Supplementary File for Memory Selection Network for Video Propagation

Ruizheng Wu[1⋆], Huaijia Lin[1⋆], Xiaojuan Qi[2], Jiaya Jia[1,3]
[1]The Chinese University of Hong Kong, [2]University of Hong Kong,
[3]SmartMore
{rzwu, linhj, leojia}@cse.cuhk.edu.hk, xjqi@eee.hku.hk

1 Multiple Frame Propagation

In our main paper, we consider only one-frame propagation among all propagation methods, and here we adopt MSN to select more frames for the multiple-frame propagation method, e.g. STM [2]. We apply MSN on the full STM, i.e. utilize every 5 previous frames as guidance ('STM-5'), and we will select the same number of frames by MSN as guidance for propagation ('STM-5 + MSN'), shown in Tab. 1.

Table 1. Multiple frame propagation on DAVIS-2017 dataset.

Methods	$\mathcal{J}$ (%)	$\mathcal{F}$ (%)
STM-5 [2]	79.2	84.3
STM-5 + MSN	79.8	84.8

2 Components of Baseline Temporal Propagation Network

As shown in Fig. 1, the baseline temporal propagation network (TPN) includes two branches: Appearance branch and Motion branch. Both branches are based on VGG16 [3]. Specifically, for the motion branch, firstly we incorporate a pre-trained FlowNet2C [1] to extract optical flow for two frames as input. Besides, we design two sub-branches for the motion branch, which involve different numbers of convolution layers. Tab. 2 demonstrates the effects of different components.

3 Memory Selection Network

Network Architecture We design a light-weight selection network that only involves several convolution layers and fully connected layers, as illustrated in

⋆ Equal Contribution.

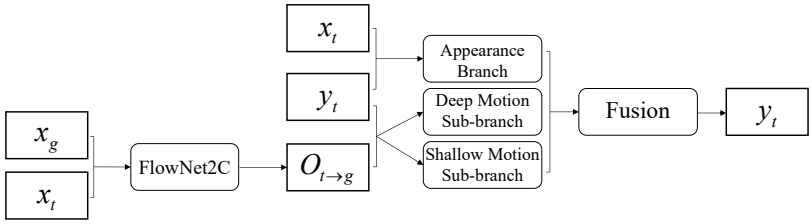

Fig. 1. Components of temporal propagation network (**TPN**).

Table 2. The effects of different components in TPN. We quantify the performance of each component in the video object segmentation task on YouTube-VOS dataset [4]. 'Only Appearance Branch' indicates only appearance branch is utilized. 'Only Motion Branch' indicates only motion branch is used. 'w/o FlowNet2C' denotes FlowNet2C [1] is fixed instead of incorporating into framework and training end-to-end. 'w/o Shallow Motion' denotes shallow motion branch is discarded.

Components	*Overall* score (%)
Only Appearance Branch	48.82
Only Motion Branch	62.02
w/o FlowNet2C [1]	50.90
w/o Shallow Motion	52.61
TPN	**63.04**

Fig. 2. In the inference stage, f_a, f_p and f_t are extracted with VGG16 [3] only once for each frame, while they might be utilized by selection network for multiple times.

Qualitative Results Fig. 3 illustrates the comparisons of using previous frames or selected frames for propagation. Frame-by-frame propagation fails to propagate the mask correctly in the occlusion (the first two rows) and large motion (the last row) scenarios. The proposed MSN selects the suitable guidance frames for propagation, where error accumulation is reduced.

4 Selection Analysis

Statistic of Guidance Frame Intervals Fig. 4 demonstrates the distribution of MSN selections. It is deduced that the adjacent frame (frame interval equals to 1) is the preferred frames to be selected by MSN in most easy scenarios, while MSN selects distant frames especially when the adjacent frame undergoes large motion and serious occlusion.

