Unsupervised Video Object Segmentation with Joint Hotspot Tracking
Supplementary Material

Anonymous ECCV submission

Paper ID 2189

1 Application in Interactive Scenarios

By default, the EG-Net and Gaze2Mask-Net are applied to the first frame to determine the target object and obtain the initial mask and hotspot for subsequent tracking. In such setting, our model can perform inference online in real-time. In addition to the automatic initialization, our method can also be extended to user-interactive scenarios. Specifically, users can provide eye gaze sensors or simple clicks on the object to indicate points of interest. Then the user input replaces the gaze map from EG-Net and serves as a guidance to produce the initial object mask and hotspot. After the object initialization, our weighted correlation siamese network will be used to generate the mask and hotspot track-lets.

We show an example case with the interactive initialization in a multi-object video, as shown in Fig. 1. The objects to be tracked are determined by the target object initialization module, and often all the salient objects will be included in the mask. As can be seen in Fig. 1 (b), for the video with multiple objects, our initialization module automatically identifies both objects as the target, and the hotspot tracking can consistently capture the focal regions of both objects. To implement the user interaction, we first apply simple click inside the specific object (*i.e., human head*) to define the initial hotspot and obtain the initial mask using Gaze2Mask-Net. Then the tracking module runs one inference pass for each initialized object to obtain its mask and hotspot track-lets. The results in Fig. 1 (c)(d) show that our weighted correlation siamese network is able to produce consistent hotspot tracking on the tracked object.

2 Object Hotspot Tracking Application

As introducted in the main paper, we propose a new task, object hotspot tracking, which aims to generate the intra-object salient spots estimation along the video sequence. Compared with video eye gaze tracking, our hotspot tracking could provide cleaner and more stable and temporally consistent results, which provide strong benefits to video editing tasks, such as video cropping, zooming. More descriptions about the comparisons can be found in the Sec.1 of the main paper. To implement our hotspot tracking on the aforementioned applications, we build a user interface, which allows automatic and user-specific video zooming and retargeting. In the automatic setting, given an input video, our system can automatically determine the primary objects and generate the corresponding hotspot and mask track-lets. Our system also allows users to define the obejcts by drawing bounding boxes or simple click. Implementations of the cropping algorithm are described in [1], which takes the object trajectories from mask and hotspot as reference, to produce consistent, visually pleasing cropping windows across time. The sample demo applications are shown in the supplementary videos (**UI_zooming.mp4**, **Retargeting.mp4** and **AUTO_zooming.mp4**). The videos, **UI_zooming.mp4** and **AUTO_zooming.mp4**, show the applications of user interactive zooming and automatic zooming, respectively. The video **Retargeting.mp4** shows the video retargeting process.

3 Architecture of Eye Gaze Network and Gaze to Mask Network

In the paper, we propose a Target Object Initialization module (TOI) to determine the target object to be segmented in the input video. The TOI contains two sub-networks, **EG-Net** and **Gaze2Mask-Net**, used for predicting a location estimation and the mask of the target object. We show the architecture of EG-Net and Gaze2Mask-Net in Fig. 2 and Fig. 3, respectively. The detailed illustration can be found in the Sec.3.1 of the main paper.

4 Hotspot Annotation

To quantitatively verify the performance of our hotspot tracking model, we annotate the hotspot ground truth on the DAVIS-2016 dataset [5] and make a comparison with video eye gaze models [7, 3] (see Tab. 5 in the main paper). The hotspot is sparsely annotated on every 10 frame of the video sequences in the DAVIS-2016 validation set. Five users are asked to first determine the salient part inside the object and then make consistent mouse clicks on that region across time. We illustrate several examples of our hotspot annotations in Fig. 4.

5 More Qualitative Results

We provide additional results on several sequences from the DAVIS-2016 [5], SegTrackv2 [4] and Youtube-Objects [2] in a supplementary video (1975.mp4).

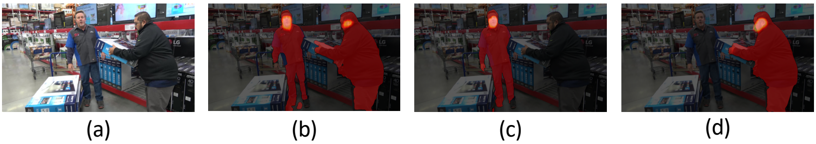

Fig. 1: The intermediate hotspot and mask tracking results under various definitions of target object. (a) Input frame. (b) The target object is automatically defined as the union of two objects by target object initialization module. (c) and (d) The target object is defined on the separate object by user click. Zoom in to see the details.

