JOHNS HOPKINS UNIVERSITY
amazon

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

Shao-Yuan Lo[1], Poojan Oza[2], Sumanth Chennupati[2], Alejandro Galindo[2], Vishal M. Patel[1]
[1]Johns Hopkins University, [2]Amazon

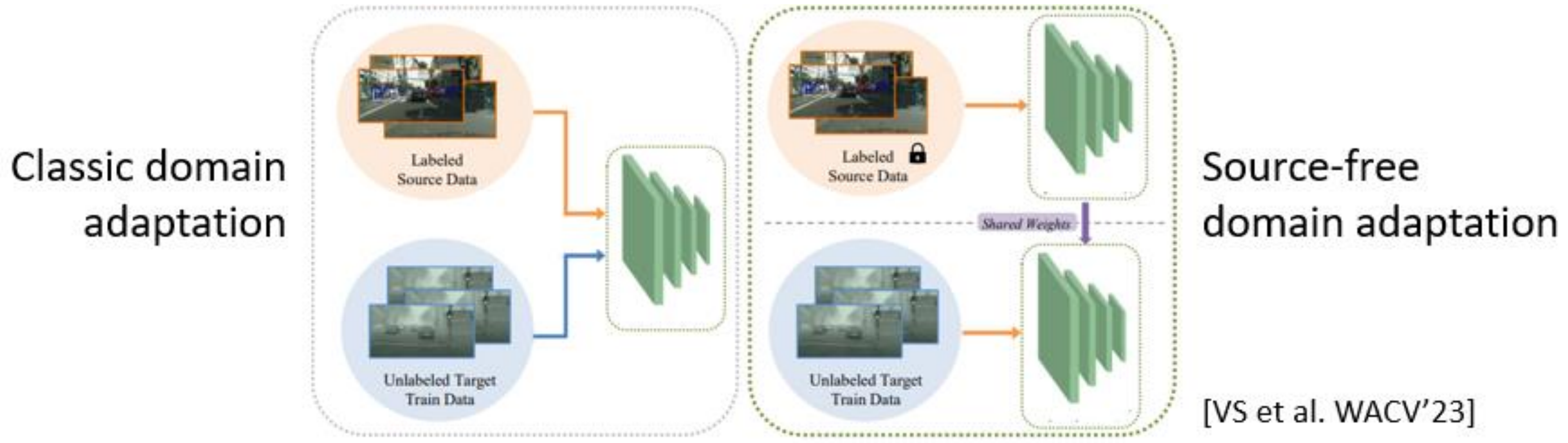
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

## 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.).

## Background

- Consider the scenario that the **training (source) data** and **test (target) data** are from different domains (i.e. datasets). This would cause accuracy drop on target data due to the **domain shift** problem.
- **UDA**: Given a **labeled source dataset** and an **unlabeled target dataset**, learn a model for the **target** domain.
- **SFDA**: Given a **source-trained model** and an **unlabeled target dataset**, adapt the model to the **target** domain.
- SFDA does not require the access to source datasets, which are usually private or restrict.
- SFDA is more transmission efficient since a source-trained model is usually much smaller than a source dataset.

