# Video Semantic Segmentation via Sparse Temporal Transformer

Jiangtong Li, Wentao Wang, Junjie Chen, Li Niu, Jianlou Si, Chen Qian, Liqing Zhang

**MoE Key Lab of Artificial Intelligence,**
**Department of Computer Science and Engineering,**
**Shanghai Jiao Tong University**

**SenseTime Research,**
**SenseTime**

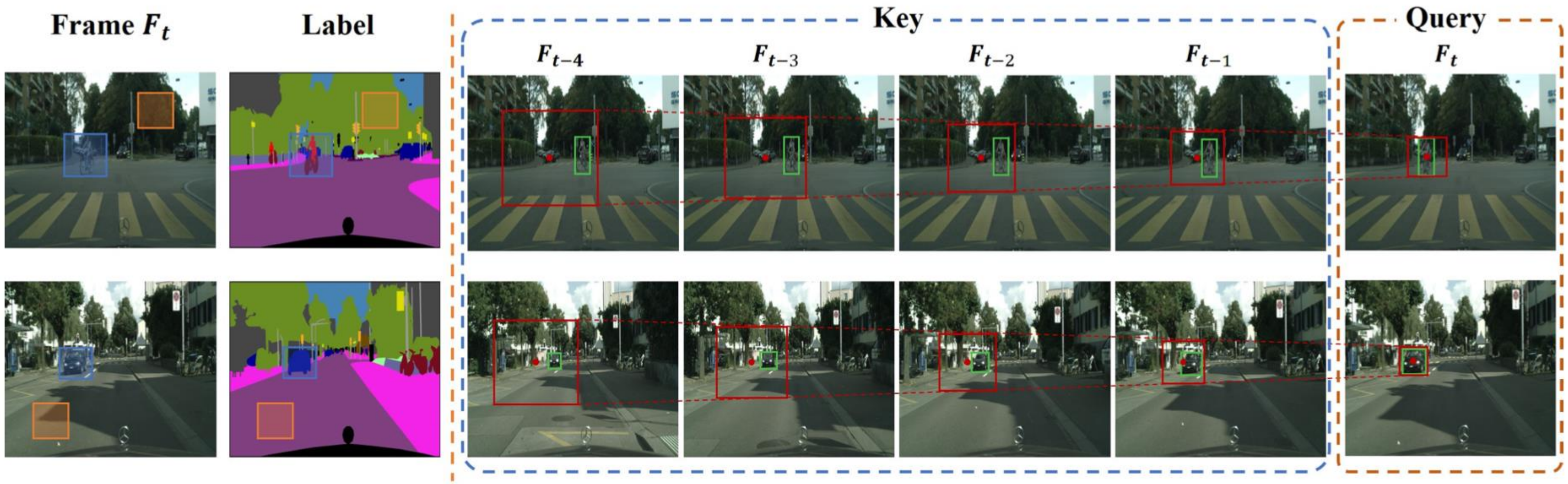
# Challenges & Previous Solution

- **Challenge 1.** The demand of temporal consistency in a semi-supervised manner
    - NetWarp [1] and GRFP [2] estimated frame-to-frame motion warping (e.g., optical flow) to segment consecutive frames
    - ETC [3] adopted warped prediction loss to constrain the prediction of current frame during training and performed single-frame prediction during inference.
- **Challenge 2.** The balance between segmentation accuracy and inference efficiency for real-time applications
    - DVSNet [4] employed large models towards the key frames, and propagate to non-key frames using optical flows.
    - Accl [5] employed large models towards the key frames, and utilized small model to process the non-key frames.
    - TDNet [6] adopted knowledge distillation from large model towards small model to improve the segmentation efficiency without increasing the computational cost.

[1] Raghudeep Gadde, Varun Jampani, and Peter V Gehler. 2017. Semantic video cnns through representation warping. In ICCV 2017.
[2] David Nilsson and Cristian Sminchisescu. 2018. Semantic video segmentation by gated recurrent flow propagation. In CVPR 2018.
[3] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. 2020. Efficient Semantic Video Segmentation with Per-frame Inference. In ECCV 2020.
[4] Yu-Syuan Xu, Tsu-Jui Fu, Hsuan-Kung Yang, and Chun-Yi Lee. 2018. Dynamic video segmentation network. In CVPR 2018.
[5] Samvit Jain, Xin Wang, and Joseph E Gonzalez. 2019. Accel: A corrective fusion network for efficient semantic segmentation on video. In CVPR 2019.
[6] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. 2020. Temporally distributed networks for fast video semantic segmentation. In CVPR 2020.

