

# TEMPORAL MEMORY ATTENTION FOR VIDEO SEMANTIC SEGMENTATION

Hao Wang<sup>1,2</sup>, Weining Wang<sup>1,2</sup>, Jing Liu<sup>1,2</sup>

<sup>1</sup> National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

<sup>2</sup> School of Artificial Intelligence, University of Chinese Academy of Sciences

## ABSTRACT

Video semantic segmentation requires to utilize the complex temporal relations between frames of the video sequence. Previous works usually exploit accurate optical flow to leverage the temporal relations, which suffer much from heavy computational cost. In this paper, we propose a Temporal Memory Attention Network (TMANet) to adaptively integrate the long-range temporal relations over the video sequence based on the self-attention mechanism without exhaustive optical flow prediction. Specially, we construct a memory using several past frames to store the temporal information of the current frame. We then propose a temporal memory attention module to capture the relation between the current frame and the memory to enhance the representation of the current frame. Our method achieves new state-of-the-art performances on two challenging video semantic segmentation datasets, particularly 80.3% mIoU on Cityscapes and 76.5% mIoU on CamVid with ResNet-50.

**Index Terms**— video semantic segmentation, memory, self-attention

## 1. INTRODUCTION

Image semantic segmentation is a dense prediction task that needs to predict a category label for each pixel of a given image. Video semantic segmentation is a much more challenging task, which needs to assign a category label for each pixel in each frame of a given video sequence.

Video semantic segmentation is an important task for visual understanding, which has attracted a lot of attention from the research community [1, 2, 3, 4]. The most straightforward solution for video semantic segmentation is to apply an image semantic segmentation model to each frame of the videos as image semantic segmentation does. However, video frames have strong relation with each other. Simply applying an image segmentation model on a video sequence frame by frame doesn't make full use of the temporal relation between video frames. Modeling the temporal relation of video frames will improve the performance of the video segmentation model. Previous works building the temporal relation of a video sequence can be categorized into two streams: optical-flow-based methods and non-optical-flow-based methods.

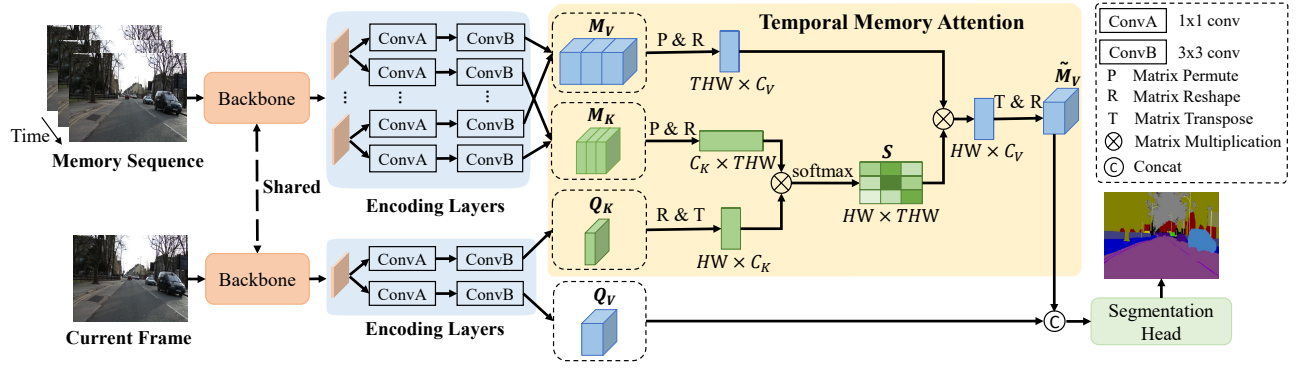


**Fig. 1.** An example of the behavior of TMANet. It collects related information from the previous frames to enhance the representation of the current frame. The orange arrows represent the highly related positions between the frames.

Optical flow represents the motion of an object between consecutive frames. The optical-flow-based methods [1, 2, 5, 3] usually contain two networks: 1) an optical flow network, which predicts the motion of objects between consecutive frames by a well pre-trained optical flow network (*e.g.* FlowNet-2.0 [6]), and 2) a segmentation network, which generates the segmentation results for the pre-defined key frame and uses the predicted optical flow to propagate the segmentation result from the key frame to other frames. Optical-flow based methods share the same point that video segmentation model needs high-quality optical flow predictions, and poor optical flow predictions will lead to poor segmentation results.

