Spatio-Temporal Pixel-Level Contrastive Learning-based Source-Free Domain Adaptation for Video Semantic Segmentation

WED-AM-220



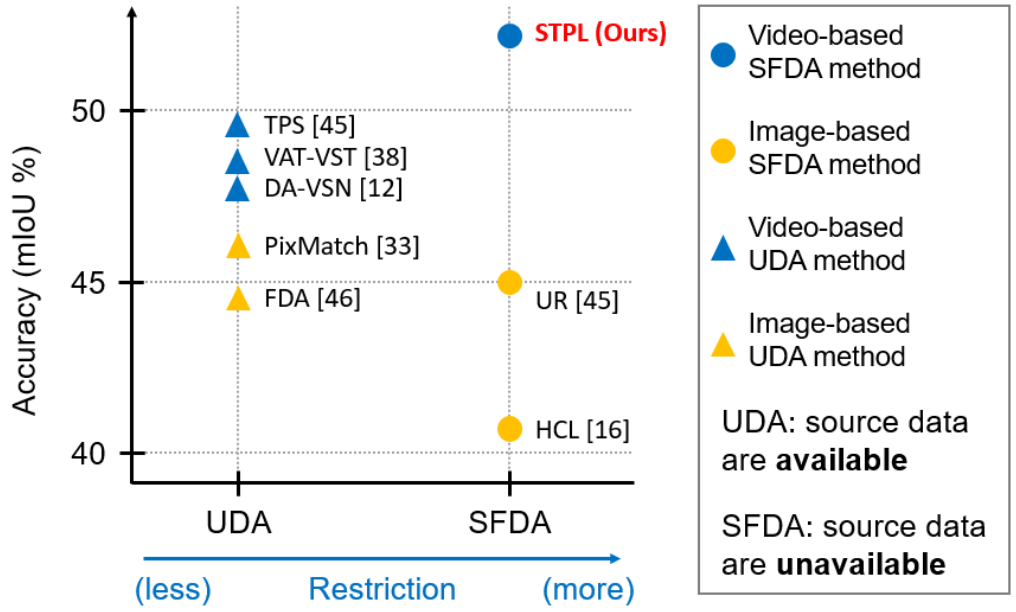


Shao-Yuan Lo Poojan Oza Sumanth Chennupati Alejandro Galindo Vishal M. Patel

CVPR 2023

Contributions

- We propose the **first** Source-Free Domain Adaptation (**SFDA**) method for Video Semantic Segmentation (**VSS**).
- The proposed method is based on a novel Spatio-Temporal Contrastive Learning (**STPL**) framework.
- The proposed STPL outperforms various state-of-the-art domain adaptation approaches (CVPR’21, ECCV’22, etc.).