# Our Solution

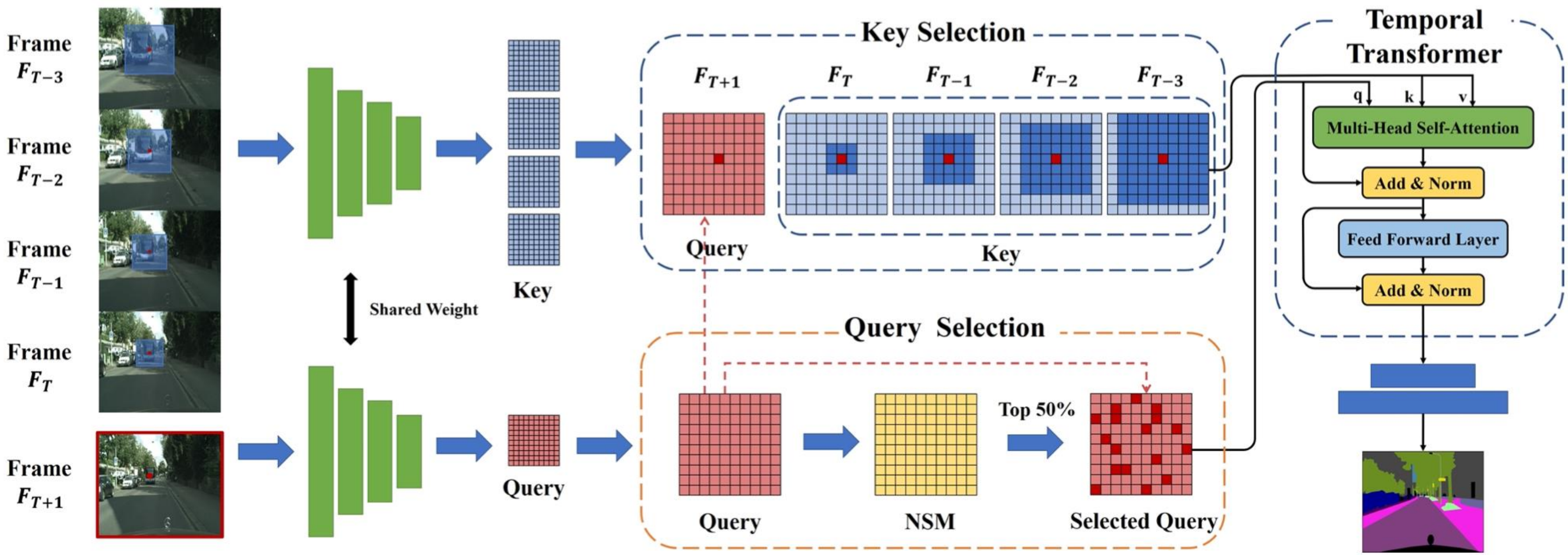
- For **Challenge 1.** The demand of temporal consistency in a semi-supervised manner
  - We propose to incorporate a temporal transformer into existing segmentation models as an adaptive module to capture the temporal relation among consecutive frames.
- For **Challenge 2.** The balance between segmentation accuracy and inference efficiency for real-time applications
  - We propose two selection strategies towards the temporal transformer framework (i.e., query selection and key selection).




(a) Query Selection (b) Key Selection

# Proposed Method

- Sparse Temporal Transformer



# Proposed Method

- Query Selection
  - Motivation: semantic boundary regions need more representation [1].