The non-optical-flow-based methods raise a new direction to generate the video representation and achieve better performance recently. Per-frame prediction method [7] on video semantic segmentation introduces a novel temporal consistency loss to improve the temporal consistency of video prediction and employs a light model with knowledge distillation to retain high performance and attain high inference speed simultaneously. TDNet [4] proposes to distribute several sub-networks over sequential frames and recombine the extracted features for segmentation via an attention propagation module. The non-optical-flow based methods discard the optical flow prediction, which is more efficient for video semantic segmentation. Our proposed method belongs to non-optical-flow-based method.

Memory networks have been introduced to enhance the reasoning ability of the model in VideoQA [8, 9] and video object segmentation [10, 11, 12], but have never been introduced in video semantic segmentation as we know. [8] uses episodic memory to conduct multiple cycles of inference by interacting the question with video features conditioned on current memory. STA [10] designs a spatial-temporal at-



**Fig. 2.** Illustration of our proposed TMANet. We select  $T$  frames from a given video as the memory sequence. The current frame and memory sequence are fed into a shared backbone to extract features. The encoding layers further embed the features to keys and values. The Temporal Memory Attention module captures temporal relation between  $Q_K$ ,  $M_K$  and  $M_V$ , generating an enhanced memory embedding  $\tilde{M}_V$ . The embedding of current frame  $Q_V$  is concatenated with  $\tilde{M}_V$  to generate final segmentation result through a segmentation head. Best viewed in color.

tention mechanism to capture the temporal information for video object segmentation. Memory networks utilize a memory component to store and retrieve information required by the query from the memory.

In video representation, it is straightforward to construct a memory that consists of the previous frames and a query represents the current frame. Then, we can retrieve information from the previous frames by computing the correlation between the previous frames and the current frame to enhance the representation of the current frame. Motivated by this, we propose a Temporal Memory Attention network (TMANet) to better capture the temporal relation of video frames and enhance the video representation without the help of optical-flow. Take the street scene in Fig.1 as an example, the person appearing on the current frame also appears in the previous frames, which exists high relationships between adjacent frames. Our model aims to adaptively integrate similar information from the previous frames, thus enhances the representation of the current frame and improves the segmentation results.

Our main contributions are as follows: (1) We propose a novel Temporal Memory Attention Network, which is the first work applying the memory and self-attention mechanism in video semantic segmentation. (2) We design a novel Temporal Memory Attention module to capture the temporal correlation in the video sequence efficiently. (3) The proposed method achieves new state-of-the-art performances on two challenging datasets, namely Cityscapes and CamVid.

## 2. METHODOLOGY

### 2.1. Overview

Given a video sequence that contains multiple frames where one frame is annotated with labels, we consider the previous frames without annotation labels as the memory frames and the current frame with annotation label as query frame. It

should be noted that the memory contains multiple frames, while the query contains one frame. Both the memory and the query frame are then fed into a shared backbone to extract features following previous works [4, 13, 14]. The output of the backbone is of high dimension but in low resolution. To reduce computational cost and encode different representation of the memory and the query, the extracted features from the backbone are fed into encoding layers for channel reduction and feature encoding. The key feature is learned to encode visual semantics for matching robust appearance variations, the value feature stores detailed information for producing semantic prediction, and the number of channel in the key feature is much smaller than that of the value feature. Next, the key and value feature go through our proposed Temporal Memory Attention (TMA) module to build the long-range temporal context information. Then, the value features of query is combined with the long-range temporal context information to enhance the query representation. After feature aggregation, a segmentation head is followed to output the final segmentation result for the current frame.