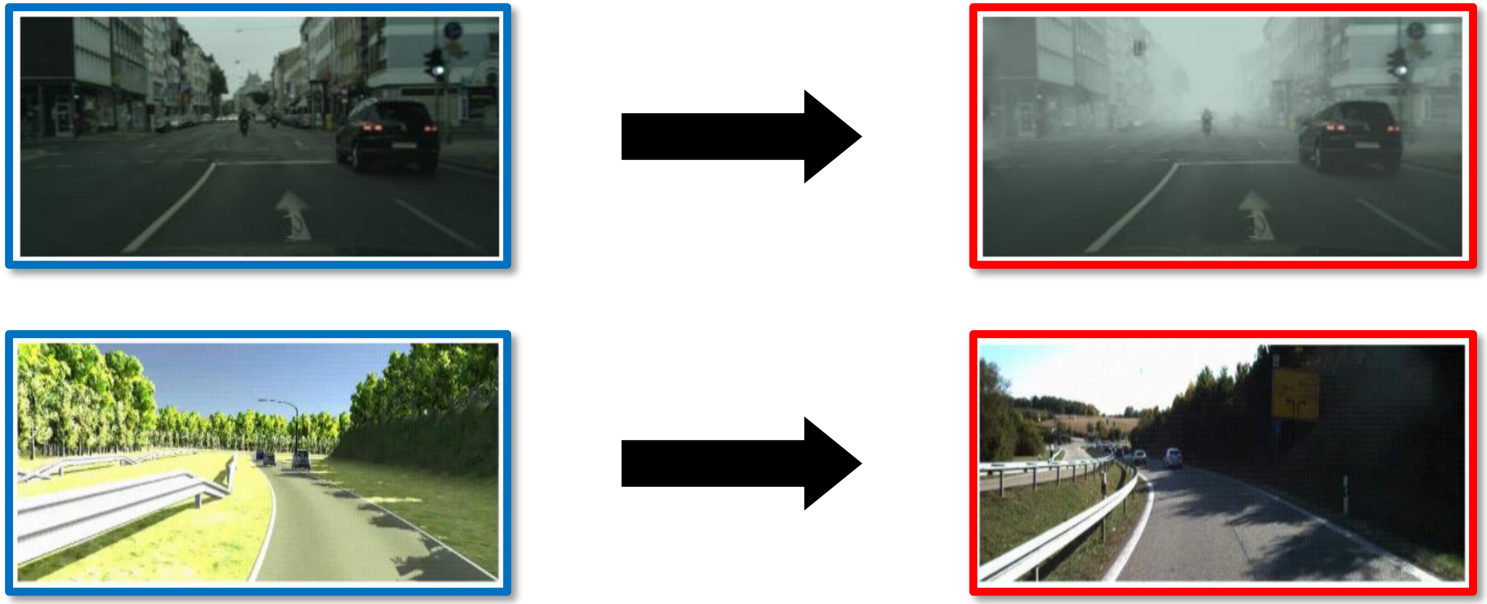
Recall: Video Semantic Segmentation

- Video semantic segmentation (VSS) aims to predict pixel-level semantics for each video frame.
- Compared to image semantic segmentation (ISS), **temporal information** can be exploited to improve either **accuracy** or **inference speed**.

[Jain et al. CVPR'19]

Recall: Domain Shifts

- **Scenario:** **Training (source) data** and **test (target) data** are from different domains (i.e. datasets).
- **Setting:** Given a **labeled source** dataset and an **unlabeled target** dataset, learn a model for the **target** domain.