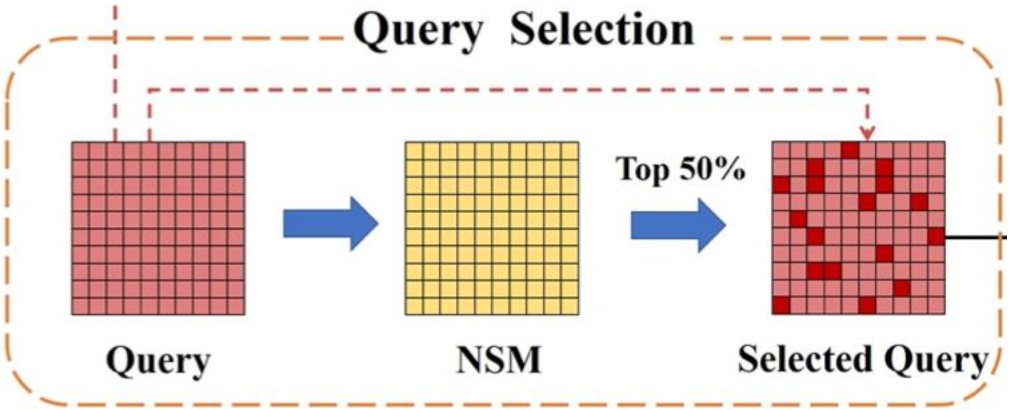
  - Identification of semantic boundary regions: the similarity between the feature region and its neighboring --- Neighboring Similarity Matrix (NSM)

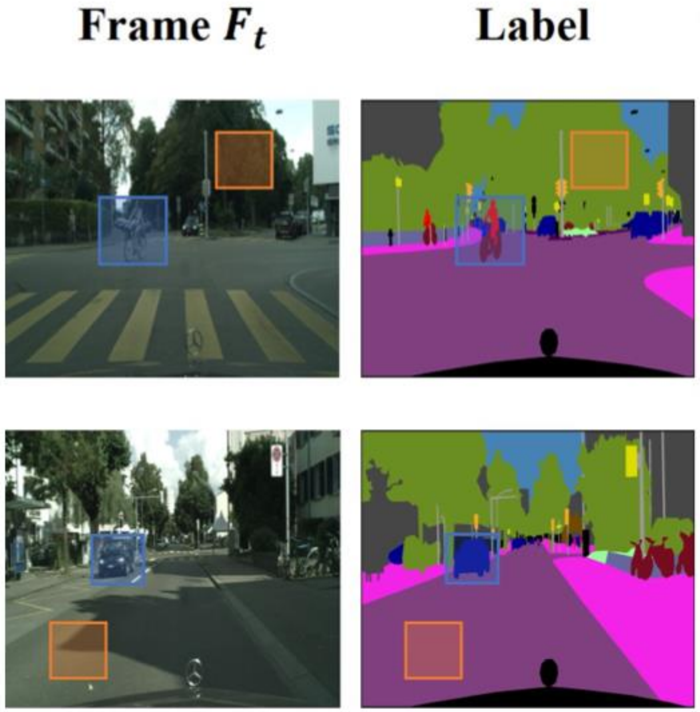
$$\mathbf{p}_{sim} = \mathrm{SoftMax}(\mathbf{Q^n} \cdot \mathbf{q}^T),$$

$$\mathcal{D}_{KL} = KL(\mathbf{p}_u || \mathbf{p}_{sim}) = \sum_{i=1}^{n_b} p_{u[i]} \log \frac{p_{sim[i]}}{p_{u[i]}},$$

$$\mathcal{D}_{cos} = \frac{1}{n_b} \sum_{i=1}^{n_b} \left(1 - \frac{\mathbf{Q^n_{[i]}} \cdot \mathbf{q}^T}{||\mathbf{Q^n_{[i]}}||_2 ||\mathbf{q}||_2}\right),$$

