Key Instance Selection for Unsupervised Video Object Segmentation

Donghyeon Cho* SK T-Brain
Sunguen Hong* SK T-Brain
Sungil Kang* SK T-Brain
Jiwon Kim SK T-Brain

Abstract

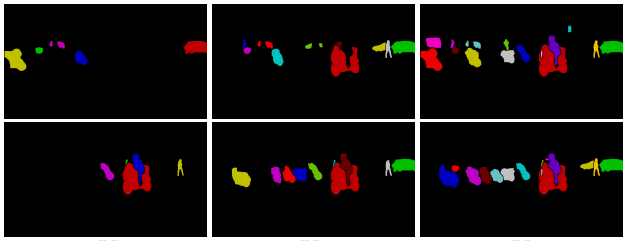
This paper proposes key instance selection based on video saliency covering objectness and dynamics for unsupervised video object segmentation (UVOS). Our method takes frames sequentially and extracts object proposals with corresponding masks for each frame. We link objects according to their similarity until the M-th frame and then assign them unique IDs (i.e., instances). Similarity measure takes into account multiple properties such as ReID descriptor, expected trajectory, and semantic co-segmentation result. After M-th frame, we select K IDs based on video saliency and frequency of appearance; then only these key IDs are tracked through the remaining frames. Thanks to these technical contributions, our results are ranked third on the leaderboard of UVOS DAVIS challenge.

1. Introduction

Given a mask in the first frame, semi-supervised video object segmentation (SVOS) is a task of generating masks in the subsequent frames. After the first SVOS challenge [11], the SVOS has been gradually getting attention. Also, datasets are continuously updated [12] or newly constructed [15]. Recently, interactive video object segmentation (IVOS) [2] and unsupervised video object segmentation (UVOS) [3] were introduced as new challenges. This paper tackles UVOS which does not require any human supervision.

The basic approach of the UVOS is to estimate a mask of the first frame, then use it to apply conventional SVOS methods [6, 8]. However, not only is it greatly influenced by the results of the first frame, it is not guaranteed that there are all targets in the first frame. To resolve these problems, we can continuously assign a new ID (i.e., instance) to an object that satisfies certain criteria. However, this increases not only the number of IDs indiscriminately but also time complexity. Therefore, it is recommended to fix the appropriate K, the maximum number of IDs. In this paper, we propose a method to select K IDs by exploring video saliency of several frames at the beginning of the video.

*They are equal contributors to this works.



(a) K=5 (b) K=10 (c) K=15

Figure 1. Segmentation results with respect to K instances on a 'scooter-black' video. Top: random instance selection. Bottom: our key instance selection.

Compared to random K instance selection, our method can capture main objects well, even in case of small K as shown in Fig. 1-(a).

The main contributions of this paper are as follows. Compared to adding new IDs continuously until the end of the video, our key instance selection approach significantly reduces time as well as memory complexity. Also, our model can capture regions of importance better than random instance selection. When we measure the similarity between assigned IDs and extracted proposals, we use all set of positive ReID descriptors for each ID. Unlike conventional methods that use optical flow [5] to handle large appearance changes in a frame sequence, we use semantic co-segmentation [9, 13]. Finally, we automatically search hyperparameters through Meta AI system developed by T-Brain under scalable environments (e.g., 144 GPUs).

2. Method

Our objective is to assign IDs to the proposals without additional inputs such as a mask of the first frame. For this, we use an object pool that manages assigning, adding, and deleting IDs. Fig. 2 illustrates the overview of the proposed method.

As a first step, we perform instance segmentation on the current frame and propagate the masks obtained in the previous frame. We then extract features and assign IDs to candidates that satisfy the criteria of the online tracker linked with object pool. Here, if the candidate matches the existing ID in the object pool, we assign the matched ID to the

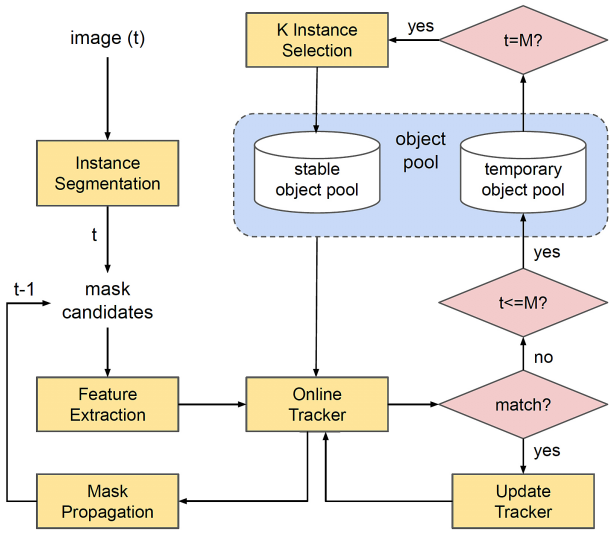


Figure 2. Overview of the proposed method.

candidate, and the online tracker is updated; otherwise, we assign a new ID to the candidate, then this new ID is added to the object pool. Adding a new ID is performed only up to the M-th frame. At the M-th frame, K instances among the accumulated IDs are selected and finally, those selected IDs are tracked in the remaining frames.

