Paper ID 1394

Video Semantic Segmentation via Sparse Temporal Transformer

Jiangtong Li[1*], Wentao Wang[1*], Junjie Chen[1], Li Niu[1†], Jianlou Si[2], Chen Qi[2], Liqing Zhang[1†]
[1]Department of Computer Science and Engineering, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[2]SenseTime Research

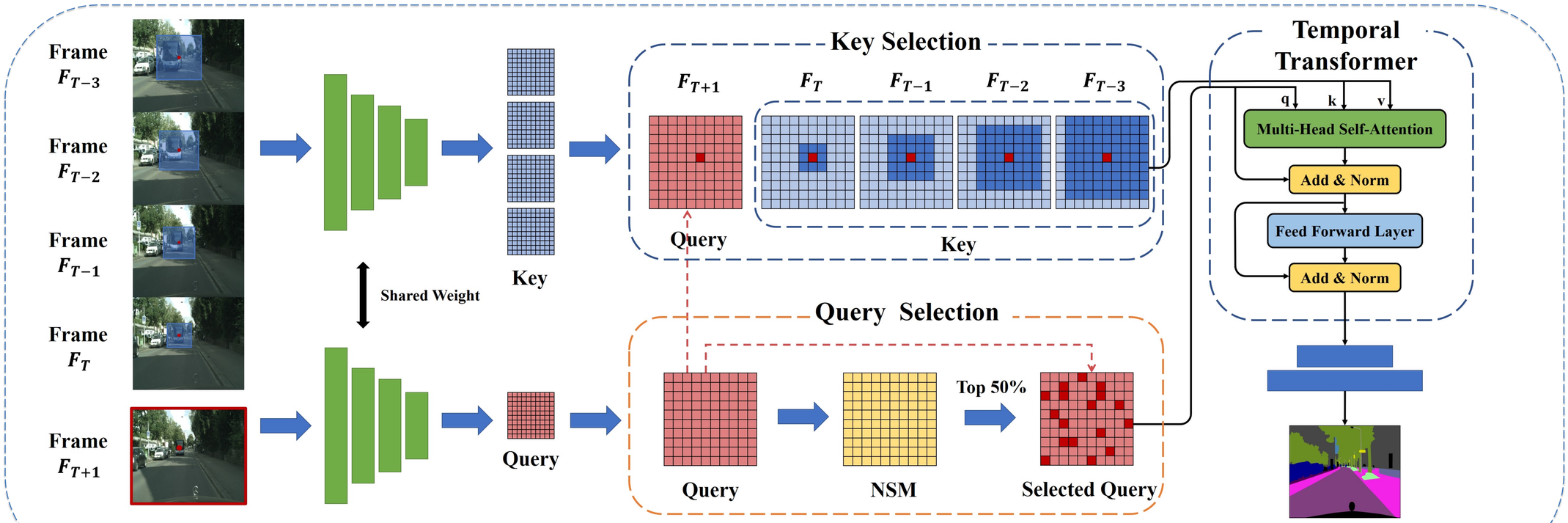

Figure 1. The flowchart of our Sparse Temporal Transformer method.

• Introduction:

➢ **Challenges in Video Semantic Segmentation:**

- The demand of temporal consistency.
- The balance between segmentation accuracy and inference efficiency.

➢ **Our Contributions:**

- We incorporate the transformer architecture into the image semantic segmentation model to capture temporal relation for video semantic segmentation task.
- We propose a novel Sparse Temporal Transformer (STT) module with key selection and query selection to balance the segmentation accuracy and inference efficiency in a good manner. Our proposed selection strategies can reduce the time complexity by a large margin without harming the segmentation accuracy and temporal consistency.
- Extensive experiments on two video semantic segmentation datasets, i.e. Cityscapes and Camvid, demonstrate the effectiveness of our method.

• Method:

➢ **Sparse Temporal Transformer Network:**

• The flowchart of our Sparse Temporal Transformer method are shown in Figure 1.

➢ **Query Selection:**

• Remove the redundancy points in query feature maps.
• Select rule -- Neighboring Similarity Matrix (NSM):

$$\boldsymbol{p}_{sim} = SoftMax(Q^n \cdot q^T) \qquad \boldsymbol{p}_u: \text{Uniform distribution}$$

$$\mathfrak{D}_{KL} = KL(\boldsymbol{p}_u || \boldsymbol{p}_{sim}) \qquad \mathfrak{D}_{cos} = \frac{1}{n_b}\sum_{i=1}^{n_b}(1 - \frac{Q^n_{[i]} \cdot q^T}{||Q^n_{[i]}||_2 ||q||_2})$$

$$\mathfrak{D}_{NSM} = \mathfrak{D}_{KL} + \mathfrak{D}_{cos}$$

➢ **Key Selection:**

• Remove the redundancy points in key feature maps.
• Select Rule: enlarge the searching regions gradually from near frame to far frame.

