He was selected for the 2022-2024 Youth Talent Support Program of the Chinese Association of Automation. His current research interests include multimodal perception and recognition, as well as light field processing.

Jing Zhang (Senior Member, IEEE) is currently a professor at the School of Computer Science, Wuhan University, Wuhan, China. He previously served as a Research Fellow at the School of Computer Science, The University of Sydney. He has published over 100 papers in leading venues such as CVPR, ICCV, NeurIPS, IEEE TPAMI, and IJCV, with research focused on computer vision and deep learning. He is an Area Chair for NeurIPS and ICPR, a Senior Program Committee member for AAAI and IJCAI, and a guest editor for IEEE TBD, while also regularly reviewing for top-tier journals and conferences.



Xiatian Zhu received the PhD degree from the Queen Mary University of London. He is a senior lecturer with Surrey Institute for People-Centred Artificial Intelligence, and Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, U.K. He won the Sullivan Doctoral Thesis Prize 2016. He was a research scientist with Samsung AI Centre, Cambridge, U.K. His research interests include computer vision and machine learning.

Dacheng Tao (Fellow, IEEE) is currently a Distinguished University Professor in the College of Computing & Data Science at Nanyang Technological University. He mainly applies statistics and mathematics to artificial intelligence and data science, and his research is detailed in one monograph and over 200 publications in prestigious journals and proceedings at leading conferences, with best paper awards, best student paper awards, and test-of-time awards. His publications have been cited over 112K times and he has an h-index 160+ in Google Scholar. He received the 2015 and 2020 Australian Eureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award. He is a Fellow of the Australian Academy of Science, AAAS, ACM and IEEE.

Tianyou Chai (Life Fellow, IEEE) received the Ph.D. degree in control theory and engineering from Northeastern University, Shenyang, China, in 1985. He became a Professor at Northeastern University in 1988. He is the Founder and the Director of the Center of Automation, Northeastern University, which became the National Engineering and Technology Research Center and the State Key Laboratory. He was the Director of the Department of Information Science, National Natural Science Foundation of China, from 2010 to 2018. He has developed control technologies with applications to various industrial processes. He has published more than 320 peer-reviewed international journal articles. His current research interests include modeling, control, optimization, and integrated automation of complex industrial processes.

Dr. Chai is a member of the Chinese Academy of Engineering and a Fellow of International Federation for Automatic Control (IFAC). His paper titled "Hybrid intelligent control for optimal operation of shaft furnace roasting process" was selected as one of the three best papers for the Control Engineering Practice Paper Prize for the term 2011–2013. For his contributions, he has won five prestigious awards of the National Natural Science, the National Science and Technology Progress, and the National Technological Innovation, the 2007 Industry Award for Excellence in Transitional Control Research from IEEE Multi-Conference on Systems and Control, and the 2017 Wook Hyun Kwon Education Award from the Asian Control Association.

Unified Domain Adaptive Semantic Segmentation Supplementary Material

Zhe Zhang, Gaochang Wu, *Member, IEEE*, Jing Zhang, Xiatian Zhu, Dacheng Tao, *Fellow, IEEE,* Tianyou Chai, *Life Fellow, IEEE*

This document presents the supplementary materials omitted from the main paper due to space limitations.

I. Training Algorithm

In this paper, our unified method follows an end-to-end learning paradigm. The training process for **UDA-VSS** is outlined in Algorithm 1. For **UDA-ISS**, the process is similar but excludes temporal information, i.e., optical flow. Please refer to our main paper for detailed equations and loss functions.

Algorithm 1 Unified Domain Adaptive Semantic Segmentation.

Input: $X_t^{\mathcal{S}}$, $X_t^{\mathcal{T}}$, $X_{t-\tau}^{\mathcal{T}}$, $o_{t-1\to t}^{\mathcal{S}}$, $o_{t-\tau-1\to t-\tau}^{\mathcal{T}}$, $o_{t'\to t}^{\mathcal{T}}$, $\lambda_{\mathcal{T}}$, λ_f, batch size B, and maximum iteration N.
Initialize our model using ImageNet pre-trained model.
for $n \leftarrow 0$ to N **do**
  if $n == 0$ **then**
    Obtain initial pseudo-label $\hat{Y}_t^T$ and mask $M_t^{\mathcal{T}^*}$.
  else
    Generate video patch templates from the source and target domains adaptively via Eq. (1).
    Create intra-mixed domains via Eq. (8) to Eq. (10).
    Create more enhanced inter-mixed domains via Eq. (11) to Eq. (12).
    Perform feature-level sequence mixing across two quad-mixed domains via Eq. (13).
    Conduct flow-guided feature aggregation for domain alignment via Eq. (18) to Eq. (20).
    Train the model with loss $\mathcal{L}_{all}$ via Eq. (21).
  end if
end for
Output: Final weights of the UDA-VSS model.

Note that the two inter-mixed domains derived from the two intra-mixed domains are also termed quad-mixed domains, as they are obtained through quad-directional mixing.

II. Differentiation of image UDA-SS Between Our Method and MoDA [1]

Recently, Pan et al. propose MoDA [1] and report segmentation mIOU on the *GTAV* $\rightarrow$ *Cityscapes-**Seq***, referring to it as domain adaptive `image` semantic segmentation. However, MoDA differs fundamentally from the concepts within mainstream methods [2]–[12] and `image UDA-SS` in this paper. The reason is as follows: MoDA utilizes temporal information from the target domain videos as geometric constraints to refine pseudo-labels. In contrast, we generate pseudo-labels using only spatial image information, following prior works.

TABLE I
Classes division regarding "things" or"stuff", and "movable" or "stationary" categories on SynthiaSeq $\rightarrow$ Cityscapes-Seq.

SynthiaSeq $\rightarrow$ Cityscapes-Seq	
thing	building, pole, light, sign, person, rider, car
stuff	road, side, vegetation, sky
movable object	person, rider, car
stationary object	road, side, building, pole, light, sign, vegetation, sky

Essentially, MoDA represents an attempt at domain adaptation from a source `image` domain to a target `video` domain, similar to I2VDA [13], which is discussed in the main paper. As for performance, MoDA achieves mIOUs of 54.9% and 75.2% on *GTAV* $\rightarrow$ *Cityscapes-**Seq***, using CNN [14] or transformer-based [15] networks, respectively. In comparison, we obtain mIOUs of 66.8% and 76.1% on *GTAV* $\rightarrow$ *Cityscapes*, respectively, as shown in Table III of the main paper. Despite leveraging temporal information, the results of MoDA remain below ours. Notably, using the CNN-based network, we outperform MoDA by 11.9%, demonstrating its effectiveness and substantial potential.

Additionally, MoDA only reports results on one video UDA-SS benchmark, Viper $\rightarrow$ Cityscapes-Seq, with an mIOU of 49.1% (CNN-based). In contrast, our method achieved 56.2% (CNN-based), significantly surpassing MoDA. Moreover, unlike previous works [13], [16]–[20] that only employ CNN-based networks, *for the first time*, we explore the vision transformer [3] in video UDA-SS, achieving a higher mIOU of 58.2%. Furthermore, we also conduct experiments on the Synthia-Seq $\rightarrow$ Cityscapes-Seq benchmark, achieving mIOUs of 61.5% and 67.2% with CNN or transformer-based networks, respectively. Based on the above analysis, we indeed pioneer unified image and video UDA-SS, setting new state-of-the-art benchmarks on four challenging benchmarks.

III. Extended Experimental Analysis

A. Qualitative Results for Image UDA-SS

Besides the qualitative results in the main paper (Fig. 7) for video UDA-SS tasks, in Fig. 1 and Fig. 2, we present the results for image UDA-SS tasks: SYNTHIA $\rightarrow$ Cityscapes and

TABLE II
CLASSES DIVISION REGARDING "THINGS" OR"STUFF", AND "MOVABLE" OR "STATIONARY" CATEGORIES ON VIPER → CITYSCAPES-SEQ.