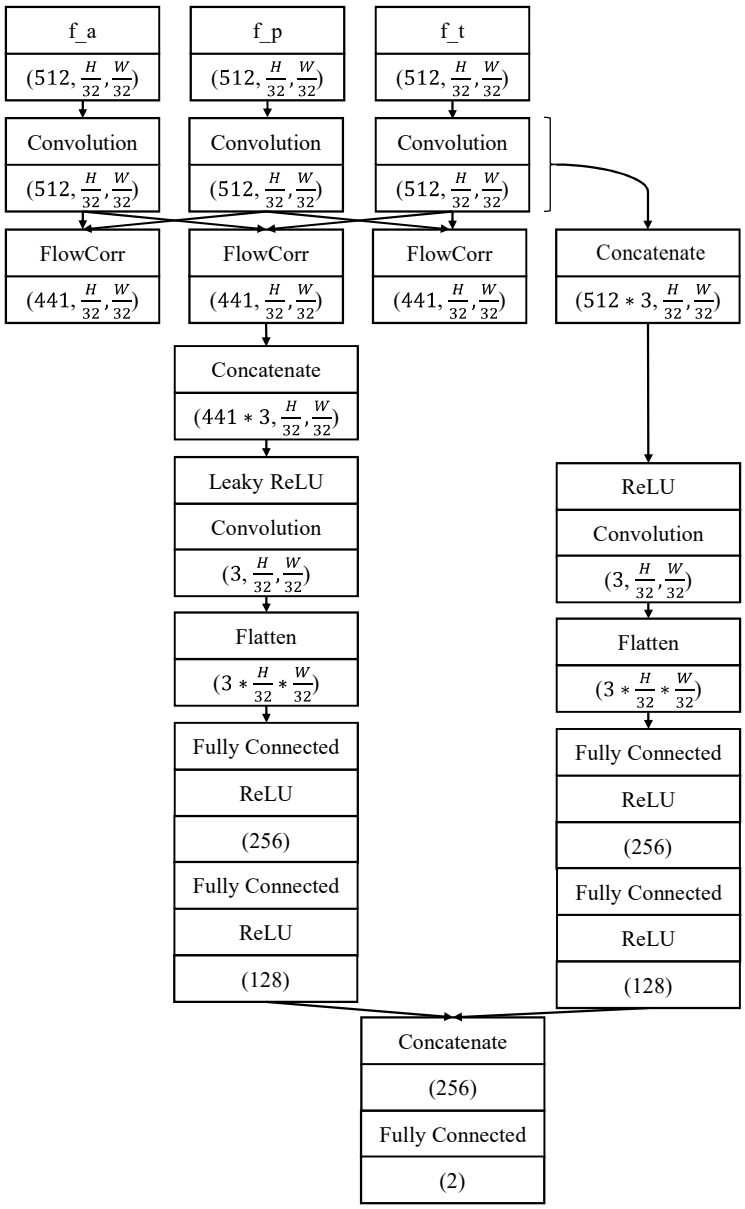

Fig. 2. The detailed network architecture of memory selection network (**MSN**). Each block is composed of the layer name and its output size, where feature maps are denoted as (channel, height, width) and vectors are denoted as (size).

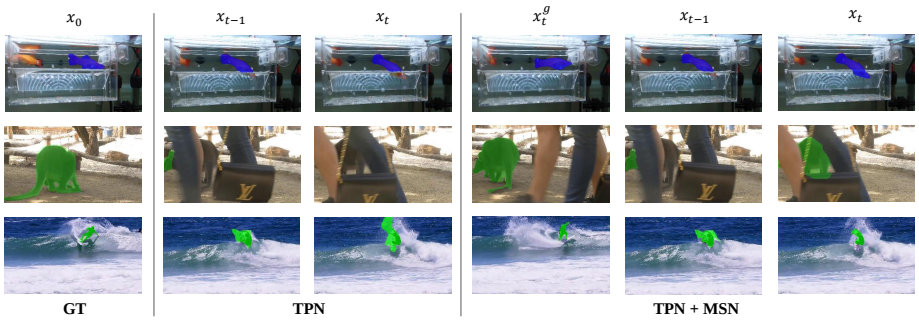

Fig. 3. Qualitative comparison between frame-by-frame propagation and MSN-guided propagation on YouTube-VOS[4]. The 1^{th} column illustrates the annotated frame. In frame-by-frame propagation, segmentation results in previous frames (the 2^{nd} column) are propagated to target frames (the 3^{rd} column) with TPN. Guided by MSN, segmentation results in selected guidance frames (the 4^nd column) instead of previous frames (the 5^{th} column) are propagated to target frames (the 6^{th} column). Best viewed in color.

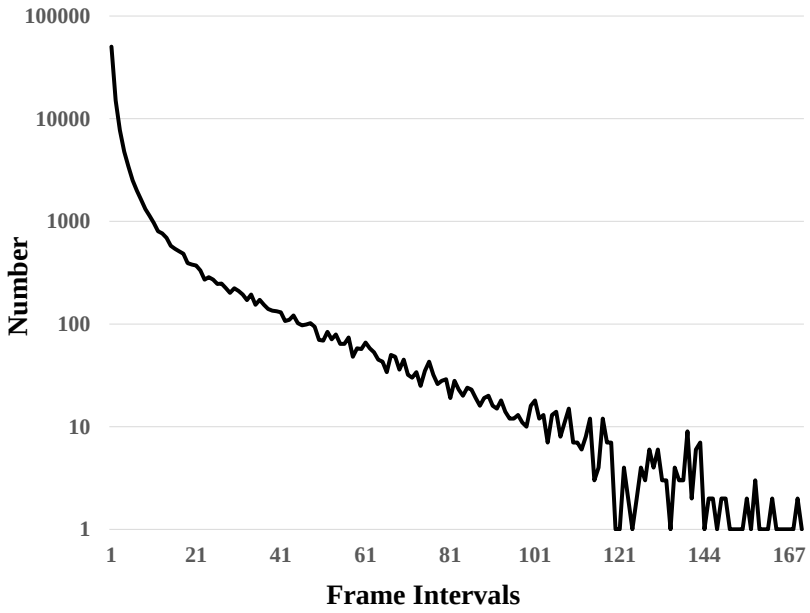

Fig. 4. Statistic of the frame intervals between the selected guidance frame and target frame on YouTube-VOS dataset [4].

References

1. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR (2017)
2. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: ICCV (2019)
3. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
4. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)