Simultaneously Short- and Long-Term Temporal Modeling for Semi-Supervised Video Semantic Segmentation

Jiangwei Lao, Weixiang Hong, Xin Guo, Yingying Zhang, Jian Wang, Jingdong Chen, Wei Chu
Ant Group
wenshuo.ljw@antgroup.com

1. Details of Spatial-Temporal Transformer

The details of 3D W-MSA and Mix-FFN are presented below:

- 3D Windows Multi-head Self-Attention is an efficient implementation for 3D input. In detail, we evenly partition the 3D input feature map into a set of non-overlapping windows. For example, if the input size is $T \times H \times W$ and the window size is $t \times h \times w$, we first partition the input into $\lceil \frac{T}{t} \rceil \times \lceil \frac{H}{h} \rceil \times \lceil \frac{W}{w} \rceil$ windows, and then perform multi-head self-attention within each window. Finally, we merge the features into a new 3D tensor that has the same shape as the input tensor.

- Mix-FFN like [15] enables more efficient computation than Video Swin Transformer [10] for information exchange between local windows. Specifically, Mix-FFN adds a depth-wise 3×3 convolution between the two MLPs as: MLP → DW-Conv → MLP. The introduction of depth-wise 3×3 convolution helps to connect non-overlapping windows [17].

2. Experiment

2.1. Details about Global Category Context

We experiment different optimization strategies (*i.e.*, MSE and Exponential Moving Averages) for GCC. The experimental results are illustrated in Figure 1, as we can see, MSE loss consistently outperforms EMA. Also, GCC shows the best performance when the number of clusters is 3 with either MSE or EMA. Therefore, the number of clusters of GCC is set to 3 throughout this paper.

2.2. Transformer-based Backbone

In order to verify the generalization of our proposed method for different types of backbones, we additionally used the Swin Transformer [9] as our backbone. As shown in Table 1, Our proposed SSLTM still demonstrates advantageous performances than all compared methods.

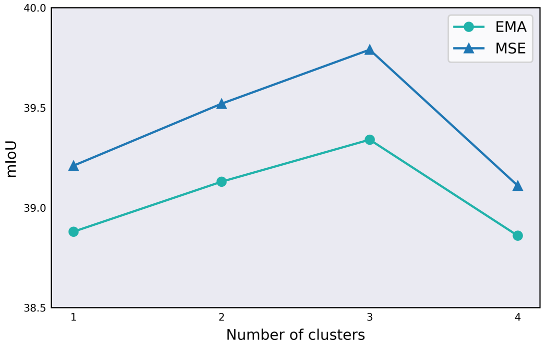


Figure 1. GCC performs the best when the number of clusters is 3. MSE loss is more suitable to update GCC, compared with EMA.

Table 1. All results are obtained with Swin-S [9] as backbone, overall higher than those of ResNet-101 [7] in our manuscript.

Image-Based	Params	mIoU	Video-Based	Params	mIoU
DeepLabv3+ [4]	59.5M	37.48	ETC [8]	55.5M	40.82
UperNet [14]	58.1M	39.49	NetWarp [5]	55.5M	40.22
PSPNet [18]	64.1M	38.18	TCB [11]	55.5M	41.42
OCRNet [16]	55.5M	39.56	SSLTM (w/o MT)	59.3M	43.11
Segmenter [12]	57.8M	39.94	SSLTM (w/ MT)	59.3M	**44.37**

2.3. Visual comparisons between w/ STT and w/o STT

The visualized results are shown in Figure 2. One can observe that the segmentation results without STT look correct in principle yet messy in details. Meanwhile, it is common to spot inconsistencies between predictions of adjacent frames, if STT is not used. As a comparison, the segmentation results with STT are smooth and consistent.

2.4. Results on the CityScapes test set

We compare with the current sota method on the cityscapes test set. The experimental results are shown in Table 2. Our proposed SSLTM demonstrates advantageous performances over all competitors.

Table 2. Comparisons on the CityScapes test set. Our proposed SSLTM demonstrates advantageous performances over all competitors. C: Cityscapes coarse annotation. V: Cityscapes video. MV: Mapillary Vistas.

Model	Extra Data	Training Method	mIoU
Zhu *et al.* [19]	V, MV	semi-supervised	83.5
Panoptic-DeepLab [3]	MV	supervised	84.1
OCR [16]	C, MV	supervised	84.2
Panoptic-DeepLab [3]	C, V, MV	semi-supervised	85.1
Tao *et al.* [13]	C, MV	semi-supervised	85.1
Naive-Student [2]	V, MV	semi-supervised	85.2
Warp-Refine Propagation [6]	C, V, MV	semi-supervised	85.3
Borse *et al.* [1]	C, MV	semi-supervised	85.6
Ours	V, MV	semi-supervised	86.4

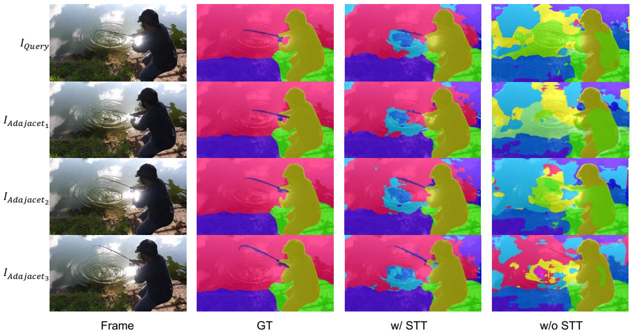


Figure 2. Visual comparisons between w/ STT and w/o STT on the validation set of VSPW. STT is beneficial to the consistency of semantic prediction.

