MoNet: Deep Motion Exploitation for Video Object Segmentation —Supplementary Material

Huaxin Xiao[1] Jiashi Feng[2] Guosheng Lin[3] Yu Liu[1] Maojun Zhang[1]
[1]National University of Defense Technology [2]National University of Singapore [3]Nanyang Technological University
{xiaohuaxin, jasonyuliu, mjzhang}@nudt.edu.cn elefjia@nus.edu.sg gslin@ntu.edu.sg

A. Unsupervised results

Tab. S1 present the results without online fine-tuning. The numbers in red denote the corresponding results with first frame fine-tuning. Other notations follow conventions in Tab. 5.

Variant	Baseline	+FA	+FA&MP	+CRF
Mean $\mathcal{G}\uparrow$	63.2 (75.7)	66.4 (80.2)	71.2 (83.8)	72.5 (84.7)

Table S1. The unsupervised results by different variants.

B. Results on DAVIS 2017

We evaluate our model on DAVIS-17 [5] validation set without using the DAVIS-17 training data. The mean $\mathcal{G}$ for variants +FA, +FA&MP and +FA&MP&CRF is 0.566, 0.581 and 0.588 respectively.

C. Detailed Results of Attribute

Tab. S2 shows the detailed results of attribute-based performance analysis. Four top-performing semi-supervised approaches, *i.e.*, OSVOS [1], MSK [4], SFL [2] and CTN [3], are selected for comparison. The $\mathcal{M}_{\mathcal{J}}$ in Tab. S2 denotes mean region similarity $\mathcal{J}$ over all sequences with the specific attribute (e.g., Shape Complexity), and "Gain" denotes the performance gain on video sequences without the specified challenging attribute. The proposed MoNet has better performance on the video sequences with all attributes and presents more stable performance—when discarding these attributes.

Tab. S3 summarizes the mean region similarity $\mathcal{J}$ performance of different components, *i.e.*, feature alignment (*+FA*), motion prior (*+FA&MP*) and fully-connected CRF (*+CRF*), under various video attributes. The red numbers in Tab. S3 denote the improvement over the previous component.

D. More Qualitative Results

Fig. S1 shows example segmentation results of the proposed two components, *i.e.*, feature alignment (*+FA*) and motion prior (*+FA&MP*). We can observe that the *+FA* provides more complete segmentation masks, especially in the second example, while the *+FA&MP* can effectively filter out the confusing instances and noisy regions.

References

[1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017. 1, 2

[2] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, 2017. 1, 2

[3] W.-D. Jang and C.-S. Kim. Online video object segmentation via convolutional trident network. In *CVPR*, 2017. 1, 2

[4] A. Khoreva, F. Perazzi, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, 2017. 1, 2