Viper → Cityscapes-Seq	
thing	building, fence, light, sign, person, car, truck, bus, motorcycle, bike
stuff	road, side, vegetation, terr, sky
movable object	person, car, truck, bus, motorcycle, bike
stationary object	road, side, building, fence, light, sign, vegetation, terr, sky

GTAV → Cityscapes, as predicted by our method. We compare these results with those predicted by ADPL [5], DAFormer [3], and SePiCo [2]. Our model's predictions are smoother and contain fewer spurious areas compared to the other models, indicating a significant performance improvement.

B. More Qualitative Comparisons

Besides Fig. 1 and Fig. 2, and Fig. 7 in the main paper, please refer to the *supplementary video* for more visual comparisons of both `image` and `video` UDA-SS segmentation results with previous methods.

C. Category Selection for Video Patch Template Generation

In Table VII of the main paper, we investigated the impact of category selection on video patch template generation ($Z_t^{\mathcal{S}^*,k}$ and $Z_t^{\mathcal{T}^*,k}$) from both domains. We explored the results when the category selection space (distinct from the long-tail category space) comprised only "things" or "stuff" [21], as well as solely movable or stationary objects instead of the whole category space. Specifically, the classes division for Synthia-Seq → Cityscapes-Seq is shown in Table I and for Viper → Cityscapes-Seq is shown in Table II.

D. Long-Tail Categories for Templates

Long-tail class learning is an aspect for tackling rare object learning [4], [5], [8]. In this paper, we also involve two selected long-tail categories in *Source to Source Mixing* ($\mathcal{S}\to\mathcal{S}$) and *Source to Intra-mixed Target Mixing* ($\mathcal{S}\to(\mathcal{T}\to\mathcal{T})$) at each training iteration. Specifically, we randomly select two long-tail classes from categories of *pole*, *light*, *sign*, and *rider* for Synthia-Seq → Cityscapes-Seq, and *fence*, *light*, *sign*, *terrain*, *motor*, and *bike* for VIPER → Cityscapes-Seq.

E. Extended Visualization of Mixing Strategies in QuadMix

In Fig. 3, we extend the visual analysis presented in Fig. 4 of the main paper, further illustrating the effects of $\mathcal{S} \to \mathcal{T}$ and $\mathcal{T} \to \mathcal{S}$ for the SYNTHIA-Seq [22] → Cityscapes-Seq [23] benchmark. For additional insights, Fig. 4 presents corresponding visualizations for the Viper [24] → Cityscapes-Seq benchmark. Moreover, Fig. 7 to Fig. 12 present an extensive set of QuadMix visualizations across both video benchmarks, covering various strategies including $\mathcal{S} \to \mathcal{S}$, $\mathcal{T} \to \mathcal{T}$, $\mathcal{S} \to (\mathcal{T} \to \mathcal{T})$, and $\mathcal{T} \to (\mathcal{S} \to \mathcal{S})$, thereby enabling a more thorough understanding.

F. Visualization of the Evolution of Target Pseudo-Labels

We visualize the evolution of pseudo-label quality throughout training to illustrate how their accuracy improves across different stages. As shown in Fig. 5 (with CNN backbone) and Fig. 6 (with ViT backbone), the pseudo-labels undergo progressive refinement, increasingly aligning with the underlying structure of the target domain.

G. Ablation Study on Image UDA-SS

For the image UDA-SS task, we also conducted ablation experiments on the GTA → Cityscapes dataset (using the CNN architecture). The comparison includes spatial feature aggregation for domain alignment, I-template, and F-template. As shown in Table III and Table IV, all three components are effective and work better in combination, resulting in complementary benefits and enhanced domain alignment.

H. Effectiveness of Temporal Information for Video UDA-SS

For the video UDA-SS task, leveraging temporal information improves both consistency and feature learning by capturing dependencies across consecutive frames. As shown in Table V, our ablation studies demonstrate the effectiveness of two key components: optical flow warp-based pseudo-label generation, as well as the optical flow-based QuadMix. By aligning features over time, these strategies improve temporal coherence and feature representation, resulting in more robust and stable performance and ultimately enhancing adaptation to the target video domain.

I. Quantitative Impact of the Number of Mixed Categories

We conducted quantitative experiments to evaluate the impact of varying the number of mixed categories (0, 1, 2, 3, 4) from each domain on the SYNTHIA-Seq → Cityscapes-Seq benchmark, as shown in Table VI. The results indicate that the proposed QuadMix maintains strong performance as the number of mixed categories varies across domains.

J. Further Illustration of Temporal Cues in Video UDA-SS

Fig. 13 further illustrates the role of temporal information in pseudo-label generation (Eq. 15, Eq. 17, and Eq. 18) and quad-directional mixed sample learning (Eq. 12). This visualization offers a clearer insight into how temporal dependencies across consecutive frames enhance pseudo-label accuracy and facilitate more effective spatio-temporal learning in video UDA-SS.

K. Visualization of Image/Video Frame Enhancement

To promote consistent learning, we apply random Gaussian blur and color jitter, as defined in Eq. 22 and Eq. 24, with an 80% probability and varying intensities for each image or video frame. Fig. 14 presents visual comparisons of QuadMix results before and after enhancement for better understanding.

L. Detailed Analysis of Category-Aware IoU Results

We present the category-aware IoU results corresponding to the ablation studies in Sections IV-C and IV-D of the main paper. These results are reported in Tables VII and VIII. For clarity, all values are rounded to one decimal place.

TABLE III
ABLATION STUDIES ON **IMAGE UDA-SS** USING THE GTAV → CITYSCAPES BENCHMARK. "I-TEMPLATE" REPRESENTS THE IMAGE PATCH TEMPLATE, "F-TEMPLATE" REFERS TO FEATURE-LEVEL TEMPLATE MIXING ACROSS THE QUAD-MIXED DOMAINS, AND "AGG." DENOTES THE AGGREGATED IMAGE FEATURE ALIGNMENT.

ID	Agg.	I-template	F-template	mIOU (Gain)
0	✓			64.41
1	✓	✓		65.67
2	✓	✓	✓	66.76

TABLE IV
CATEGORY-AWARE IOU RESULTS CORRESPONDING TO THE ABLATION STUDY PRESENTED IN TABLE III.

ID	Road	Side.	Buil.	Wall	Fence	Pole	Light	Sign	Vege.	Terr.	Sky	Pers.	Rider	Car	Truck	Bus	Train	Moto.	Bike	mIOU
0	96.5	74.1	89.1	39.8	31.3	50.7	59.9	69.5	90.6	48.5	92.8	78.4	51.4	92.9	66.1	63.3	1.8	58.3	68.7	64.41
1	96.6	77.7	90.4	47.3	37.9	54.2	62.8	71.4	90.8	50.3	90.6	78.2	43.4	93.3	70.2	64.6	0.3	59.3	68.3	65.67
2	97.1	78.0	90.4	49.7	40.3	53.4	61.7	70.9	90.7	49.7	92.9	77.9	53.9	93.5	72.4	65.9	0.7	60.5	68.5	66.76

TABLE V
ANALYSIS OF TEMPORAL INFORMATION IN VIDEO UDA-SS ON THE SYNTHIA-SEQ → CITYSCAPES-SEQ BENCHMARK.