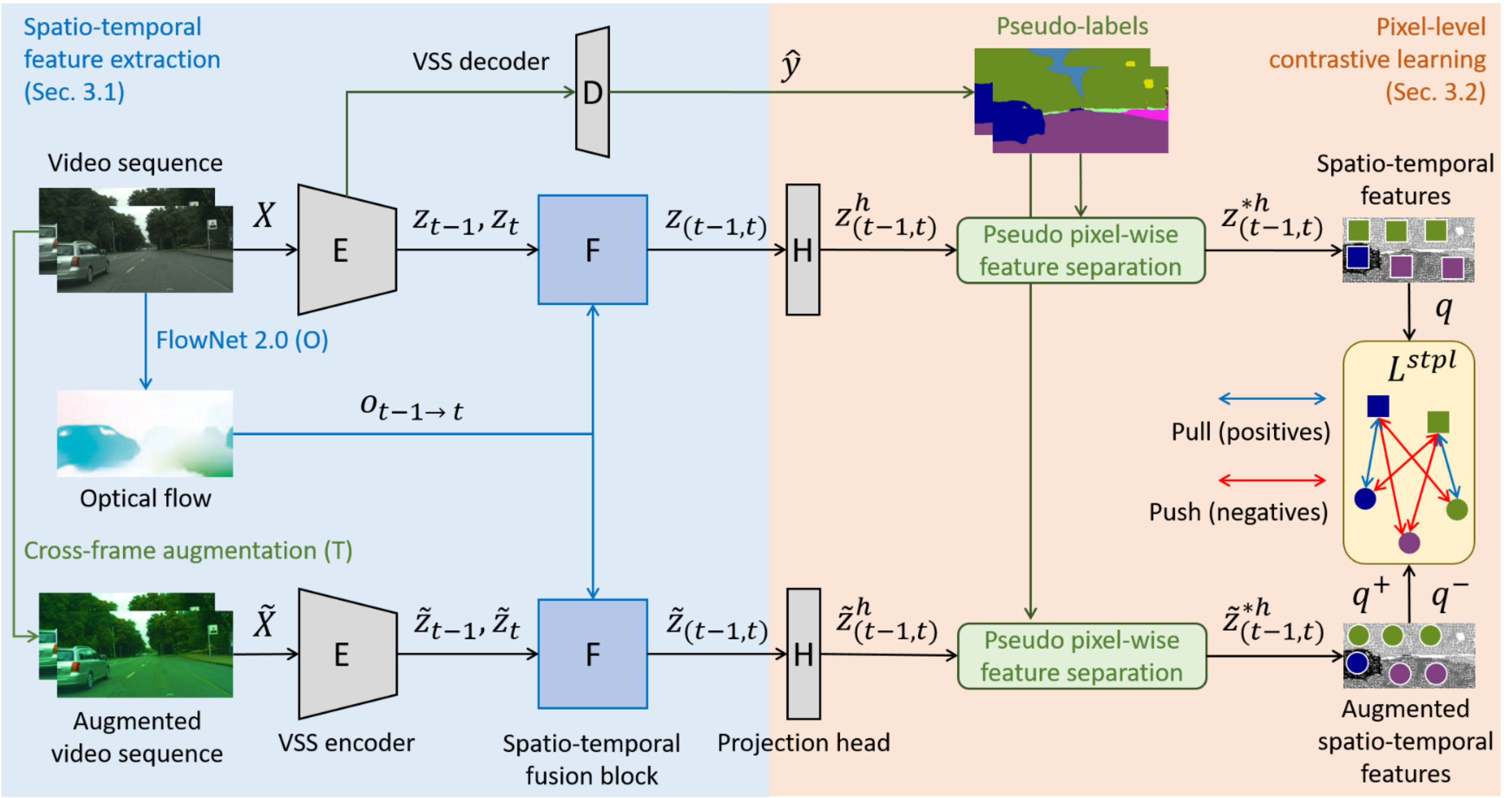



[VS et al. WACV'23]

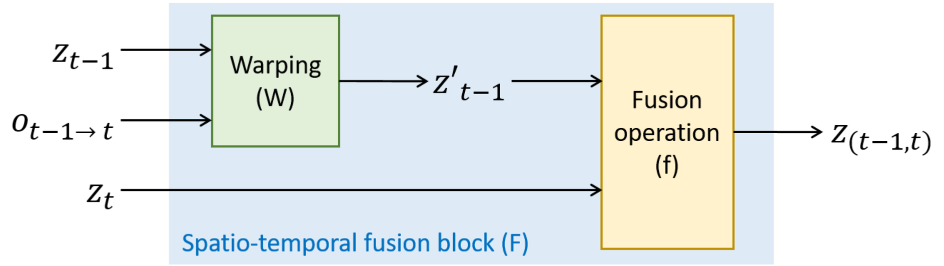
## Challenges

- SFDA has not been explored for video data.
- Existing UDA for VSS methods are **not applicable** to the SFDA setting.
- SFDA methods for image data do not consider **temporal information**.
- No access to **any** labeled training data.

## Method




- Spatio-temporal feature extraction
  - Feature warping by **optical flow** (temporal information)
  - Fusion operation: concatenation, 1x1 convolution, attention module, etc.