$$\mathcal{D}_{NSM} = \mathcal{D}_{KL} + \mathcal{D}_{cos},$$



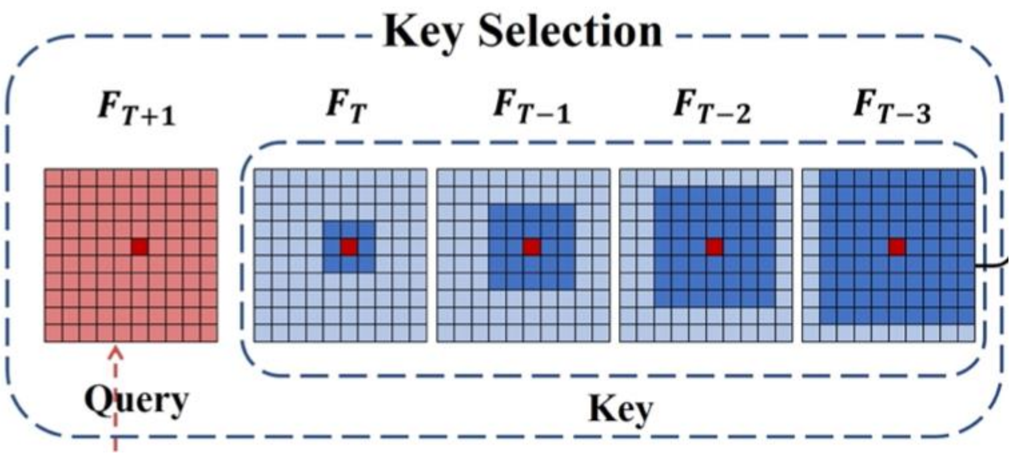

[1] Dmitrii Marin, Zijian He, Peter Vajda, Priyam Chatterjee, Sam Tsai, Fei Yang, and Yuri Boykov. 2019. Efficient segmentation: Learning downsampling near semantic boundaries. In ICCV 2019.

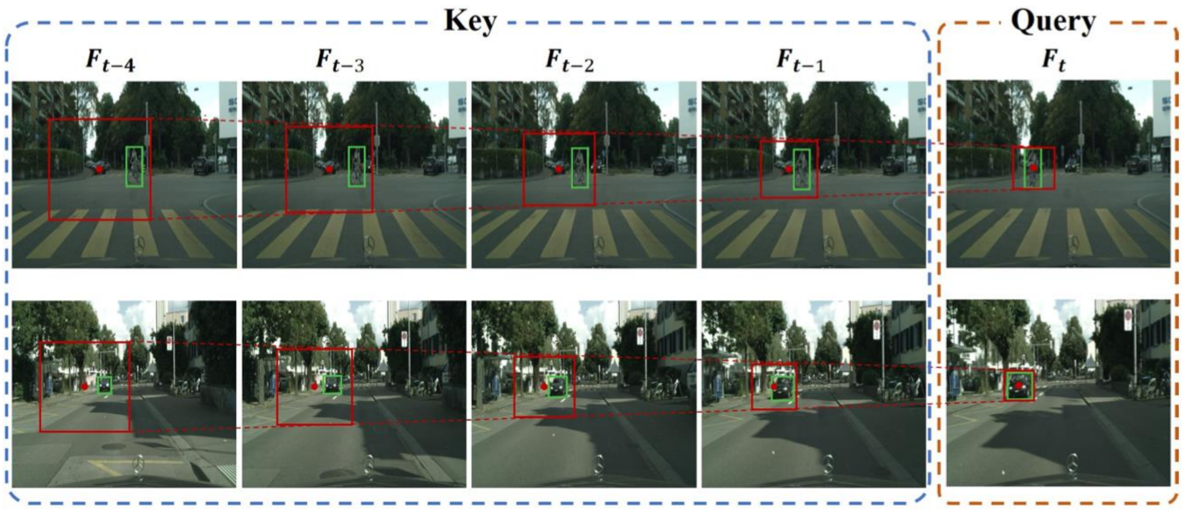
# Proposed Method

- Key Selection
  - Motivation: in consecutive frames, tracking the corresponding small regions in previous frames can bring much useful temporal information.

  - Rules for enlarging the searching regions:
    - The key frame farther from the current frame should have larger key region;
    - The size of key regions should vary within a proper range.

$$l_t = \begin{cases} s + (T - t) * \epsilon & , \quad \text{if} \quad s + (T - t) * \epsilon < e; \\ e & , \quad \text{otherwise}. \end{cases}$$



# Experimental Results

- Experiment Setup
  - Dataset
    - Cityscapes
      - Training: 2,975 video clips
      - Validation: 500 video clips
      - Test: 1,525 video clips
    - Camvid
      - Training: 367 video clips
      - Validation: 100 video clips
      - Test: 233 video clips
  - Evaluation Metrics
    - For segmentation accuracy: mean Intersection-over-Union (mIoU)
    - For temporal consistency: TC following ETC [1], which measures the consistency based on the mean flow warping error between all consecutive frames.