ID	Warp	Flow Mix	Road	Side.	Buil.	Pole	Light	Sign	Vege.	Sky	Pers.	Rider	Car	mIOU
0	✓		88.2	18.6	79.8	29.0	26.5	44.7	82.2	83.8	58.6	35.3	84.4	57.39
1		✓	89.2	43.8	81.0	30.1	20.9	45.0	82.3	82.6	60.1	26.9	87.2	59.01
2	✓	✓	90.8	39.9	83.2	33.2	30.1	50.7	84.8	82.3	61.2	32.7	87.4	61.48

TABLE VI
ABLATION STUDIES ON THE NUMBER OF MIXED CATEGORIES FROM SOURCE AND TARGET DOMAINS ON THE SYNTHIA-SEQ → CITYSCAPES-SEQ.

Mixed Number	Road	Side.	Buil.	Pole	Light	Sign	Vege.	Sky	Pers.	Rider	Car	mIOU
0	89.4	40.3	80.7	28.2	20.8	50.8	81.5	83.8	60.8	28.7	87.5	59.31
1	91.7	47.1	80.8	29.5	27.2	47.2	81.0	81.5	58.3	26.3	88.0	59.87
2	90.8	39.9	83.2	33.2	30.1	50.7	84.8	82.3	61.2	32.7	87.4	61.48
3	92.5	51.5	81.9	29.8	25.9	47.5	84.0	82.6	60.0	28.5	87.3	61.05
4	91.5	45.4	82.1	31.3	22.2	16.5	84.5	82.9	61.9	27.6	87.3	60.29

REFERENCES

[1] F. Pan, X. Yin, S. Lee, A. Niu, S. Yoon, and I. S. Kweon, "Moda: Leveraging motion priors from videos for advancing unsupervised domain adaptation in semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2024, pp. 2649–2658.

[2] B. Xie, S. Li, M. Li, C. H. Liu, G. Huang, and G. Wang, "Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9004–9021, 2023.

[3] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9924–9935.

[4] T.-D. Truong, N. Le, B. Raj, J. Cothren, and K. Luu, "Fredom: Fairness domain adaptation approach to semantic scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19 988–19 997.

[5] Y. Cheng, F. Wei, J. Bao, D. Chen, and W. Zhang, "Adpl: Adaptive dual path learning for domain adaptation of semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9339–9356, 2023.

[6] H. Ma, X. Lin, and Y. Yu, "I2f: A unified image-to-feature approach for domain adaptive semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1695–1710, 2024.

[7] D. Kim, M. Seo, K. Park, I. Shin, S. Woo, I. S. Kweon, and D.-G. Choi, "Bidirectional domain mixup for domain adaptive semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, 2023, pp. 1114–1123.

[8] L. Gao, J. Zhang, L. Zhang, and D. Tao, "Dsp: Dual soft-paste for unsupervised domain adaptive semantic segmentation," in *Proc. ACM Int. Conf. Multimed.*, 2021, pp. 2825–2833.

[9] N. Araslanov and S. Roth, "Self-supervised augmentation consistency for adapting semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15 384–15 394.

[10] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "Dacs: Domain adaptation via cross-domain mixed sampling," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1379–1389.

[11] J. Huang, D. Guan, A. Xiao, and S. Lu, "Rda: Robust domain adaptation via fourier adversarial attacking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 8988–8999.

[12] L. Melas-Kyriazi and A. K. Manrai, "Pixmatch: Unsupervised domain adaptation via pixelwise consistency training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12 435–12 445.

[13] X. Wu, Z. Wu, J. Wan, L. Ju, and S. Wang, "Is it necessary to transfer temporal knowledge for domain adaptive video semantic segmentation?" in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 357–373.

[14] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.

[15] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12 077–12 090.

[16] D. Guan, J. Huang, A. Xiao, and S. Lu, "Domain adaptive video segmentation via temporal consistency regularization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 8053–8064.

[17] Y. Xing, D. Guan, J. Huang, and S. Lu, "Domain adaptive video

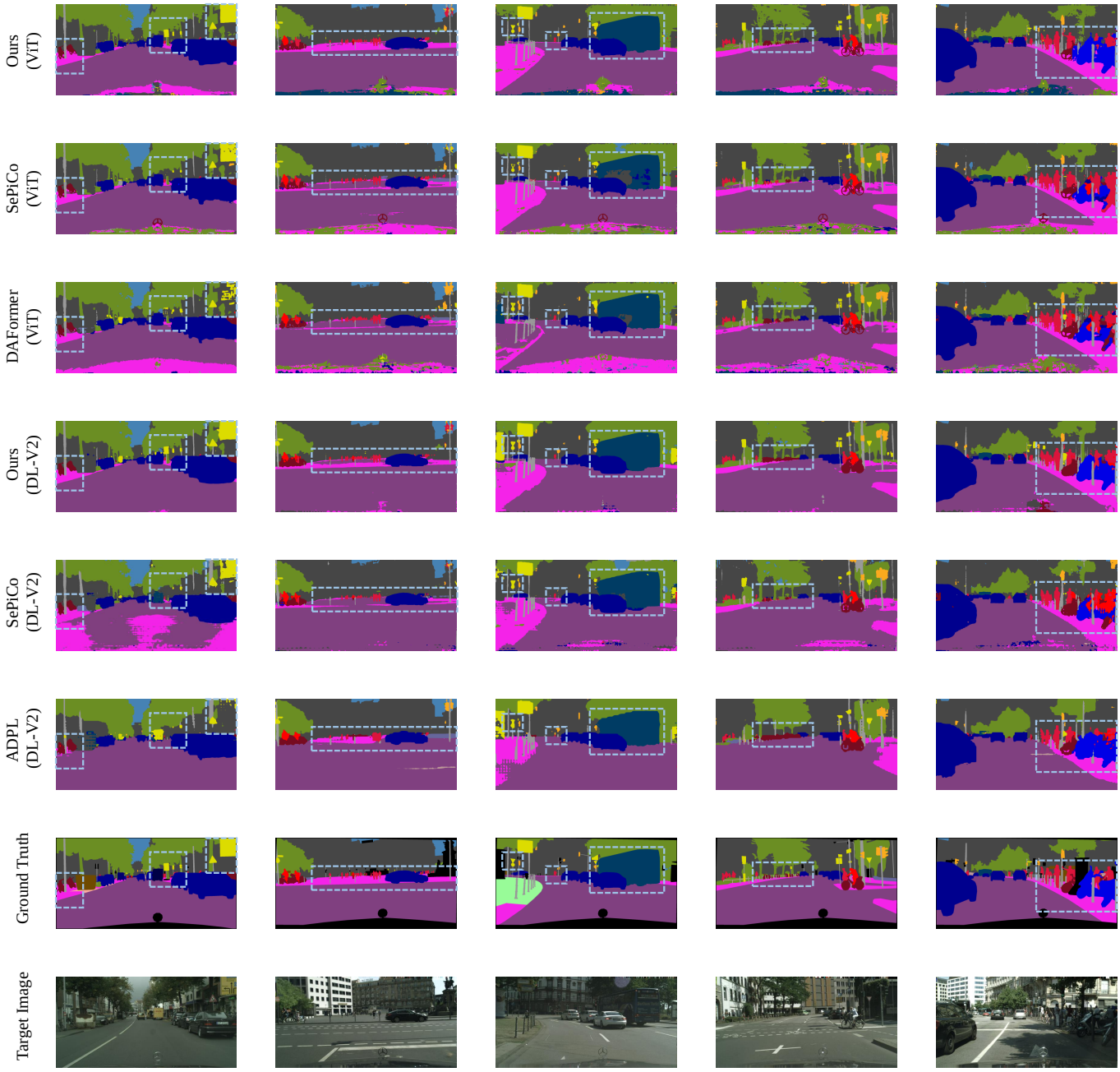


Fig. 1. Qualitative comparison of results on the image UDA-SS task (SYNTHIA → Cityscapes). We compare our method with ADPL [5], DAFormer [3], and SePiCo [2], including results using CNN [14] and Vision Transformer [3] architectures. Please zoom in for details.

segmentation via temporal pseudo supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 621–639.