Cityscapes → Foggy Cityscapes

Virtual KITTI → KITTI

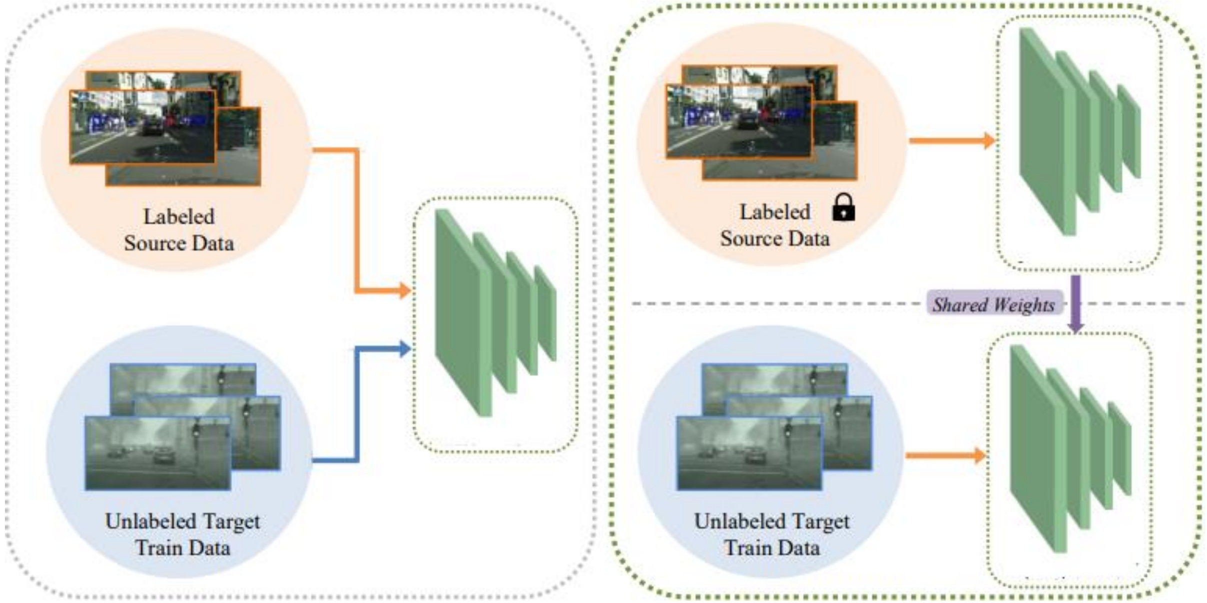
Source domain

Target domain

Source-Free Domain Adaptation

- **Scenario:** Training (source) and test (target) data are from different domains, and **we cannot access to the source data** (e.g. privacy).
- **Setting:** Given a **source-trained model** and an **unlabeled target dataset**, adapt the model to the **target** domain.

Classic domain adaptation



Source-free domain adaptation

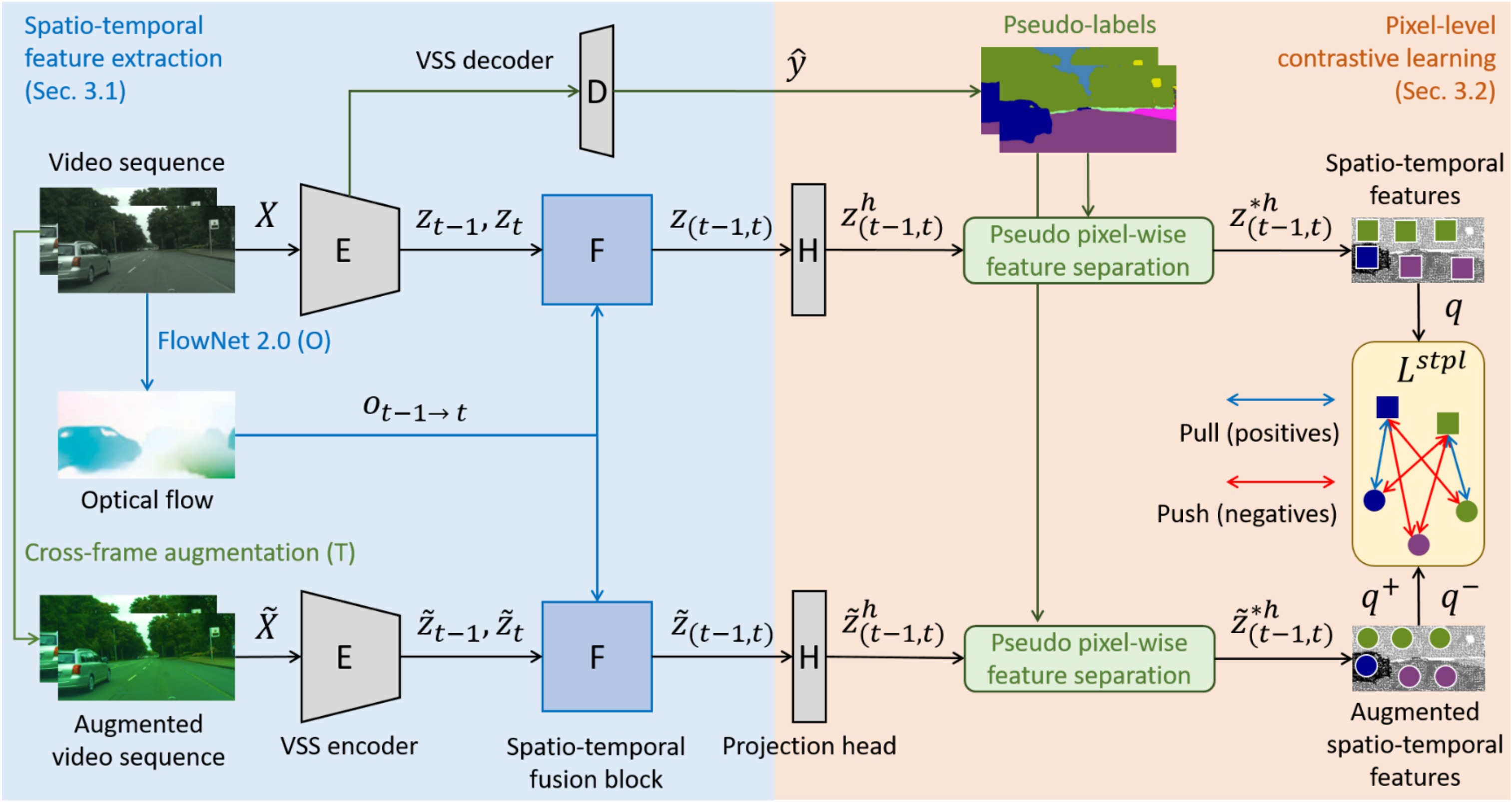
[VS et al. WACV'23]