- Comparison with Existing Methods
  - High-Speed Methods

| Method | Backbone | mIoU (%) ↑ | TC (%)↑ | fps (frame/s) ↑ |
|---|---|---|---|---|
| DVSNet [50] | ResNet18 | 63.2 | - | 30.3 |
| ICNet [55] | ResNet50 | 67.7 | - | 50.0 |
| LadderNet [31] | DenseNet121 | 72.8 | - | 30.3 |
| SwiftNet [41] | ResNet18 | 75.4 | - | 43.5 |
| BiSeNet18 [53] | ResNet18 | 73.8 | - | 50.0 |
| BiSeNet34 [53] | ResNet34 | 76.0 | - | 37.0 |
| TDNet-BiSe18 [25] | ResNet18 | 75.0 | 70.2 | 47.6 |
| TDNet-BiSe34 [25] | ResNet34 | 76.4 | 71.1 | 38.5 |
| ETC-Mobi [37] | MobileNetV2 | 73.9 | 69.9 | 20.8 |
| STT-BiSe18 | ResNet18 | **75.8** | **71.4** | 44.2 |
| STT-BiSe34 | ResNet34 | **77.3** | **72.0** | 33.8 |

[1] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. 2020. Efficient Semantic Video Segmentation with Per-frame Inference. In ECCV 2020.

# Experimental Results

- Comparison with Existing Methods
  - High-Quality Methods

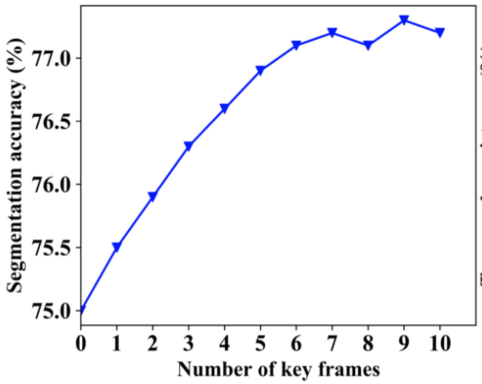
| Method | Backbone | Cityscapes | | | Camvid | | |
|---|---|---|---|---|---|---|---|
| | | mIoU (%) ↑ | TC (%) ↑ | fps (frame/s) ↑ | mIoU (%) ↑ | TC (%) ↑ | fps (frame/s) ↑ |
| NetWarp [20] | ResNet101 | 80.6 | - | 0.3 | 67.1 | - | 2.8 |
| DFF [61] | ResNet101 | 68.7 | 71.4 | 9.7 | - | - | - |
| GRFP [40] | ResNet101 | 69.4 | - | 3.2 | 66.1 | - | 4.4 |
| LVS [33] | ResNet101 | 76.8 | - | 5.9 | - | - | - |
| Accel [28] | ResNet101/18 | 72.1 | 70.3 | 3.6 | 66.7 | - | 7.6 |
| PSPNet18 [56] | ResNet18 | 75.5 | 68.5 | 10.8 | 71.0 | - | 24.4 |
| PSPNet50 [56] | ResNet50 | 78.1 | - | 4.2 | 74.7 | - | 8.5 |
| PSPNet101 [56] | ResNet101 | 79.4 | 69.7 | 2.1 | 77.6 | 77.1 | 4.1 |
| TDNet-PSP18 [25] | ResNet18 | 76.8 | 70.4 | 11.8 | 72.6 | 73.2 | 25.2 |
| TDNet-PSP50 [25] | ResNet50 | 79.9 | 71.1 | 5.6 | 76.0 | 77.4 | 11.1 |
| ETC-PSP18 [37] | ResNet18 | 73.1 | 70.6 | 10.8 | 75.2 | 77.3 | 24.4 |
| ETC-PSP101 [37] | ResNet101 | 79.5 | 71.7 | 2.1 | 79.4 | 78.6 | 4.1 |
| STT-PSP18 | ResNet18 | **77.3** | **73.0** | 11.5 | **76.1** | **81.4** | 24.7 |
| STT-PSP101 | ResNet101 | **82.5** | **73.9** | 2.2 | **80.2** | **82.3** | 4.2 |