As illustrated in Fig.2, given a memory sequence containing  $T$  frames and a query with a single frame  $X \in R^{3 \times H \times W}$ , we concatenate the memory frames along the temporal dimension to get a new memory  $M \in R^{T \times 3 \times H \times W}$ . First, features are extracted via a shared deep backbone. Then, we feed them into different encoding layers to generate features with different semantic information,  $M_K \in R^{T \times C_K \times H \times W}$ ,  $M_V \in R^{T \times C_V \times H \times W}$  for memory and  $Q_K \in R^{C_K \times H \times W}$ ,  $Q_V \in R^{C_V \times H \times W}$  for query. After that, the key and value are input to the Temporal Memory Attention module to capture the long-range temporal relations. We add a simple feature aggregation following [14, 10] to aggregate the temporal information in memory and important information in query. Finally, We add a segmentation head implemented by 1x1 convolution to generate segmentation map ( $R^{C \times H \times W}$ ), where  $C$  is the number of classes.

## 2.2. Encoding Layer

Directly using the original output of the backbone is computationally expensive because of the high-dimensional channel. The simplest way for channel reduction is applying a 1x1 convolution on the feature maps. However, 1x1 convolution is not able to capture the spatial information and leads to performance decreasing. The 3x3 convolution or larger kernel can capture spatial information with a larger receptive field, but it will bring more parameters and computational cost. Therefore, we propose to apply a 1x1 convolution for channel reduction and add a 3x3 convolution for spatial information encoding to balance the performance and computation.

## 2.3. Temporal Memory Attention Module

As for images, long-range context refers to the relation between a unique pixel and other pixels [14], while the long-range context of videos is the relation between different frames [15, 10]. As represented in Fig.2, we propose a Temporal Memory Attention module to build the temporal relations of video frames for video semantic segmentation.

After embedding the memory sequence as mentioned above, we accordingly obtain  $T$  key features and  $T$  value features. We then concatenate them along the temporal dimension generating a 4-dimension matrix, and then permute and reshape them to  $M_K \in R^{C_K \times M}$  and  $M_V \in R^{M \times C_V}$ , respectively.  $M = T \times H \times W$  is the number of pixels in the memory. Similarly, we reshape and transpose the key of query to  $Q_K \in R^{N \times C_K}$ , where  $N = H \times W$  is the number of pixels in the query. Next, we multiply  $M_K$  and  $Q_K$ , and then apply a softmax layer to calculate the temporal memory attention  $S \in R^{N \times M}$ ,

$$S_{ij} = \frac{\exp(Q_K^i \cdot M_K^j)}{\sum_{j=1}^M \exp(Q_K^i \cdot M_K^j)} \quad (1)$$

where  $S_{ij}$  measures the impact of the  $i^{th}$  position in the key of query on the  $j^{th}$  position in the key of memory. It should be noted that larger impact from the query to the memory indicates greater relation between them. After obtaining the temporal attention map  $S$ , we multiply  $S$  and  $M_V$  to integrate the temporal relation to memory, thus enhancing the embedding of memory.

## 2.4. Feature Aggregation

After obtaining the long-range temporal context information via temporal memory attention module, we combine the long-range temporal context information with the information from current frame, as follows:

$$f = \Theta(\tilde{M}_V, Q_V) \quad (2)$$

where  $f$  is the aggregated feature, and  $\Theta$  is the employed feature aggregation method.

Feature aggregation can be implemented by a decoder structure [16], feature concatenation or feature summation.

In this paper, we employ feature concatenation for simplicity. After feature aggregation, we exploit a segmentation head to generate the final segmentation result for the current frame.

## 3. EXPERIMENTS

### 3.1. Dataset and Implementation Details

To evaluate our proposed method, we carry out comprehensive experiments on two benchmark datasets Cityscapes[17] and CamVid[18].

Cityscapes [17] contains 5000 high-quality fine annotated images, which can be split into 2975, 500 and 1275 snippets for training, validation and testing, respectively. Each snippet contains 30 frames, and only the  $20^{th}$  frame of each snippet is annotated with 19 classes for semantic segmentation. CamVid [18] contains 4 videos with 11 category labels for semantic segmentation and is annotated every 30 frames. The annotated frames are grouped into 467, 100 and 233 snippets for training, validation and testing, respectively. We adopt mean Intersection-over-Union (mIoU) as our evaluation metric on Cityscapes and CamVid.