[18] Y. Gao, Z. Wang, J. Zhuang, Y. Zhang, and J. Li, "Exploit domain-robust optical flow in domain adaptive video semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 641–649.

[19] H. Mai, R. Sun, Y. Wang, T. Zhang, and F. Wu, "Pay attention to target: Relation-aware temporal consistency for domain adaptive video semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4162–4170.

[20] K. Cho, S. Lee, H. Seong, and E. Kim, "Domain adaptive video semantic segmentation via cross-domain moving object mixing," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 489–498.

[21] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, "Panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9404–9413.

[22] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3234–3243.

[23] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[24] S. R. Richter, Z. Hayder, and V. Koltun, "Playing for benchmarks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2213–2222.

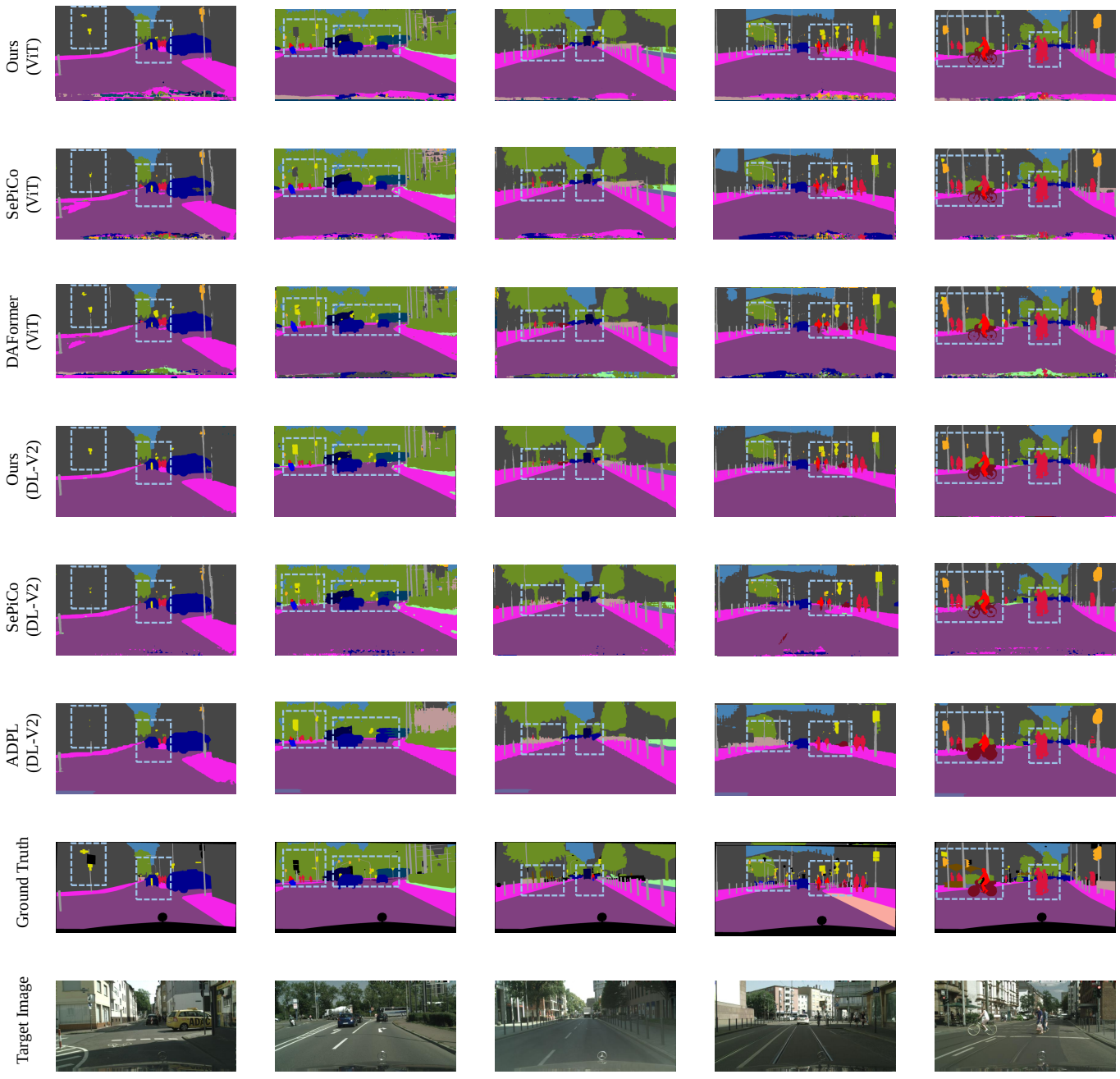


Fig. 2. Qualitative comparison of results on the image UDA-SS task (GTAV → Cityscapes). We compare our method with ADPL [5], DAFormer [3], and SePiCo [2], including results using CNN [14] and Vision Transformer [3] architectures. Please zoom in for details.

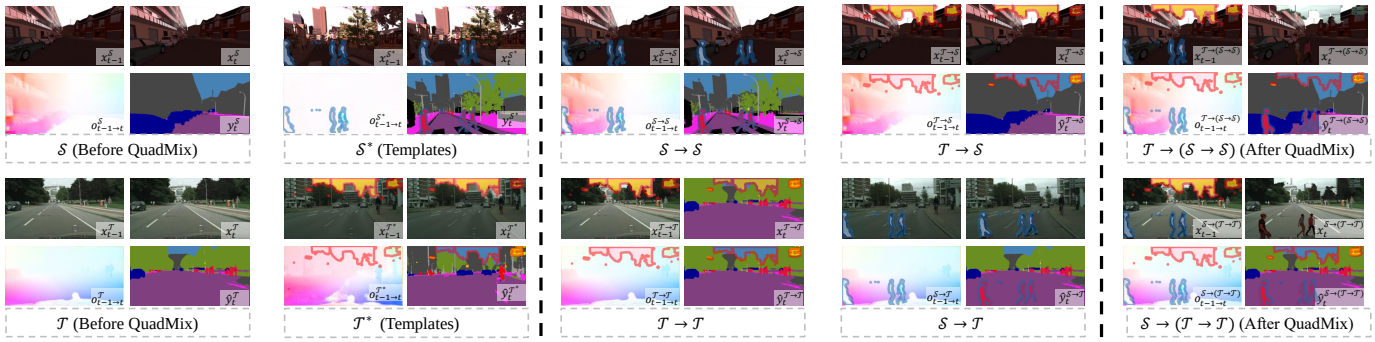


Fig. 3. Examples of various mixing strategies in QuadMix on the **SYNTHIA-Seq** $\to$ **Cityscapes-Seq** benchmark: (i) $\mathcal{S}$ and $\mathcal{T}$ (before QuadMix), (ii) $\mathcal{S}^*$ and $\mathcal{T}^*$ (source templates: *person* and *rider*, target templates: *sign* and *sky*), (iii) $\mathcal{S} \to \mathcal{S}$ and $\mathcal{T} \to \mathcal{T}$ (intra-domain mixing), (iv) $\mathcal{S} \to \mathcal{T}$ and $\mathcal{T} \to \mathcal{S}$ (not directly applied in this work), (v) $\mathcal{S} \to (\mathcal{T} \to \mathcal{T})$ and $\mathcal{T} \to (\mathcal{S} \to \mathcal{S})$ (further with inter-domain mixing, i.e. after QuadMix). The effects of these strategies on video frames, optical flow, and labels/pseudo-labels are illustrated. We present $x_t^{\mathcal{T} \to (\mathcal{S} \to \mathcal{S})}$ and $x_t^{\mathcal{S} \to (\mathcal{T} \to \mathcal{T})}$ without masks for better understanding. Notably, the patch templates required in training iteration n are generated online adaptively from iteration $n-1$. Please zoom in for details.