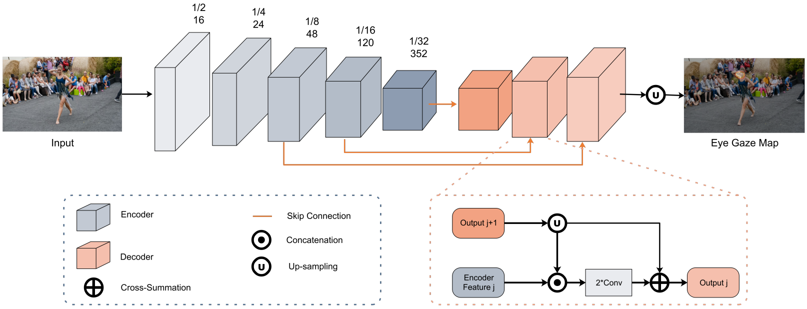

Fig. 2: Overall framework of Eye Gaze Network (EG-Net). Give an input for initialization, the EG-Net first exploits the EfficientNet [6] to extract the corresponding multi-level features. The three residual decoder blocks (shown in the orange dotted box) are stacked to generate the eye gaze map.

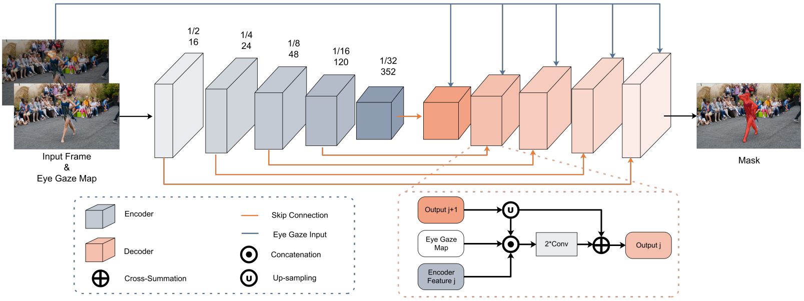

Fig. 3: Overall framework of Gaze to Mask Network (Gaze2Mask-Net). The Gaze2Mask-Net takes the input as the initial frame and the corresponding eye gaze map generated by EG-Net. The EfficientNet [6] is used to extract multi-level features. We stack five residual decoder blocks (shown in the orange dotted box) to produce the object mask in a coarse to fine manner.

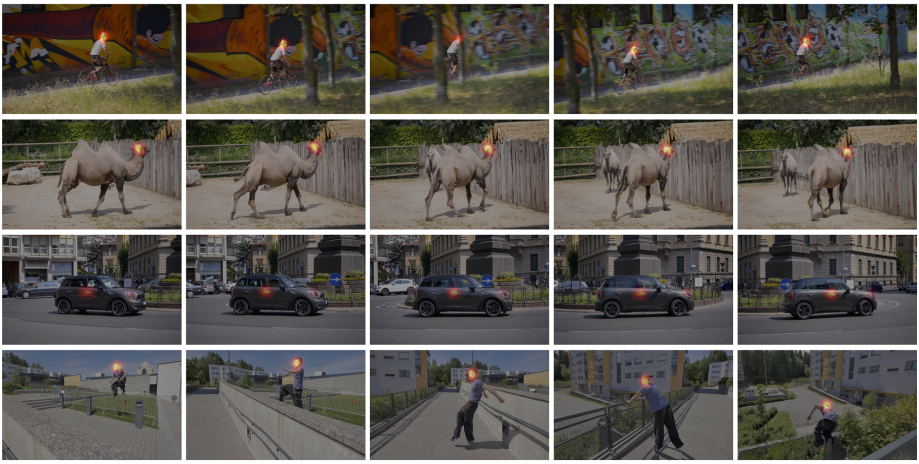

Fig. 4: Visual examples of hotspot annotations on DAVIS-2016 dataset.

,

References

1. Anonymous, et al: A system for smart video cropping, zooming and retargeting (2019)
2. Ferrari, V., Schmid, C., Civera, J., Leistner, C., Prest, A.: Learning object class detectors from weakly annotated video. In: CVPR (2012)
3. Jiang, L., Xu, M., Liu, T., Qiao, M., Wang, Z.: Deepvs: A deep learning based video saliency prediction approach. In: Proceedings of European Conference on Computer Vision (2018)
4. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: Proceedings of the IEEE International Conference on Computer Vision (2013)
5. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2016)
6. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
7. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C.H., Ling, H.: Learning unsupervised video object segmentation through visual attention. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2019)