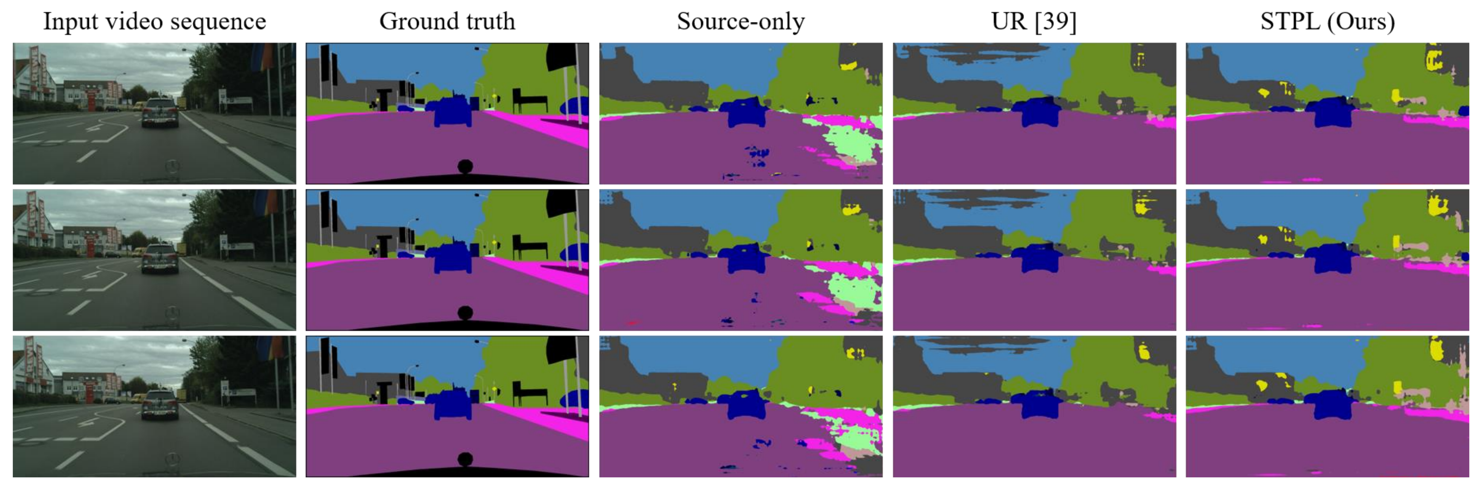



$$\mathcal{L}_q^{stpl} = \frac{-1}{|P_q|} \sum_{q^+ \in P_q} \log \frac{\exp(q \cdot q^+/\tau)}{\sum_{q^- \in N_q} \exp(q \cdot q^-/\tau)}$$

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

## Results

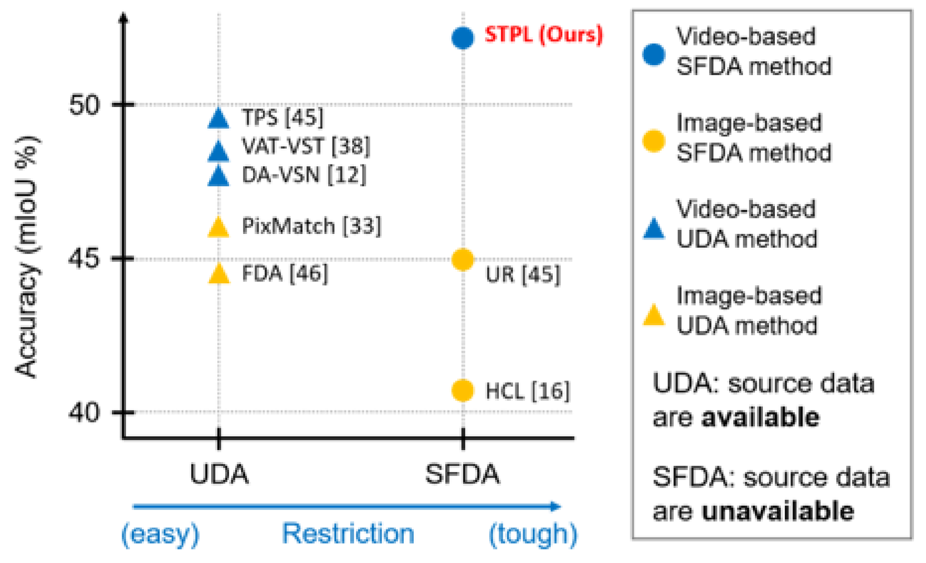
- Qualitative results




## Results

- Quantitative results

| 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 |




- Ablation study

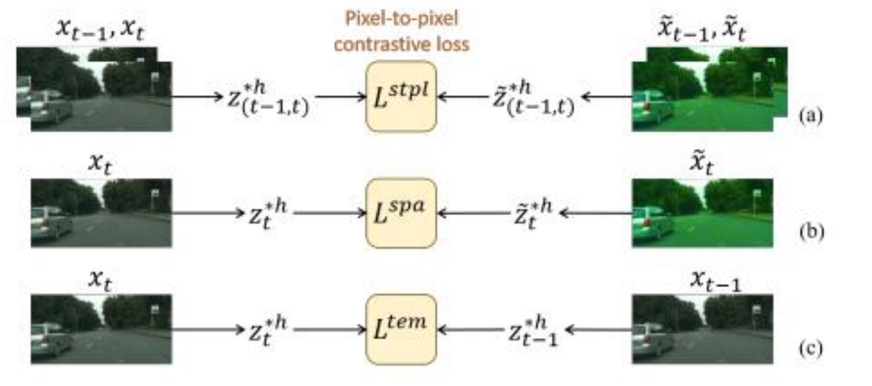



Figure 4. Illustration of (a) the proposed spatio-temporal contrast $\mathcal{L}^{stpl}$ (Eq. (3), (4)), (b) spatial-only contrast $\mathcal{L}^{spa}$, and (c) temporal-only contrast $\mathcal{L}^{tem}$.

| Method / Objective function | mIoU |
|---|---|
| Source-only | 37.1 |
| Vanilla Self-training | 45.4 (+8.3) |
| Duplicate CL | 45.7 (+8.6) |
| Temporal-only CL ($\mathcal{L}^{tem}$) | 47.4 (+10.3) |
| Spatial-only CL ($\mathcal{L}^{spa}$) | 51.1 (+14.0) |
| Naïve T+S CL ($\mathcal{L}^{tem} + \mathcal{L}^{spa}$) | 51.4 (+14.3) |
| STPL (Ours; $\mathcal{L}^{stpl}$) | **52.5 (+15.4)** |

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