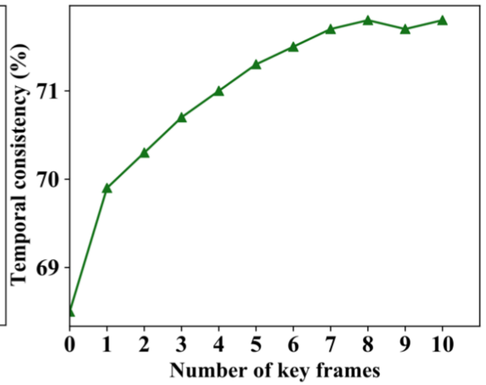
# Experimental Results

- Ablation Study
  - Key selection

| | SS ($s$) | ES ($e$) | EC ($\epsilon$) | key size | mIoU (%) | TC (%) | fps (frame/s) |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 5 | 1 | 527 | 77.2 | 73.0 | 11.5 |
| 2 | 2 | 5 | 1 | 639 | 77.3 | 72.9 | 11.1 |
| 3 | 3 | 5 | 1 | 735 | 77.1 | 73.0 | 10.7 |
| 4 | 1 | 3 | 1 | 279 | 76.5 | 72.1 | 11.9 |
| 5 | 1 | 7 | 1 | 679 | 77.3 | 72.8 | 11.0 |
| 6 | 1 | 5 | 2 | 663 | 77.1 | 72.8 | 11.0 |
| 7 | 1 | 5 | 3 | 695 | 77.2 | 72.9 | 11.0 |
| 8 | - | - | - | 57344 | 75.1 | 69.9 | 0.2 |

  - Numbers of key frames



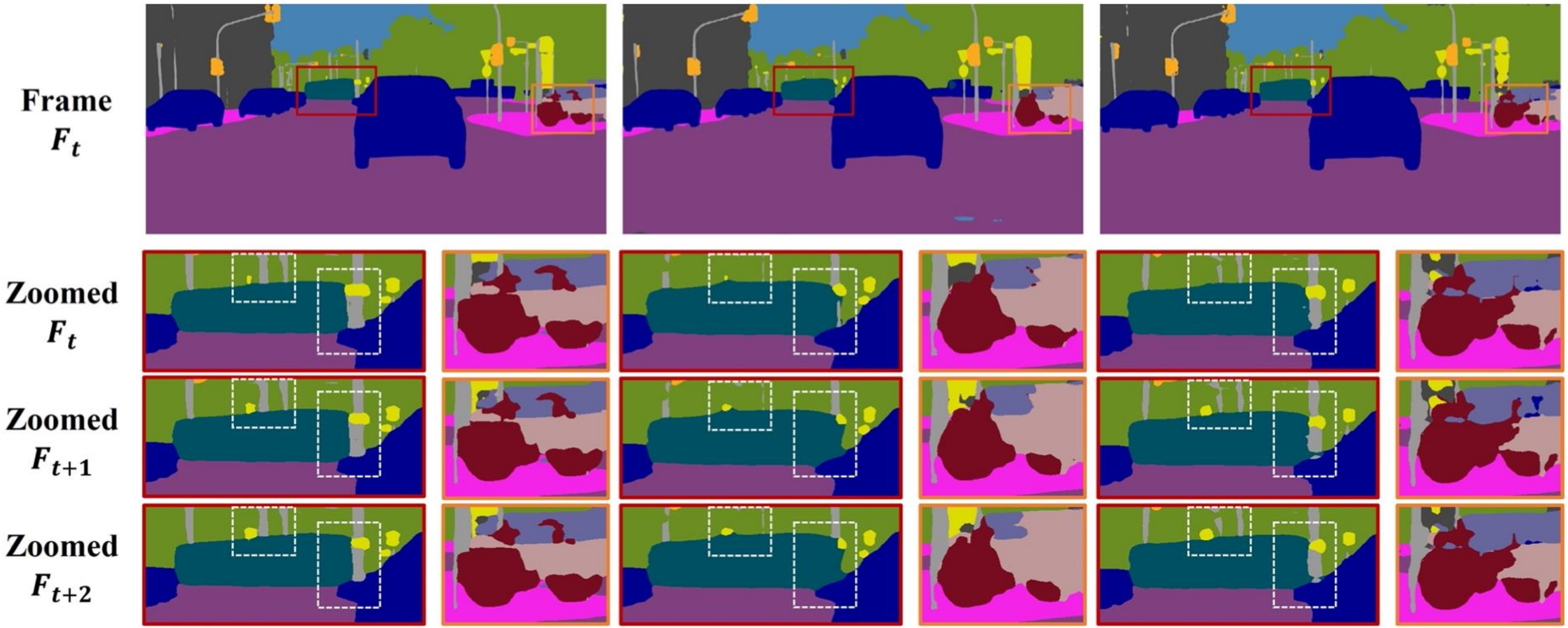




- Query selection

| | NR ($r$) | TR | mIoU (%) | TC (%) | fps (frame/s) |
|---|---|---|---|---|---|
| 1 | 1 | 50 % | 76.1 | 71.2 | 11.5 |
| 2 | 3 | 50 % | 77.1 | 72.8 | 11.5 |
| 3 | 5 | 50 % | 77.3 | 73.0 | 11.5 |
| 4 | 7 | 50 % | 77.2 | 72.9 | 11.5 |
| 5 | 9 | 50 % | 76.8 | 72.1 | 11.5 |
| 6 | 5 | 0 % | 75.3 | 68.7 | 13.6 |
| 7 | 5 | 25 % | 76.7 | 72.4 | 12.6 |
| 8 | 5 | 75 % | 77.3 | 72.7 | 10.5 |
| 9 | 5 | 100 % | 77.2 | 72.9 | 9.4 |

