The Emergence of Objectness: Learning Zero-Shot Segmentation from Videos

Runtao Liu[1,2*] **Zhirong Wu**[1*] **Stella X. Yu**[3] **Stephen Lin**[1]

Microsoft Research Asia[1] John Hopkins University[2] UC Berkeley / ICSI[3]
runtao219@gmail.com stellayu@berkeley.edu {wuzhiron,stevelin}@microsoft.com

In this supplementary, we provide the network architecture details in Section 1. In Section 2, we present more qualitative video object segmentation results on 3 datasets, DAVIS 2016 [1], FBMS59 [2] and SegTrackv2 [3]. We also present the per class quantitative results on these datasets.

1 Network Details

The network details are shown in Table 1 and Table 2. Table 1 shows the detailed network architecture for the segment prediction head of our segmentation network. Our correspondence network adopt the similar framework as PWCNet [4] which contains a feature extractor, a flow estimator and a context network. The feature extractor is the same as that of PWCNet while we don't use the context network in our correspondence network. Table 2 shows the detailed layers of the flow estimator.

2 More Results

More video object segmentation results are shown in Figure 1, Figure 2 and Figure 3 for SegTrackv2, DAVIS 2016 and FBMS59 correspondingly. We choose those samples from different videos as much as possible. In Figure 4, more saliency detection results from DUTS [5] dataset are represented.

We also include the videos containing all our predictions in the submission .zip file. For FBMS59, the annotation is given per 20 frames and we only predict those frames since the video is too long.

3 Broader Impact

We proposed a self-supervised pretraining method for zero-shot object segmentation. The central idea of decomposing appearance and motion can be implemented with other network architectures, and even training losses. However, we have not studied the implications of these variations of the approach. There will also be unpredictable failures, where the generalization of the self-supervised framework still needs deeper understanding. This method is data-driven thus the data bias problem should be careful during data collection in both pretraining and downstream tasks. As this method can be applied to a wide range of videos without annotation, privacy should be also careful during the data utilization.

[*]Equal contribution. Work done when Runtao was a StarBridge intern at MSRA.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

References

[1] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[2] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200, 2013.

[3] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2192–2199, 2013.

[4] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[5] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017.

Table 1: Details about the prediction head in our segmentation network. Our segmentation network consists of a backbone, ResNet50, and a prediction head which predicts the segments through the features from the backbone. Here c is a hyperparameter which represents the segment number.

Layer	Output size
Input Feature	$2048 \times 48 \times 48$
Conv($3 \times 3, 2048 \rightarrow 256$) + BN + ReLU	$256 \times 48 \times 48$
Conv($3 \times 3, 256 \rightarrow 256$) + BN + ReLU	$256 \times 48 \times 48$
Conv($3 \times 3, 256 \rightarrow c$)	$c \times 48 \times 48$

Table 2: Architecture details about our correspondence network. As it processes the input at different pyramid levels, here H and W represents the size of input in a certain level. And c is a hyperparameter about the segment number.

Index	Layer	Output size
1.	Input Feature	$115 \times H \times W$
2.	Conv($3 \times 3, 115 \rightarrow 128$) + ReLU	$128 \times H \times W$
3.	Conv($3 \times 3, 128 \rightarrow 128$) + ReLU	$128 \times H \times W$
4.	Concatenate 2. and 3.	$256 \times H \times W$
5.	Conv($3 \times 3, 256 \rightarrow 96$) + ReLU	$96 \times H \times W$
6.	Concatenate 3. and 5.	$224 \times H \times W$
7.	Conv($3 \times 3, 224 \rightarrow 64$) + ReLU	$64 \times H \times W$
8.	Concatenate 5. and 7.	$160 \times H \times W$
9.	Conv($3 \times 3, 160 \rightarrow 32$) + ReLU	$32 \times H \times W$
10.	Concatenate 7. and 9.	$96 \times H \times W$
11.	Average Pooling	$96 \times c$
12.	FC ($96 \rightarrow 2$)	$2 \times c$

Figure 1: Qualitative results of SegTrackv2

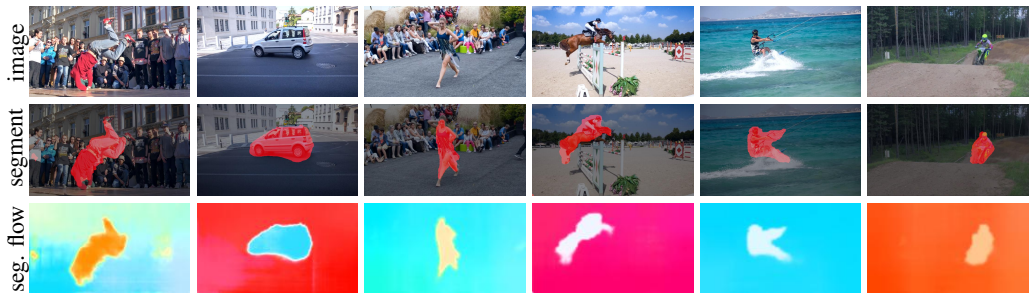

Figure 2: Qualitative results of DAVIS 2016

Figure 3: Qualitative results of FBMS59

Figure 4: Qualitative salient object detection results. Our model can detect multiple primary objects and even static object like the chair and the rocks.