We implement our method based on PyTorch on 4 GPUs of Tesla V100. Inspired by [19], we employ the poly learning rate policy and employ SGD as the optimizer, where the initial learning rate is multiplied by  $(1 - \frac{iter}{total\_iter})^{0.9}$  for each iteration. Momentum and weight decay are set to 0.9 and 5e-4 for all experiments on Cityscapes and CamVid. We train our model with Sync-BN [20], where batch size and learning rate are set to 8 and 0.01 for both datasets, respectively. We set the total iteration to 80,000 for all experiments. For data augmentation, we apply random resize with a ratio between 0.5 and 2, random cropping (768x768, 640x640 for Cityscapes and CamVid respectively) and random horizontal flipping for input images and sequences for all experiments. We apply sliding window strategy to generate video snippets in the testing stage. Following [13], we add the auxiliary segmentation loss at the low-level feature of the backbone (e.g. the stage 3 output of ResNet). We adopt the above settings for all experiments if without specific clarification.

### 3.2. Ablation Study

All the ablation experiments are conducted on the Cityscapes dataset. We use FCN-50[22] as our baseline. To save computational resources and training time, we adopt ResNet-50 as the backbone and set output stride to 16 for all ablation experiments.

Following [10], we set the channel of value features in both memory and query as four times than that of key features (e.g. when the channel of key features is set to 64, the channel of value features is set to 256). Besides, it is important to determine how many past frames should be selected into memory. We conduct experiments to analyze different numbers of channel and different memory lengths. As shown in Table 1,

**Table 1.** Comparison results of different channel numbers of key features and memory lengths on the Cityscapes validation set. Sequence2 denotes 2 frames in the Memory. Key256 denotes the channel number of key is 256.

Method	mIoU (%)
Baseline	70.69
Sequence2-Key256	77.77
Sequence2-Key128	77.95
Sequence2-Key64	77.87
Sequence2-Key32	77.65
Sequence1-Key64	78.08
Sequence4-Key64	78.26
Sequence6-Key64	78.28

**Table 2.** Comparison results of different feature aggregation methods and sampling methods (default is random) on the Cityscapes validation set.

Method	Sample	mIoU (%)
Sequence4-Key64, concat	random	78.39
Sequence4-Key64, concat	continuous	78.45
Sequence4-Key64, sum	random	78.18
Sequence4-Key64, sum	continuous	78.33

we can observe that a significant improvement from 77.52 to 78.28 is obtained when the length of memory increases from 2 to 4. While when the length of memory increases to 6, the improvement is too slight to be ignored. It can be interpreted as information redundancy that the memory storing the information of 4 frames is enough for the feature representation enhancement of the current frame. Though the model performs best when the channel number is set to 128, we choose 64 as the number of channel in key features for computational efficiency which has similar performance as 128 channels.

To build the memory, we need to select multiple frames from the past video sequence. There exist two selection methods as follows: 1) random selecting multiple frames from the past video sequence (random selecting  $n$  frames from last 10 frames), 2) continuously selecting multiple frames from the past video sequence. As shown in Table 2, the continuous selecting strategy performs better than random selecting strategy. The possible reason is that random selecting strategy may involve some long-range relation which is harmful to the current frame representation because of the long distance from the current frame. While the continuous selecting strategy selects multiple frames from the current frame continuously, the representation between frames is highly related, which will enhance the feature representation of the current frame. We also analyze different feature aggregation methods, e.g. concatenation and summation. As shown in Table 2, feature concatenation performs better than feature summation. The main reason appears to be that concatenated features involve more channels and can represent more information.

The encoding layer plays an important role in the frame-

**Table 3.** Comparison results of different encoding layers on the Cityscapes validation set.

Method	mIoU (%)
Sequence4-Key64, 3x3 conv	78.26
Sequence4-Key64, 1x1 conv	77.88
Sequence4-Key64, 1x1 conv, 3x3 conv	78.39

**Table 4.** Comparison results with state-of-the-arts on Cityscapes and CamVid validation set.

Method	mIoU (%)		GFLOPs
	Cityscapes	CamVid	
DFF [2]	69.2	-	>919
GRFP [5]	73.6	66.1	-
Netwarp [21]	-	67.1	>919
LVS [3]	76.8	-	-
TDNet-50 [4]	79.9	76.0	1082
<b>Ours-50</b>	<b>80.3</b>	<b>76.5</b>	<b>754</b>

work, thus we also compare different encoding layers and the results are listed in Table 3. It can be seen that a combination of 1x1 convolution and 3x3 convolution performs best than other configurations.