Challenges

- Classic domain adaptation (UDA) for VSS methods are **not applicable** to the source-free domain adaptation (SFDA) setting.
- SFDA for ISS methods do not consider the **temporal information**.
- No access to any labeled training data.

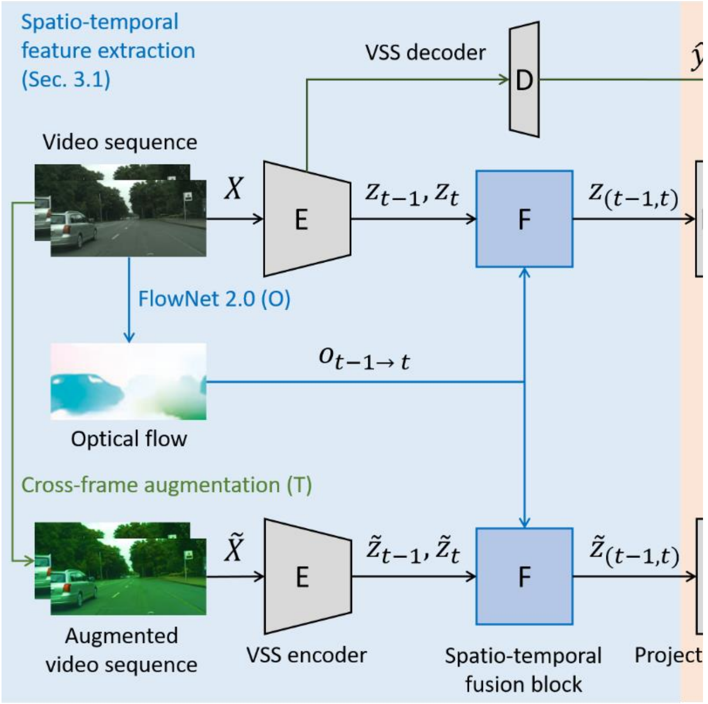
Spatio-Temporal Pixel-Level Contrastive Learning

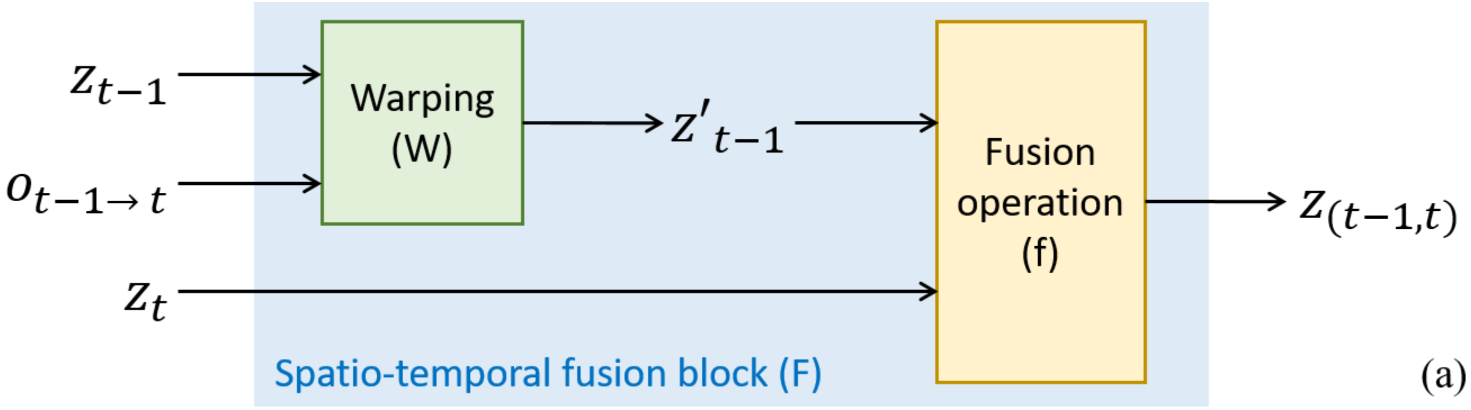
- Spatio-temporal feature extraction
- Pixel-level contrastive learning