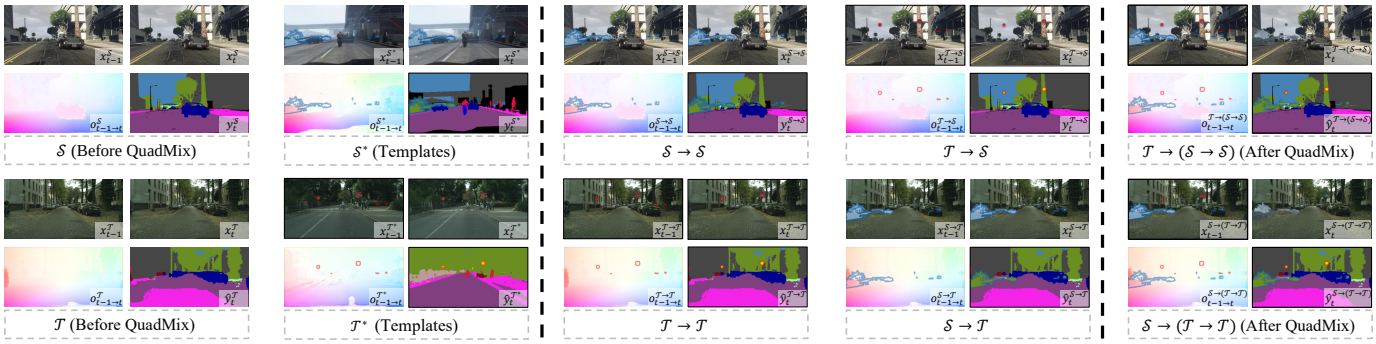


Fig. 4. Examples of various mixing strategies in QuadMix on the **Viper** $\to$ **Cityscapes-Seq** benchmark: (i) $\mathcal{S}$ and $\mathcal{T}$ (before QuadMix), (ii) $\mathcal{S}^*$ and $\mathcal{T}^*$ (source templates: *vegetation* and *terrain*, target templates: *sign* and *building*), (iii) $\mathcal{S} \to \mathcal{S}$ and $\mathcal{T} \to \mathcal{T}$ (intra-domain mixing), (iv) $\mathcal{S} \to \mathcal{T}$ and $\mathcal{T} \to \mathcal{S}$ (not directly applied in this work), (v) $\mathcal{S} \to (\mathcal{T} \to \mathcal{T})$ and $\mathcal{T} \to (\mathcal{S} \to \mathcal{S})$ (further with inter-domain mixing, i.e. after QuadMix). The effects of these strategies on video frames, optical flow, and labels/pseudo-labels are illustrated. We present $x_t^{\mathcal{T} \to (\mathcal{S} \to \mathcal{S})}$ and $x_t^{\mathcal{S} \to (\mathcal{T} \to \mathcal{T})}$ without masks for better understanding. Notably, the patch templates required in training iteration n are generated online adaptively from iteration $n-1$. Please zoom in for details.

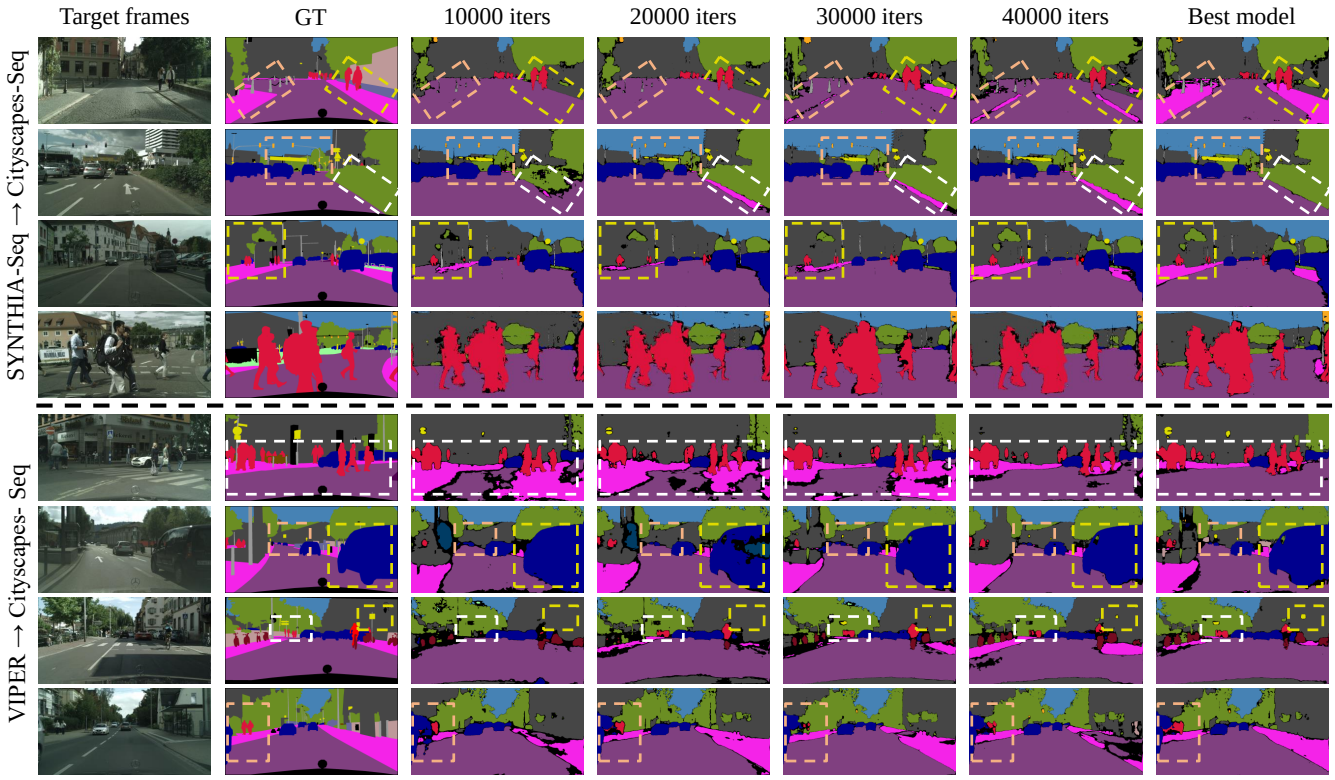


Fig. 5. Visualization of the evolution of pseudo-labels generated by our model with a **CNN** backbone on the target domain (Cityscapes-Seq). The pseudo-label quality improves progressively throughout training.