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

➢ **Temporal Transformer:**

• Multi-head Attention:

$$\boldsymbol{H}_j = SoftMax(\frac{(\widetilde{\boldsymbol{Q}}\boldsymbol{W}_j^Q)(\widetilde{\boldsymbol{K}}\boldsymbol{W}_j^K)^T}{\sqrt{\hat{d}}}\widetilde{\boldsymbol{K}}\boldsymbol{W}_j^V)$$

$$MH(\widetilde{\boldsymbol{Q}}, \widetilde{\boldsymbol{K}}) = [\boldsymbol{H}_i, \dots, \boldsymbol{H}_j, \dots, \boldsymbol{H}_{n_h}]\boldsymbol{W}^O$$

• Temporal transformer encoder:

$$\boldsymbol{X} = LN(\widetilde{\boldsymbol{Q}} + MH(\widetilde{\boldsymbol{Q}}, \widetilde{\boldsymbol{K}}))$$

$$FFH(\boldsymbol{X}) = \max(0, \boldsymbol{X}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2$$

$$TFE(\widetilde{\boldsymbol{Q}}, \widetilde{\boldsymbol{K}}) = LN(\boldsymbol{X} + FFH(\boldsymbol{X}))$$

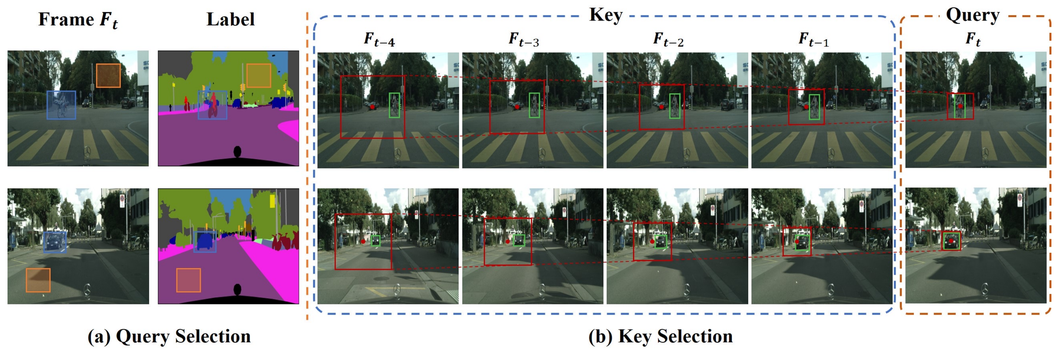

Figure 2. Illustration of query selection and key selection.

Experiments:

➢ **Qualitative Results:**

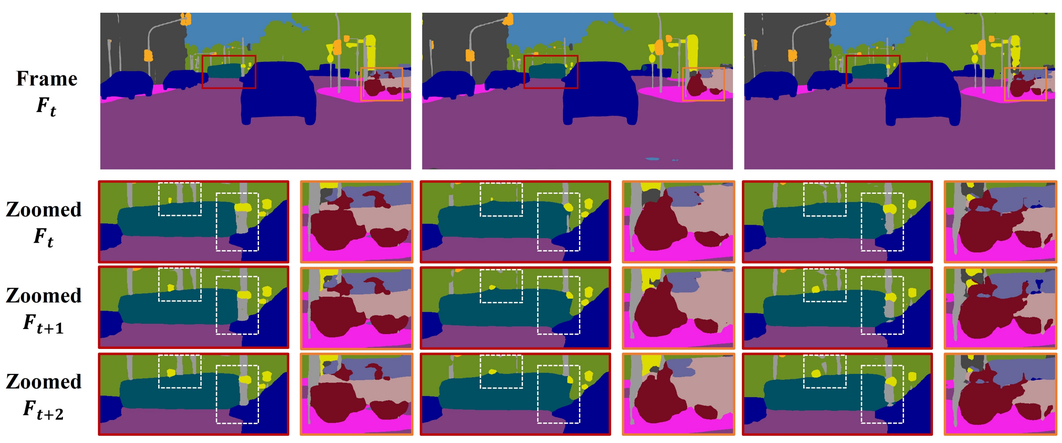

Figure 3. The qualitative comparison with two baseline methods.

➢ **Quantitative Results:**

• Comparison with 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

• Comparison with 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