Spatio-Temporal Feature Extraction

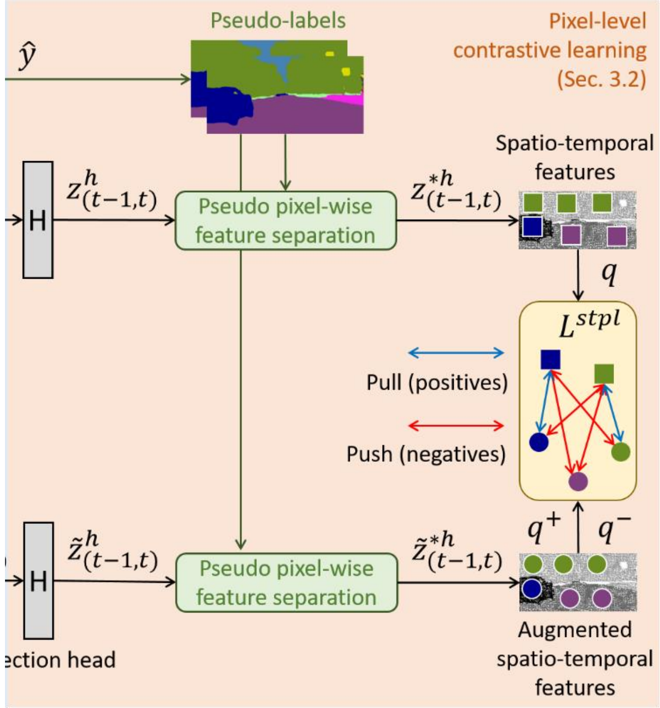
- Spatio-temporal fusion block
  - Feature warping by **optical flow** (temporal information)
  - Fusion operation: concatenation, element-wise addition, 1x1 convolution, attention module, etc.





Pixel-Level Contrastive Learning

- Pseudo-labels are used for **pseudo pixel-wise feature separation**
- **Positive samples**: Pixels of the **same** semantic class
- **Negative samples**: Pixels of **different** semantic classes