Candidate Generation: Given a frame at time I^t ($t \in T$), we perform instance segmentation to get bounding box and mask by using Mask R-CNN [7] and DeepLab [4]. Motion blur or occlusion, which often occurs in videos, can result in poor segmentation results for certain frames. To address this issue, we use masks propagated from the previous frame as mask candidates. Concretely, we utilize RGMP [9, 13] in our experiments.

ID Assignment: To assign an ID to each candidate, we compute specific scores by comparing the candidates and the registered IDs in the object pool. The first score is $S_{iou}(l, n)$ which is IOU between a mask from ID ($l \in L$) and a mask of the candidate ($n \in N$). Here, L refers to the number of IDs registered in the object pool and N is the number of candidates. We omit notation time t for simplicity. The second score is $S_{traj}(l, n)$ that measures how far the candidate is from the predicted bounding box of ID as

$$S_{traj}(l, n) = 1 - \min\left(\frac{dot(\vec{b}_l, \vec{b}_n)}{\alpha_{traj}}, 1\right), \tag{1}$$

where α_{traj} is a normalization factor. Also, $\vec{b}_l$ and $\vec{b}_n$ are vectors from bounding boxes of ID and candidate, respectively. The third score $S_{reid}(l, n)$ considers a distance of

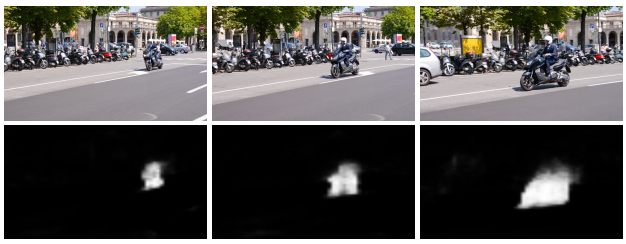


Figure 3. Video saliency results. Note that a motorcycle with motion is only focused among a lot of objects in a scene.

ReID descriptors [10] between ID l and candidate n. Unlike [6, 8], our method use the nearest ReID descriptor among the all set of positive ReID descriptors of ID l as

$$S_{reid}(l, n) = 1 - \min\left(\frac{\min_j ||d_l(j) - r_n||}{\alpha_{reid}}, 1\right), \tag{2}$$

where r_n is a ReID descriptor for candidate n and $d_l(\cdot)$ are positive ReID descriptors for ID l. The last score is a relative ReID score $S_{rel}(l, n)$:

$$S_{rel}(l, n) = \frac{S_{reid}(l, n)}{\max_l S_{reid}(l, n)}. \tag{3}$$

Total score is weighted summation of above four scores:

$$\begin{aligned} S_{total}(l, n) = w_{iou} \cdot S_{iou}(l, n) + w_{traj} \cdot S_{traj}(l, n) \\ + w_{reid} \cdot S_{reid}(l, n) + w_{rel} \cdot S_{rel}(l, n), \end{aligned} \tag{4}$$

where w_{iou}, w_{traj}, w_{reid}, and w_{rel} are weight factors of each term. Finally, we assign ID l to the candidate object as follows:

$$\hat{n} = \begin{cases} \underset{n}{\text{argmax}}\, S_{total}(l, n), & \text{if } \geq \tau_c \\ \text{None}, & \text{otherwise} \end{cases}, \tag{5}$$

where $\tau_{c=\{1,2\}}$ is a threshold value for cutting off object with low confidence. Before selecting the K instances ($t <= M$) as in Sec. 2, c is 1, and after that c is 2. If one of candidates is assigned to ID l, then $d_{\hat{n}}$ is added to pool of positive ReID descriptors for ID l. Also $\vec{b}_l$ is updated by using $\vec{b}_{\hat{n}}$.

Key Instances Selection: Basically, the pipeline mentioned above is iterated at each frame. There is ID addition process at the beginning of the video. Especially, in the first frame, object candidates with high confidence are added as new IDs. From the second frame and M-th frame, new IDs are added when object candidates have high objectness score and are not overlapped with objects assigned to existing IDs. After M-th frame, we select at most K IDs. As selection criteria, we use weighted summation of video saliency score [14] and frequency of each ID as

$$S_{sel}(l) = w_{sal} \cdot S_{sal}(l) + w_{freq} \cdot S_{freq}(l), \tag{6}$$

Table 1. Segmentation results on UVOS DAVIS challenge dataset.