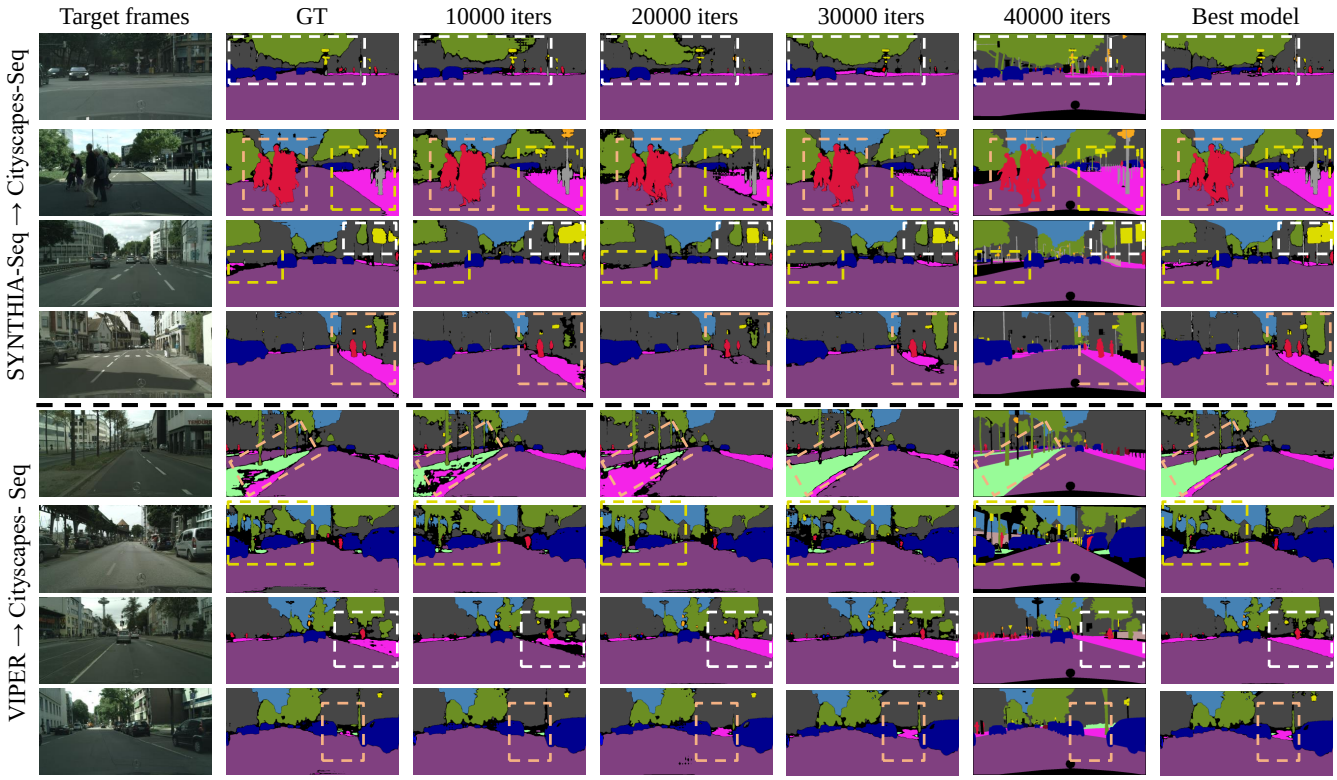


Fig. 6. Visualization of the evolution of pseudo-labels generated by our model with a **ViT** backbone on the target domain (Cityscapes-Seq). The pseudo-label quality improves progressively throughout training.

Fig. 7. Visual illustration 1 of the various mixing strategies employed in QuadMix on the **Viper** $\rightarrow$ **Cityscapes-Seq** benchmark. The final augmentation of frame t is shown without masks ($x_t^{\mathcal{T}\to(\mathcal{S}\to\mathcal{S})}$ and $x_t^{\mathcal{S}\to(\mathcal{T}\to\mathcal{T})}$) to better highlight the effects of QuadMix.

Fig. 8. Visual illustration 2 of the various mixing strategies employed in QuadMix on the **Viper** $\rightarrow$ **Cityscapes-Seq** benchmark. The final augmentation of frame t is shown without masks ($x_t^{\mathcal{T}\to(\mathcal{S}\to\mathcal{S})}$ and $x_t^{\mathcal{S}\to(\mathcal{T}\to\mathcal{T})}$) to better highlight the effects of QuadMix.

Fig. 9. Visual illustration 3 of the various mixing strategies employed in QuadMix on the **Viper** $\to$ **Cityscapes-Seq** benchmark. The final augmentation of frame t is shown without masks ($x_t^{\mathcal{T}\to(\mathcal{S}\to\mathcal{S})}$ and $x_t^{\mathcal{S}\to(\mathcal{T}\to\mathcal{T})}$) to better highlight the effects of QuadMix.

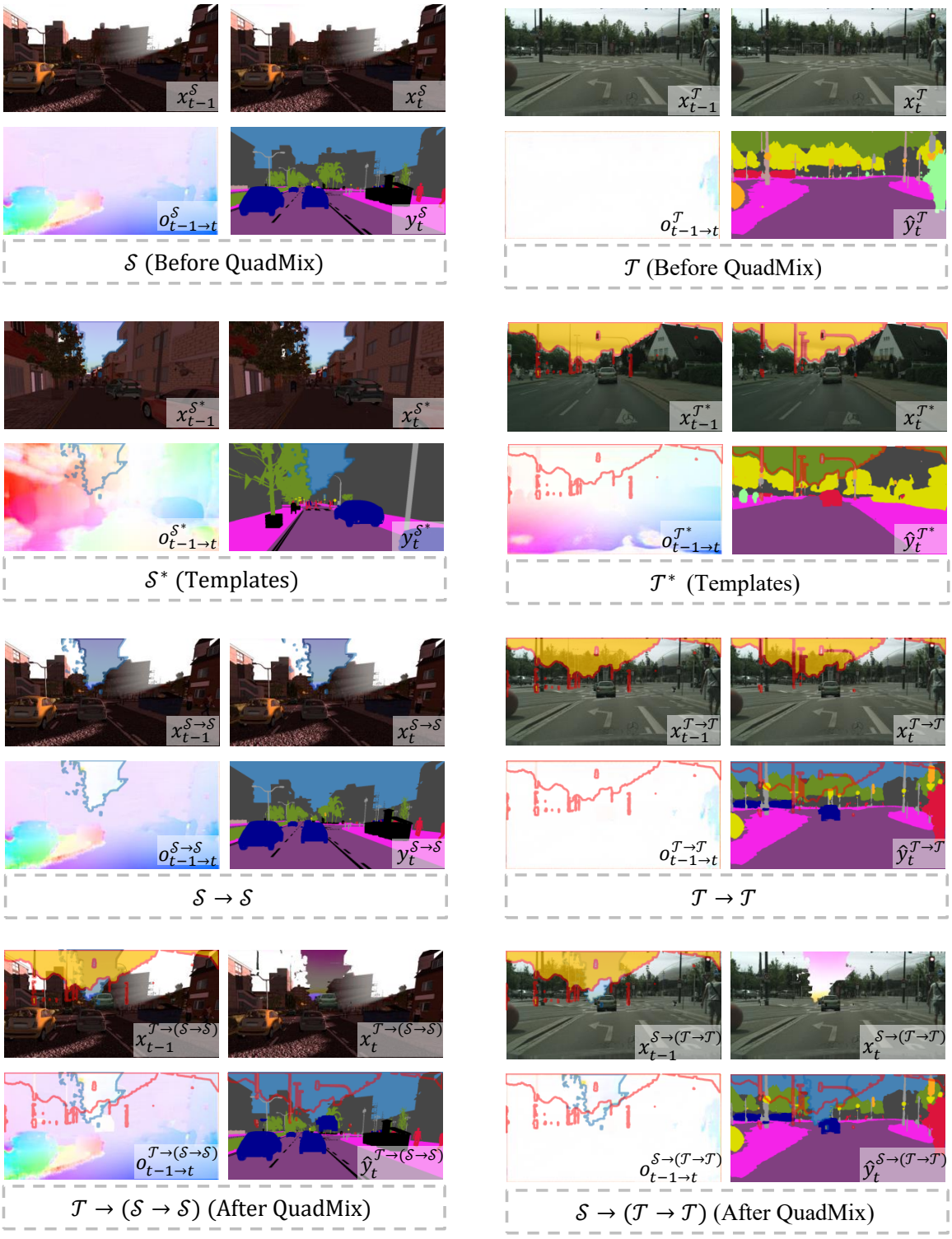


Fig. 10. Visual illustration 1 of the various mixing strategies employed in QuadMix on the **SYNTHIA-Seq** $\rightarrow$ **Cityscapes-Seq** benchmark. The final augmentation of frame t is shown without masks ($x_t^{\mathcal{T}\to(\mathcal{S}\to\mathcal{S})}$ and $x_t^{\mathcal{S}\to(\mathcal{T}\to\mathcal{T})}$) to better highlight the effects of QuadMix.