# Experimental Results

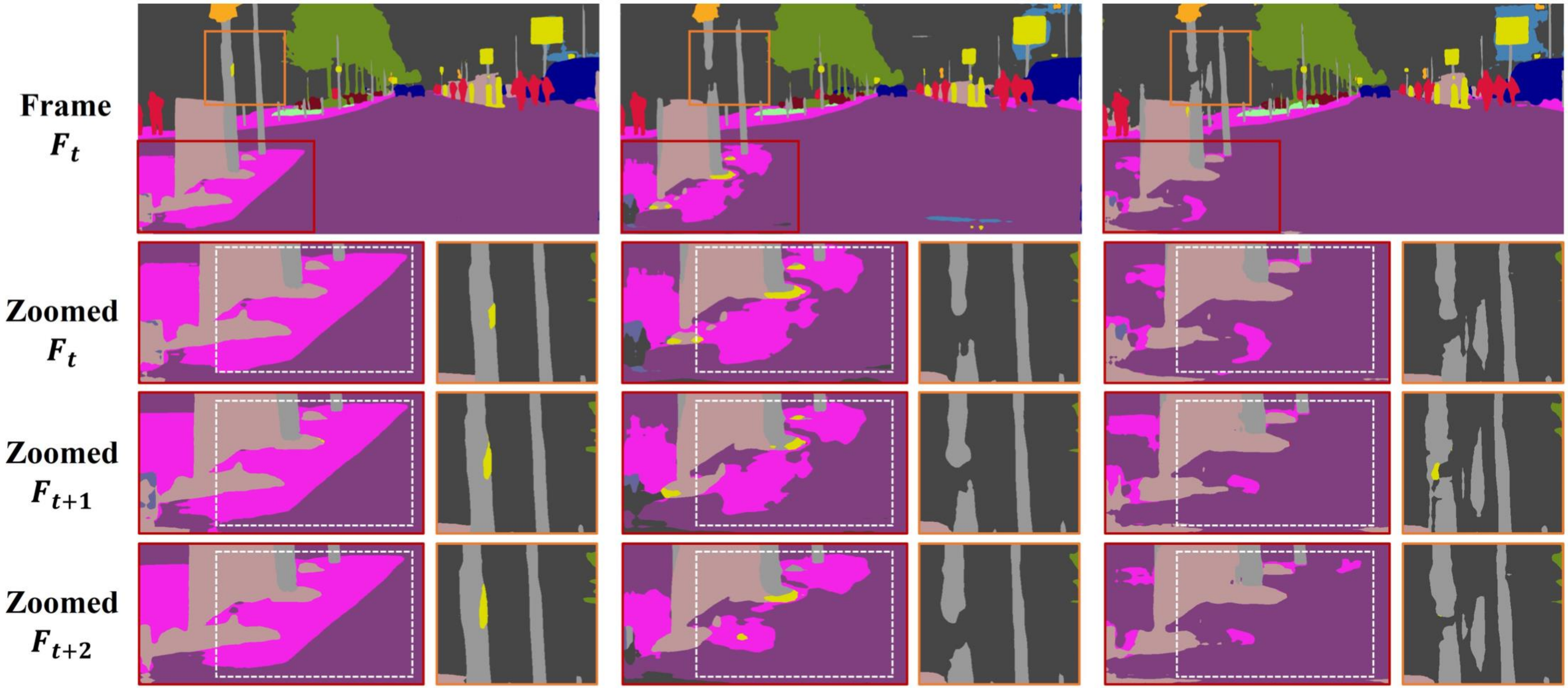
- Case study




(a) Our method (b) TDNet (c) ETC

# Experimental Results

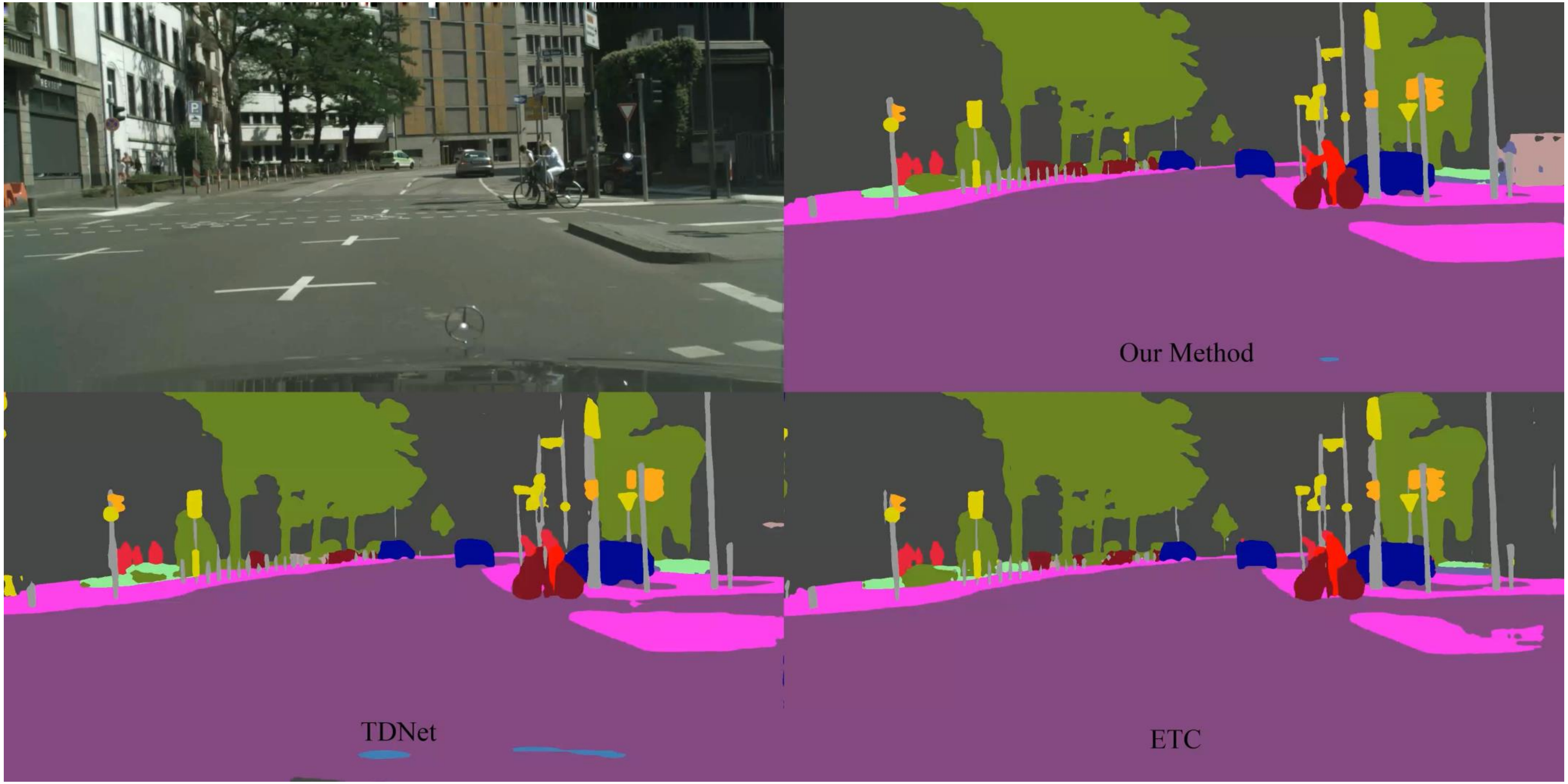
- Case study




(a) Our method (b) TDNet (c) ETC

# Experimental Results

- Case study



# Thanks for watching!

# Video Semantic Segmentation via Sparse Temporal Transformer

Jiangtong Li, Wentao Wang, Junjie Chen, Li Niu, Jianlou Si, Chen Qian, Liqing Zhang

**MoE Key Lab of Artificial Intelligence,**
**Department of Computer Science and Engineering,**
**Shanghai Jiao Tong University**

**SenseTime Research,**
**SenseTime**