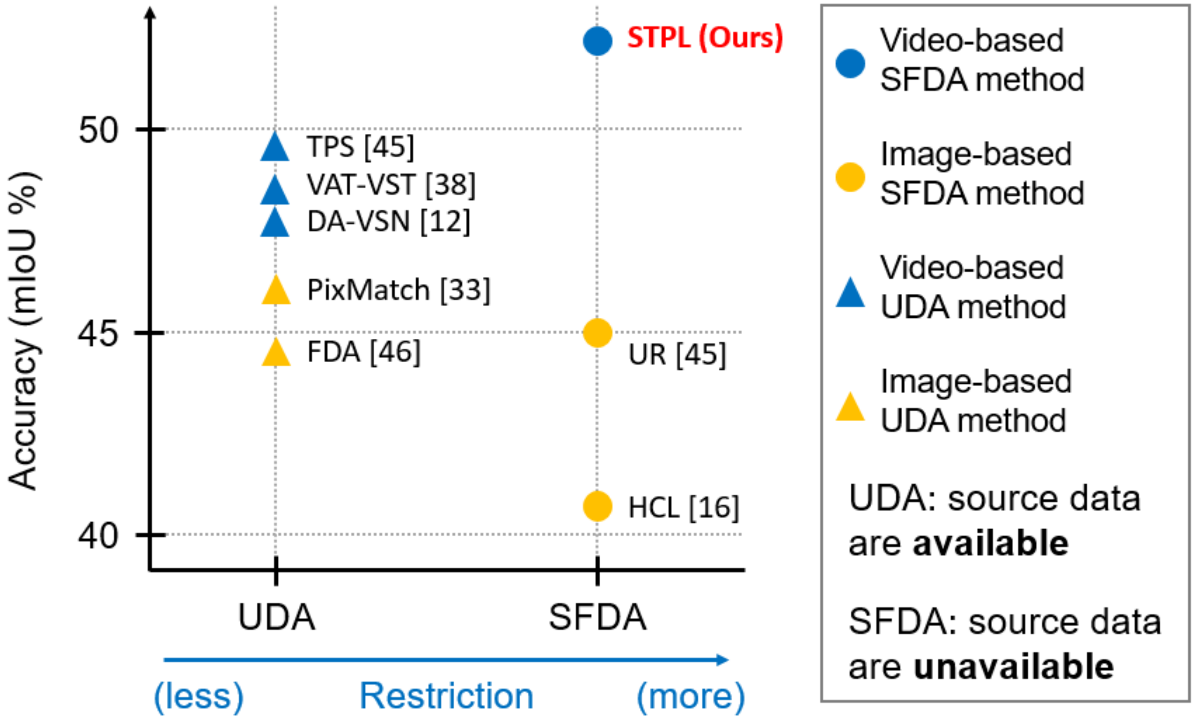
Pixel-wise SimCLR

$$\mathcal{L}_q^{stpl} = \frac{-1}{|P_q|} \sum_{q^+ \in P_q} \log \frac{\exp(q \cdot q^+ / \tau)}{\sum_{q^- \in N_q} \exp(q \cdot q^- / \tau)}$$