### 3.3. Comparison with State-of-the-arts

Table 4 shows the performance and GFLOPs of our method and other state-of-the-art methods. Considering the inference time varies from different hardware environments, we provide the computational cost of GFLOPs for fair comparison. Compared with other optical-flow based methods and non optical-flow based methods, our method achieves better performance on both Cityscapes and Camvid datasets with lower computational cost.

## 4. CONCLUSIONS

In this paper, we propose a Temporal Memory Attention Network (TMANet) for video semantic segmentation, which is the first work using memory and self-attention to build the temporal relation in video semantic segmentation. Specially, we introduce a Temporal Memory Attention module to capture the temporal relations between frames. Our method achieves state-of-the-art performance on Cityscapes and CamVid dataset without complicated testing augmented skills. In the future, we will continue to decrease the computation complexity and enhance the efficiency of the model.

## 5. ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (61922086, 61872366) and Beijing Natural Science Foundation (4192059, JQ20022).

## 6. REFERENCES

- [1] Evan Shelhamer, Kate Rakelly, Judy Hoffman, and Trevor Darrell, “Clockwork convnets for video semantic segmentation,” in *ECCV*, 2016, pp. 852–868.
- [2] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei, “Deep feature flow for video recognition,” in *CVPR*, 2017, pp. 2349–2358.
- [3] Yule Li, Jianping Shi, and Dahua Lin, “Low-latency video semantic segmentation,” in *CVPR*, 2018, pp. 5997–6005.
- [4] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi, “Temporally distributed networks for fast video semantic segmentation,” in *CVPR*, 2020, pp. 8818–8827.
- [5] David Nilsson and Cristian Sminchisescu, “Semantic video segmentation by gated recurrent flow propagation,” in *CVPR*, 2018, pp. 6819–6828.
- [6] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, “FlowNet 2.0: Evolution of optical flow estimation with deep networks,” in *CVPR*, 2017, pp. 2462–2470.
- [7] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang, “Efficient semantic video segmentation with per-frame inference,” *arXiv preprint arXiv:2002.11433*, 2020.
- [8] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia, “Motion-appearance co-memory networks for video question answering,” in *CVPR*, 2018, pp. 6576–6585.
- [9] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang, “Heterogeneous memory enhanced multimodal attention model for video question answering,” in *CVPR*, 2019, pp. 1999–2007.
- [10] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim, “Video object segmentation using space-time memory networks,” in *ICCV*, 2019, pp. 9226–9235.
- [11] Yu Li, Zhuoran Shen, and Ying Shan, “Fast video object segmentation using the global context module,” *arXiv preprint arXiv:2001.11243*, 2020.
- [12] Zhishan Zhou, Lejian Ren, Pengfei Xiong, Yifei Ji, Peisen Wang, Haoqiang Fan, and Si Liu, “Enhanced memory network for video segmentation,” in *ICCV*, 2019, pp. 0–0.
- [13] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, “Pyramid scene parsing network,” in *CVPR*, 2017, pp. 2881–2890.
- [14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, “Dual attention network for scene segmentation,” in *CVPR*, 2019, pp. 3146–3154.
- [15] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, “Non-local neural networks,” in *CVPR*, 2018, pp. 7794–7803.
- [16] 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, pp. 801–818.
- [17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *CVPR*, 2016, pp. 3213–3223.
- [18] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla, “Segmentation and recognition using structure from motion point clouds,” in *ECCV*, 2008, pp. 44–57.
- [19] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *TPAMI*, vol. 40, no. 4, pp. 834–848, 2017.
- [20] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Amrith Tyagi, and Amit Agrawal, “Context encoding for semantic segmentation,” in *CVPR*, 2018, pp. 7151–7160.
- [21] Raghudeep Gadde, Varun Jampani, and Peter V Gehler, “Semantic video cnns through representation warping,” in *ICCV*, 2017, pp. 4453–4462.
- [22] Jonathan Long, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation,” in *CVPR*, 2015, pp. 3431–3440.