Measure	RWTH Vision	Oxford-CASIA	**SK T-Brain (ours)**	UVOS-test	RWTH Vision 2	ZX$_V IP$	VIG	UOC-UPC-BSC
Ranking	1	2	3	4	5	6	7	8
Global Mean ↑	0.564	0.562	0.516	0.504	0.481	0.471	0.448	0.412
$\mathcal{J}$ Mean ↑	0.564	0.535	0.487	0.475	0.460	0.435	0.422	0.379
$\mathcal{J}$ Recall ↑	0.609	0.613	0.551	0.542	0.514	0.490	0.476	0.413
$\mathcal{J}$ Decay ↓	0.015	-0.021	0.040	0.032	0.084	0.035	0.035	0.076
$\mathcal{F}$ Mean ↑	0.594	0.590	0.545	0.533	0.503	0.506	0.474	0.444
$\mathcal{F}$ Recall ↑	0.641	0.632	0.594	0.569	0.538	0.543	0.506	0.473
$\mathcal{F}$ Decay ↓	0.058	0.001	0.077	0.055	0.118	0.067	0.068	0.117

where w_{sal} and w_{freq} are weight factors of each term. $S_{sal}(l)$ and $S_{freq}(l)$ are computed over frames. Fig. 3 shows effectiveness of video saliency based approach. The frequency of each ID means the number of each ID's appearance up to the M-th frame. Because there are several IDs that are not connected to any proposal at a certain time by τ_1, $S_{freq}(l)$ can vary by ID.

Hyperparameter Search: We perform hyperparameter search for w_{iou}, w_{traj}, w_{reid}, w_{rel}, τ_1, τ_2, w_{sal} and w_{freq} on DAVIS validation dataset. Global mean ($\mathcal{J}\&\mathcal{F}$) of results is 0.599 and searched weighting factors are $0.12, 0.575, 0.3, 0.0065, 0.55, 0.35, 0.5$ and 1.0, respectively.

3. Experiments

We evaluate the proposed method on UVOS Davis challenge dataset [3], which contains 30 videos without a mask from the first frame in each video. We directly submit our results to the CodaLab site [1] to get segmentation results. Evaluation metrics are Region Jaccard ($\mathcal{J}$) and Boundary F measure ($\mathcal{F}$) for each instance. As shown in Table. 1, our method achieves the third rank with respect to Global Mean as well as all the other specific metrics.

Fig. 4 shows qualitative results on DAVIS validation set. In the first row of Fig. 4, our method well captures *blackswan* over time. The proposed method also shows robust video segmentation results even in the relatively dynamic *Parkour* example. In addition, our method faithfully works on multiple object segmentation (from the third row to the last low in Fig. 4). Note that changes in scale (*lab-coat*) and appearance (*mbike-trick*) are handled appropriately by our method.

The proposed key instance selection method is quantitatively compared with random instance selection according to the K as shown in Table. 2. Since the proposed scheme focuses on semantically meaningful regions, it shows promising results even at a small K.

4. Conclusion

We have presented key instance selection for unsupervised video object segmentation (UVOS). Our method allows maximum K instances to be tracked by considering video saliency and frequency of appearance. Focusing on objects that are in the spotlight enables to reduce time complexity. In addition, objects in a frame sequence are linked by specific scores from ReID descriptor, trajectories, and co-segmentation result. Finally, we perform hyperparameter search to find out the optimal hyperparameters in our model by Meta AI system. Our method showed competitive results in UVOS DAVIS challenge.

Table 2. Global Mean ($\mathcal{J}\&\mathcal{F}$) scores on the DAVIS validation dataset according to the number of IDs. We compare our key instance selection method with random instance selection.

# of IDs	5	10	15	20
Random	0.474	0.529	0.587	0.594
Ours	0.576	0.579	0.591	0.599

5. Acknowledgements

We thank Youngsoon Lee (Sunny) for help to use Meta AI system for searching hyperparameters.

References

[1] Codalab. `https://competitions.codalab.org/competitions/21739#results`.

[2] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv:1803.00557*, 2018.

[3] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig" Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2018.