[5] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 1

Attribute	MoNet		OSVOS [1]		MSK [4]		SFL [2]		CTN [3]	
	$\mathcal{M}_\mathcal{J}\uparrow$	Gain $\downarrow$	$\mathcal{M}_\mathcal{J}\uparrow$	Gain $\downarrow$	$\mathcal{M}_\mathcal{J}\uparrow$	Gain $\downarrow$	$\mathcal{M}_\mathcal{J}\uparrow$	Gain	$\mathcal{M}_\mathcal{J}\uparrow$	Gain $\downarrow$
Appearance Change	**86.7**	**-3.1**	80.6	-1.2	79.8	-0.1	77.6	-2.3	72.8	+1.0
Background Clutter	**84.5**	+0.2	83.2	**-4.2**	82.9	-4.0	78.0	-2.4	71.4	+2.7
Camera-Shake	**85.3**	-0.9	78.4	+2.2	77.4	+3.6	79.7	**-5.6**	72.4	+1.7
Deformation	**83.1**	+4.4	78.6	+3.4	79.2	+1.4	76.4	**-0.8**	72.5	+2.8
Dynamic Background	**75.5**	+10.8	74.3	**+6.5**	74.1	+6.6	54.9	+24.9	62.7	+12.7
Edge Ambiguity	**81.2**	+7.7	76.7	**+6.8**	74.5	+11.6	71.6	+9.9	67.2	+14.0
Fast Motion	**82.4**	**+3.5**	76.5	+5.1	74.8	+7.6	71.9	+6.4	65.1	+12.9
Heterogeneus Object	**80.6**	**+13.5**	74.9	+16.3	75.6	+13.6	70.8	+17.6	68.4	+17.1
Interacting Objects	**80.0**	+9.4	74.5	+10.5	75.5	+8.4	71.9	**+8.3**	67.5	+12.1
Low Resolution	**83.2**	**+1.9**	77.2	+3.5	76.3	+4.6	70.6	+7.3	66.8	+9.0
Motion Blur	80.2	+8.1	73.7	+11.0	73.4	+11.4	74.1	**+3.6**	67.6	+10.7
Occlusion	**81.0**	+5.3	77.2	**+3.7**	75.5	+6.0	71.0	+7.2	70.2	+4.7
Out-of-view	**82.9**	+2.2	71.7	+10.0	71.3	+10.5	78.8	**-3.5**	67.4	+7.7
Shape Complexity	**75.3**	+14.4	70.6	**+14.2**	70.2	+14.6	66.5	+14.6	63.6	+15.3
Scale Variation	**80.5**	**+6.9**	74.3	+9.1	73.6	+10.2	69.3	+11.2	63.7	+16.3
Mean	**81.5**	+5.6	76.2	+6.5	75.6	+7.1	72.2	+6.4	68.0	9.4

Table S2. Attribute-based analysis on DAVIS validation set. We compare the proposed MoNet with 4 top-performing CNN-based methods, *i.e.*, OSVOS, MSK, SFL and CTN. For each method, we calculate the mean $\mathcal{J}$ over all sequences with specified attribute labeled, and "Gain" denotes the performance gain on video sequences without the specified challenging attribute. The up-arrow $\uparrow$ means larger is better while the down-arrow $\downarrow$ means smaller is better.

Attribute	Baseline	+FA	+FA&MP	+CRF
Appearance Change	77.4	81.5 **+4.1**	84.6 +3.1	**86.7** +2.1
Background Clutter	78.2	79.4 +1.1	83.5 **+4.2**	**84.5** +1.0
Camera-Shake	78.3	81.6 **+3.3**	83.8 +2.2	**85.3** +1.5
Deformation	73.8	77.7 **+3.9**	80.2 +2.5	**83.1** +2.9
Dynamic Background	62.6	65.7 +3.1	71.3 **+5.6**	**75.5** +4.3
Edge Ambiguity	74.4	74.7 +0.3	78.2 **+3.5**	**81.2** +3.0
Fast Motion	73.5	78.0 **+4.6**	80.4 +2.4	**82.4** +2.0
Heterogeneus Object	69.9	73.7 **+3.8**	77.3 +3.7	**80.6** +3.3
Interacting Objects	69.7	71.8 +2.1	76.1 **+4.3**	**80.0** +3.8
Low Resolution	74.5	78.5 **+4.0**	81.7 +3.2	**83.2** +1.5
Motion Blur	71.5	74.3 +2.8	76.8 +2.5	**80.2** **+3.4**
Occlusion	69.5	73.3 +3.8	76.5 +3.3	**81.0** **+4.4**
Out-of-view	71.5	79.1 **+7.6**	81.0 +1.9	**82.9** +2.0
Shape Complexity	64.6	67.6 +3.0	70.5 +2.8	**75.3** **+4.8**
Scale Variation	71.2	72.5 +1.2	77.1 **+4.6**	**80.5** +3.4
Mean	72.1	75.3+3.2	78.6**+3.3**	**81.5**+2.9

Table S3. Attribute-based component analysis on the DAVIS validation set. We compare the mean $\mathcal{J}$ performance of different components, *i.e.*, feature alignment (+*FA*), motion prior (+*FA*&*MP*) and fully-connected CRF (+*CRF*), under various video attributes.

Figure S1. Segmentation results (red masks) of the proposed feature alignment (+*FA*) and motion prior (+*FA*&*MP*). Best viewed in color with 2× zoom.