Fig. 11. Visual illustration 2 of the various mixing strategies employed in QuadMix on the **SYNTHIA-Seq** $\rightarrow$ **Cityscapes-Seq** benchmark. The final augmentation of frame t is shown without masks ($x_t^{\mathcal{T}\to(\mathcal{S}\to\mathcal{S})}$ and $x_t^{\mathcal{S}\to(\mathcal{T}\to\mathcal{T})}$) to better highlight the effects of QuadMix.

Fig. 12. Visual illustration 3 of the various mixing strategies employed in QuadMix on the **SYNTHIA-Seq** $\rightarrow$ **Cityscapes-Seq** benchmark. The final augmentation of frame t is shown without masks ($x_t^{\mathcal{T}\to(\mathcal{S}\to\mathcal{S})}$ and $x_t^{\mathcal{S}\to(\mathcal{T}\to\mathcal{T})}$) to better highlight the effects of QuadMix.

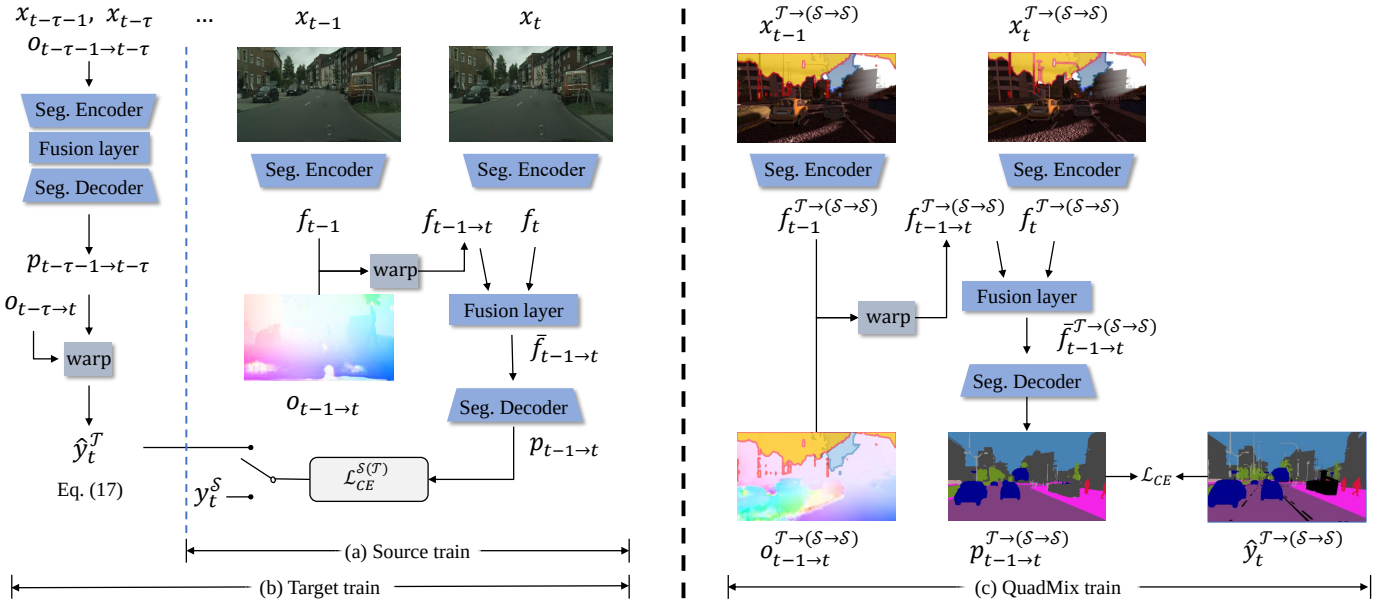


Fig. 13. Illustration of video pseudo-label generation, and the role of temporal information (i.e., optical flow) in the model's learning of QuadMixed video features.

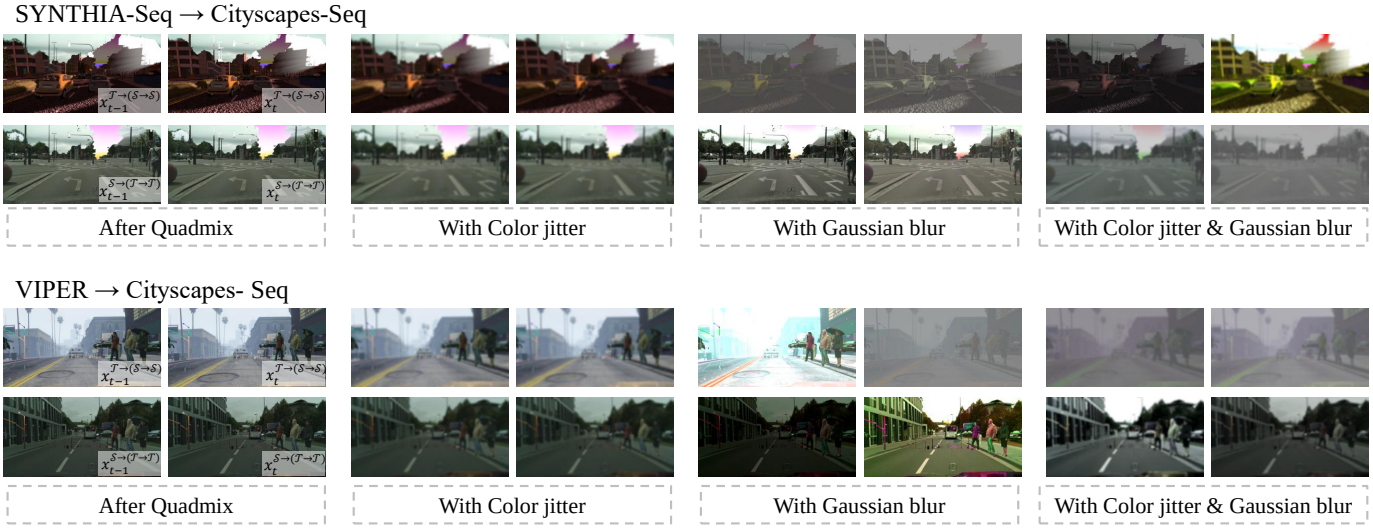


Fig. 14. Visualization of video frame enhancement (random Gaussian blur or color jitter) on two video UDA-SS benchmarks.

TABLE VII
DETAILED EXPERIMENTAL RESULTS REGARDING CATEGORY-AWARE IOU AND MIOU ON SYNTHIA-SEQ → CITYSCAPES-SEQ.