[5] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox". Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] B. Leibe J. Luiten, P. Voigtlaender. Premvos: Proposal-generation, refinement and merging for the davis challenge

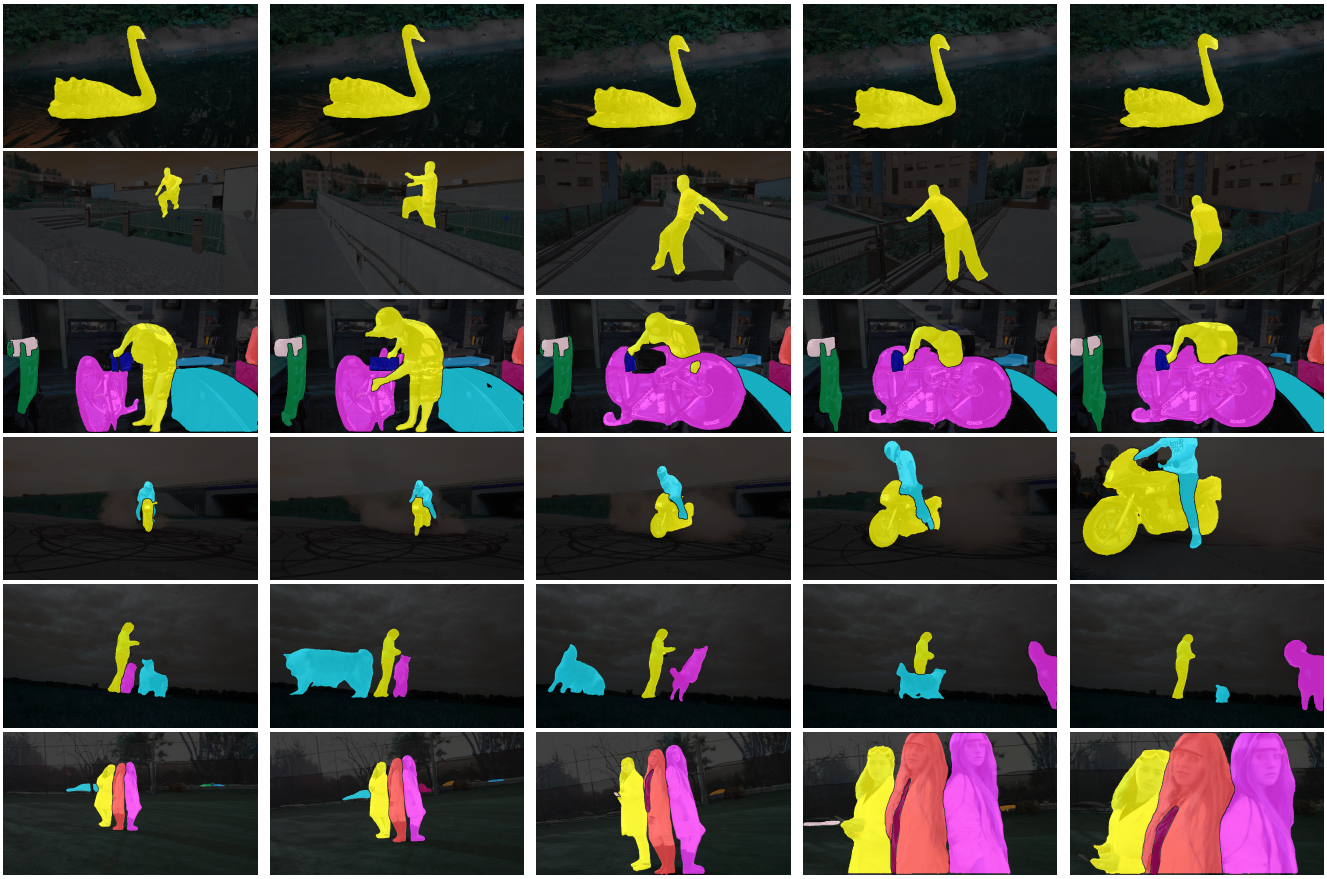


Figure 4. Qualitative results on DAVIS validation set that contains a single object and multiple objects.

on video object segmentation 2018. *The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2018.

[7] P. Dollr K. He, G. Gkioxari and R. Girshick. Mask r-cnn. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2017.

[8] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Proc. of Asian Conf. on Computer Vision (ACCV)*, 2018.

[9] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] Aljosa Osep, Paul Voigtlaender, Jonathon Luiten, Stefan Breuers, and Bastian Leibe. Large-scale object discovery and detector adaptation from unlabeled video. *CoRR*, abs/1712.08832, 2017.

[11] 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.

[12] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[13] N. Xu S. Joo Kim S. Wug Oh, J. Lee. Fast user-guided video object segmentation by deep networks. *The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2018.

[14] J. Shen W. Wang and L. Shao. Video salient object detection via fully convolutional networks. *IEEE Trans. Image Processing (TIP)*, 27(1):38–49, 2018.

[15] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018.