2.5. Analysis of Mean Teacher

We report the comparison of Mean Teacher semi-supervised training method with iterative semi-supervised [2] training method in Figure 3. It can be observed that Mean Teacher can significantly save disk I/O and training time, while also achieving better segmentation performance. It is worth noting that the iterative semi-supervised training consists of three sequential steps: training of the teacher model, label generating, training of the student model. In the label generating step, additional disk I/O is compulsory to store pseudo labels. In contrast, Mean Teacher training method only requires one training procedure, without consuming additional disk access.

2.6. Reference Frame Selection and Model Robustness

As shown in Figure 4a, the mIoU increases with enlarging temporal distance, till saturates at around 60 frames. It validates our ad-hoc choice to some extent, as the average frame number per video in the VSPW dataset is 71. As for model robustness, we evaluate performance on two subsets from the original VSPW validation set, *i.e.*, "Noisy Data", where reference frame contains different categories from the query frame, and the rest "Clean Data". As shown in Figure 4b, the RFCE performs consistently better than the baseline (QFCE), and adding GCC further improves the performance in both cases. Notably, RFCE still works well even when the reference frame contains noise.

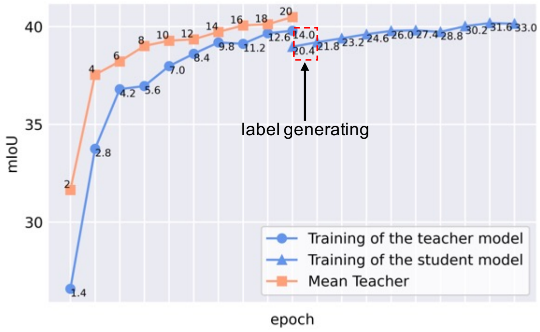
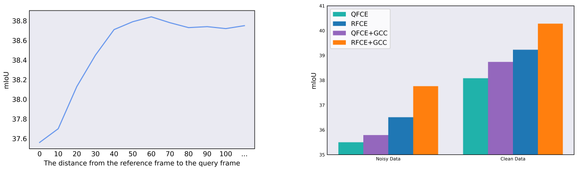


Figure 3. Comparison of Mean Teacher and Iterative Semi-Supervised Learning. The texts next to the dots represent the required time (hours), Mean Teacher takes much less time to reach better performance than iterative semi-supervised learning.



(a)

(b)

Figure 4. (a) Reference frame selection in RFCE. (b) Performance with noisy and clean reference frames.

2.7. Failure Modes

As shown in Figure 5, when the temporal distance between the first frame and the last frame of the video is very long, the video scene often changes a lot, and the temporal correlation between the query frame and the reference frame may break. Although we greatly reduce this effect using GCC modules, VSS is still a challenge in long video scenarios.

References

[1] Shubhankar Borse, Ying Wang, Yizhe Zhang, and Fatih Porikli. Inverseform: A loss function for structured boundary-aware segmentation. In *CVPR*, 2021. 2

[2] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging

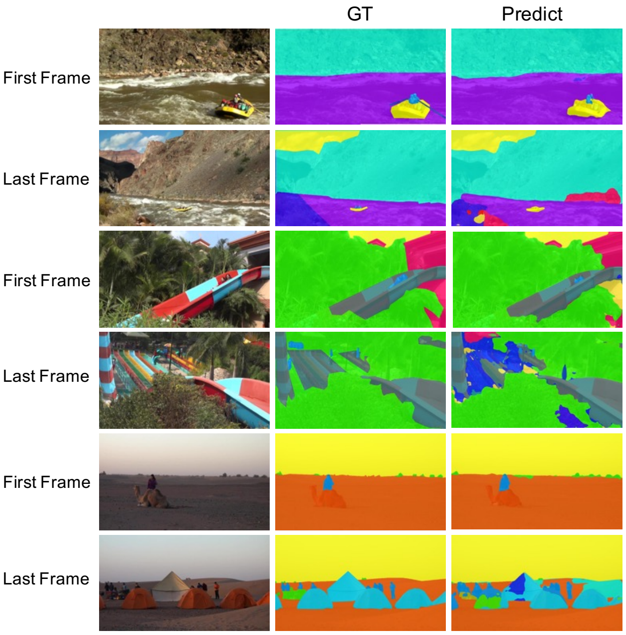


Figure 5. Some failure cases in long video scenarios.

semi-supervised learning in video sequences for urban scene segmentation. In *ECCV*, 2020. 2

[3] Liang-Chieh Chen, Huiyu Wang, and Siyuan Qiao. Scaling wide residual networks for panoptic segmentation. *CoRR*, abs/2011.11675, 2020.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1

[5] Raghudeep Gadde, Varun Jampani, and Peter V. Gehler. Semantic video cnns through representation warping. In *ICCV*, 2017. 1

[6] Aditya Ganeshan, Alexis Vallet, Yasunori Kudo, Shin-ichi Maeda, Tommi Kerola, Rares Ambrus, Dennis Park, and Adrien Gaidon. Warp-refine propagation: Semi-supervised auto-labeling via cycle-consistency. In *ICCV*, 2021. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[8] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *ECCV*, 2020. 1

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[10] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *CoRR*, abs/2106.13230, 2021. 1

[11] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. VSPW: A large-scale dataset for video scene parsing in the wild. In *CVPR*, 2021. 1

[12] Robin Strudel, Ricardo Garcia Pinel, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 1

[13] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *CoRR*, abs/2005.10821, 2020. 2

[14] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 1

[15] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 1

[16] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 1, 2

[17] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. *CoRR*, abs/2110.09408, 2021. 1

[18] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1

[19] Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn D. Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *CVPR*, 2019. 2