Results

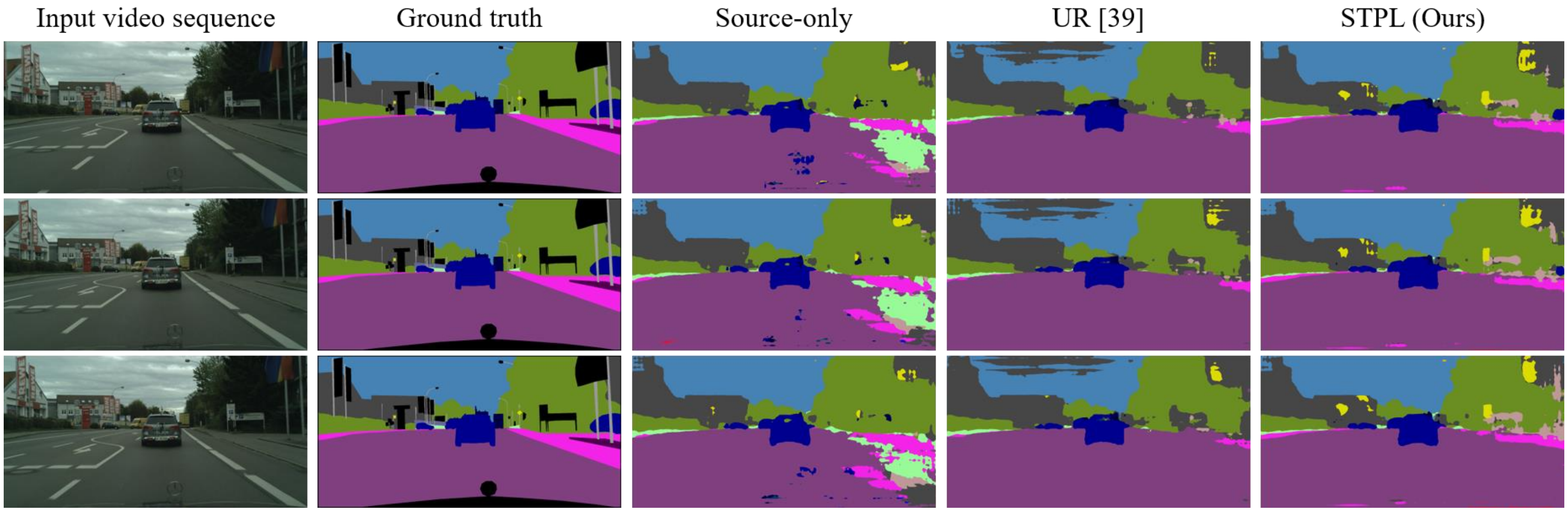
- Benchmark: VIPER → Cityscapes-Seq

Method	Design	DA	mIoU
Source-only	-	-	37.1
FDA [46] (CVPR'20)	Image	UDA	44.4
PixMatch [33] (CVPR'21)	Image	UDA	46.7
RDA [17] (ICCV'21)	Image	UDA	44.4
UR [39] (CVPR'21)	Image	SFDA	45.0
HCL [16] (NeurIPS'21)	Image	SFDA	41.5
DA-VSN [12] (ICCV'21)	Video	UDA	47.8
VAT-VST [38] (AAAI'22)	Video	UDA	48.7
TPS [45] (ECCV'22)	Video	UDA	48.9
DA-VSN* [12] (ICCV'21)	Video	SFDA	45.3
VAT-VST* [38] (AAAI'22)	Video	SFDA	43.6
TPS* [45] (ECCV'22)	Video	SFDA	27.8
STPL (Ours)	Video	SFDA	52.5
Oracle	-	-	69.9



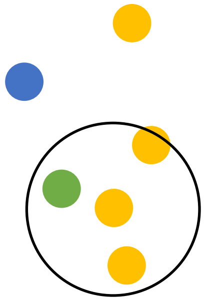
Results

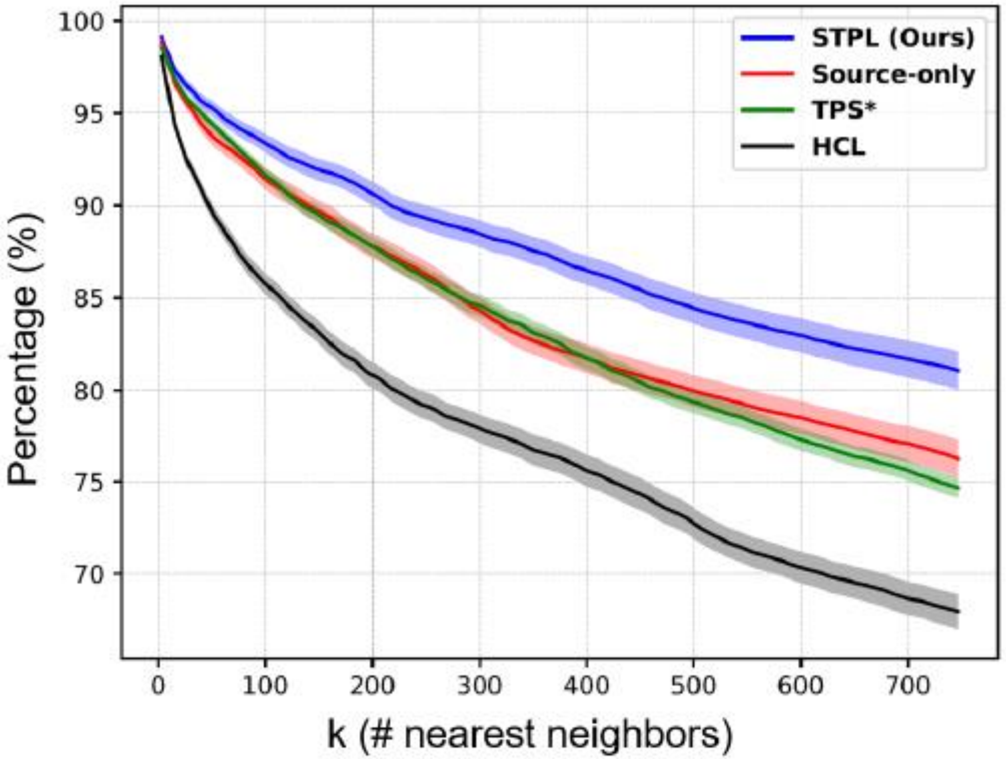
- Benchmark: VIPER → Cityscapes-Seq



Analysis

- The percentage of same-class pixel representations among the k-nearest neighbors in the feature space.





Conclusion

- We propose the **first** Source-Free Domain Adaptation (**SFDA**) method for Video Semantic Segmentation (**VSS**).
- The proposed method is based on a novel Spatio-Temporal Contrastive Learning (**STPL**) framework.
- The proposed STPL outperforms various state-of-the-art domain adaptation approaches (CVPR'21, ECCV'22, etc.).