Synthia-Seq → Cityscapes-Seq												
Methods	road	side.	buil.	pole	light	sign	vege.	sky	pers.	rider	car	mIOU
QuadMix (ViT)	94.1	61.9	82.9	36.9	41.0	59.1	85.2	85.6	64.3	37.8	90.3	**67.19**
QuadMix(DL-V2)	90.8	39.9	83.2	33.2	30.1	50.7	84.8	82.3	61.2	32.7	87.4	61.48
w/o Long-tail (DL-V2)	90.2	48.3	82.8	31.5	27.6	46.6	83.5	81.7	60.5	27.2	88.0	60.72
Table V-0*	89.7	47.1	73.2	20.2	14.0	29.6	69.4	69.2	55.6	19.6	84.5	52.01
Table V-1	90.0	48.3	73.0	29.9	18.1	41.6	73.7	72.6	58.0	22.5	85.1	55.71
Table V-2	87.5	44.7	75.4	34.1	23.2	44.7	78.9	76.8	58.7	24.9	80.9	57.25
Table V-3	85.7	43.4	75.3	25.6	19.0	37.1	73.2	69.3	58.9	21.1	85.5	54.01
Table V-4	92.5	43.1	81.7	28.9	22.1	49.5	80.1	84.4	62.1	26.4	85.0	59.62
Table V-5	90.1	44.0	78.3	30.5	26.0	42.1	79.2	72.5	54.6	33.4	86.7	57.95
Table V-6	89.2	43.8	78.2	36.4	22.9	38.1	72.0	70.7	54.5	25.3	84.6	55.97
Table V-7	91.9	54.1	79.5	32.7	25.2	42.5	80.9	76.1	60.3	34.4	85.3	60.26
Table V-8	90.8	39.9	83.2	33.2	30.1	50.7	84.8	82.3	61.2	32.7	87.4	61.48
Table VI-0*	89.2	43.8	78.2	36.4	22.9	38.1	72.0	70.7	54.5	25.3	84.6	55.97
Table VI-1	89.7	55.0	80.8	30.1	25.0	42.2	80.4	69.6	57.7	25.9	82.8	58.11
Table VI-2	89.3	29.9	81.5	24.3	25.3	45.0	82.8	83.7	60.5	26.8	87.2	57.84
Table VI-3	91.2	47.2	81.3	27.0	16.6	40.5	82.9	83.1	58.4	30.8	87.5	58.77
Table VI-4	90.4	44.1	79.4	33.4	26.9	45.7	80.6	73.8	60.4	27.7	86.4	58.98
Table VI-5	92.1	52.5	78.3	21.2	16.4	42.8	78.5	75.9	56.6	23.1	86.0	56.67
Table VI-6	89.0	49.1	80.2	30.9	20.9	45.7	80.3	79.2	59.2	31.9	85.2	59.24
Table VI-7	91.3	53.2	81.5	28.3	22.3	42.0	80.3	74.3	57.8	25.6	85.7	58.39
Table VI-8	90.8	43.9	80.4	28.6	24.5	45.3	81.5	81.1	60.2	30.6	88.1	59.55
Table VI-9	91.5	45.4	81.1	31.3	22.2	46.5	84.5	82.9	61.9	27.6	87.3	60.20
Table VI-10	92.2	54.4	80.3	24.5	21.3	45.8	81.8	83.7	60.4	21.5	87.4	59.39
Table VI-11	91.9	54.1	79.5	32.7	25.2	42.5	80.9	76.1	60.3	35.8	85.3	60.39
Table VI-12	90.8	39.9	83.2	33.2	30.1	50.7	84.8	82.3	61.2	32.7	87.4	61.48
Table VII-Things	90.6	51.6	75.2	31.2	32.1	49.9	75.9	62.4	61.6	29.2	84.1	58.53
Table VII-Stuffs	91.9	54.6	82.2	23.8	23.4	40.8	81.2	75.3	55.0	18.1	86.2	57.50
Table VII-Move	92.2	58.4	77.7	30.7	23.7	40.3	78.0	67.1	60.6	29.3	83.2	58.29
Table VII-Station	90.5	53.7	71.0	30.1	29.3	52.0	71.7	38.4	61.1	27.3	85.9	55.55
Table VII-All	90.8	39.9	83.2	33.2	30.1	50.7	84.8	82.3	61.2	32.7	87.4	61.48
Table VIII-w/o $\mathcal{A}$	89.9	38.0	82.2	25.0	28.0	50.5	83.6	83.5	61.1	33.0	87.7	60.23
Table VIII-global	91.0	50.9	81.0	23.9	24.7	43.1	81.1	76.6	57.2	24.5	86.7	58.24
Table X-ResNet-Origin	77.8	24.4	70.2	4.7	0.0	0.3	50.6	36.1	30.6	0.5	57.0	32.01
Table X-ResNet	90.6	41.2	81.1	29.3	23.1	47.5	82.7	83.8	61.1	28.4	87.0	59.62
Table X-DA-VSN	89.0	25.8	81.6	34.2	23.2	50.3	82.1	83.9	62.4	30.8	86.5	59.07
Table X-CMOM	90.7	42.6	83.2	29.3	27.2	48.8	84.4	81.2	59.4	30.2	86.7	60.33
Table X-TPS	90.8	39.9	83.2	33.2	30.1	50.7	84.8	82.3	61.2	32.7	87.4	61.48

TABLE VIII
DETAILED EXPERIMENTAL RESULTS REGARDING CATEGORY-AWARE IOU AND MIOU ON VIPER → CITYSCAPES-SEQ.

VIPER → Cityscapes-Seq																
Methods	road	side.	buil.	fenc.	light	sign	vege.	terr.	sky	pers.	car	truc.	bus	mot.	bike	mIOU
QuadMix (ViT)	87.3	43.8	87.3	25.2	40.0	36.9	86.7	20.8	90.3	65.8	86.8	48.6	65.6	37.6	49.7	**58.16**
QuadMix (DL-V2)	91.6	51.4	87.0	24.1	32.3	37.2	84.1	28.4	84.8	64.4	85.7	41.4	46.5	34.0	49.6	56.17
w/o Long-tail (DL-V2)	91.9	51.8	85.7	25.2	26.4	36.8	83.1	28.9	86.2	63.3	84.6	45.4	45.7	32.7	41.8	55.31
Table VIII-w/o $\mathcal{A}$	91.9	50.2	86.5	26.1	27.3	35.6	82.4	27.9	85.8	64.2	83.9	46.1	42.3	33.4	45.2	55.25
Table VIII-global	91.2	43.4	85.4	19.6	29.5	35.9	83.5	28.4	85.2	65.6	85.6	38.5	35.1	33.5	43.9	53.62
Table IX-1	91.2	40.0	85.8	22.3	25.7	33.9	83.8	28.7	86.8	62.6	85.4	42.1	49.7	34.6	45.9	54.56
Table IX-2	91.6	51.4	87.0	24.1	32.3	37.2	84.1	28.4	84.8	64.4	85.7	41.4	46.5	34.0	49.6	56.16
Table IX-3	90.5	49.2	83.7	21.5	22.3	30.3	81.8	28.4	87.0	58.7	86.1	47.6	51.3	33.2	45.7	54.49
Table IX-4	90.8	49.8	86.2	25.5	26.6	33.3	83.6	29.0	85.4	63.5	84.5	37.4	46.0	34.2	42.5	54.56
Table IX-5	92.4	51.6	86.8	25.2	29.4	29.4	84.7	32.5	86.7	63.9	83.1	29.8	43.6	33.0	46.6	54.58
Table IX-6	92.2	54.8	86.1	29.2	27.0	35.0	85.0	32.8	85.6	63.6	82.0	25.0	48.1	34.0	47.7	55.21
Table X-ResNet-Origin	75.1	6.5	72.8	0.0	0.0	0.0	71.1	1.1	58.4	11.2	63.0	9.8	0.0	0.0	0.0	24.60
Table X-ResNet	90.8	39.4	84.6	27.2	23.9	34.2	82.4	29.2	85.8	64.3	84.5	43.6	50.6	34.7	48.8	54.94
Table X-DA-VSN	90.9	52.1	85.9	24.2	26.6	37.2	82.5	27.6	85.6	60.1	86.4	46.4	45.3	33.6	45.3	55.31
Table X-CMOM	91.6	53.6	86.1	17.3	28.0	42.1	86.4	38.8	88.0	64.7	83.6	41.3	47.9	30.8	41.4	56.10
Table X-TPS	91.6	51.4	87.0	24.1	32.3	37.2	84.1	28.4	84.8	64.4	85.7	41.4	46.5	34.0	49.6	56.16