The 2019 DAVIS Challenge on Video Object Segmentation - CVPR 2019 Workshops

UnOVOST: Unsupervised Offline Video Object Segmentation and Tracking for the 2019 Unsupervised DAVIS Challenge

Idil Esen Zulfikar* Jonathon Luiten* Bastian Leibe
Computer Vision Group, RWTH Aachen University
idil.esen.zuelfikar@rwth-aachen.de {luiten,leibe}@vision.rwth-aachen.de

Abstract

We address Unsupervised Video Object Segmentation (UVOS), the task of automatically generating accurate pixel masks for salient objects in a video sequence and of tracking these objects consistently through time, without any information about which objects should be tracked. Towards solving this task, we present UnOVOST (Unsupervised Offline Video Object Segmentation and Tracking) as a simple and generic algorithm which is able to track a large variety of objects. This algorithm hierarchically builds up tracks in five stages. First, object proposal masks are generated using Mask R-CNN. Second, masks are sub-selected and clipped so that they do not overlap in the image domain. Third, tracklets are generated by grouping object proposals that are strongly temporally consistent with each other under optical flow warping. Fourth, tracklets are merged into long-term consistent object tracks using their temporal consistency and an appearance similarity metric calculated using an object re-identification network. Finally, the most salient object tracks are selected based on temporal track length and detection confidence scores. We evaluate our approach on the DAVIS 2017 Unsupervised dataset and obtain state-of-the-art performance with a mean $\mathcal{J}\&\mathcal{F}$ score of 58% on the test-dev benchmark. Our approach further achieves first place in the DAVIS 2019 Unsupervised Video Object Segmentation Challenge with a mean of $\mathcal{J}\&\mathcal{F}$ score of 56.4% on the test-challenge benchmark.

1. Introduction

Video Object Segmentation (VOS) aims at automatically generating accurate pixel masks for objects of each frame in a video, then associating those proposed object pixel masks in the successive frames to obtain temporally consistent tracks. VOS has been studied mostly as a semi-supervised task which allows using the ground truth object masks that are given in the first frame of the video sequences. However, VOS can be also studied as an unsupervised task in which object tracks are generated without any knowledge about which objects or how many objects should be tracked. In this paper, we present the UnOVOST (Unsupervised Offline Video Object Segmentation and Tracking) method for the unsupervised VOS task and evaluate it on the unsupervised DAVIS benchmark dataset [4]. An overview of UnOVOST can be seen in Figure 1. Moreover, our method achieves the first place in the DAVIS 2019 Unsupervised Video Object Segmentation Challenge.

*Equal Contribution

2. Related Work

Currently, many methods have been presented for semi-supervised VOS with multiple objects [16, 26, 25, 24] where the ground-truth of the first frame is given. Modern deep learning based methods are able to approach this task both accurately and efficiently. On the other hand, the deep learning based unsupervised VOS methods mainly work on videos with only a single object [12, 13, 11, 22, 23], with a few methods tackling the multi-object task [2, 31, 6]. FSEG [11] and LVO [23] concatenate appearance and motion features from a two-stream FCNs, then segment videos via a fusion model and a visual memory model, bidirectional convolutional GRU, respectively. In the multi-object scenario, [2, 31, 6] are based on motion segmentation. CCG [2] develops a two-stage statistical model that estimates piece-wise rigid motion, then merges rigid motion into objects using semantic segmentation with CNNs. [6] adapts a Mask R-CNN [7] architecture to the video object segmentation domain by combining both appearance and motion features to an RPN model. [31] predicts foreground masks of moving objects using pixel-wise-features extracted from appearance and motion cues, then learns and clusters feature embeddings of the foreground masks to segment different objects. As seen, the proposed methods rely on appearance and motion change throughout a video sequence; as a result, they are only able to segment the moving objects and omit the static objects.

3. Approach

We propose UnOVOST as a new method which tackles the unsupervised VOS task. Our approach consists of five stages. The first stage is object proposal mask generation. Here, we use an instance segmentation network to obtain a set of object proposals per frame without distinguishing objects as foreground and background objects or static and moving objects. This is in contrast to previous approaches [31, 2, 6] which focus on finding foreground objects and assume foreground objects are moving objects. Instead of using this assumption, we pursue a broader approach to identify objects in the foreground and

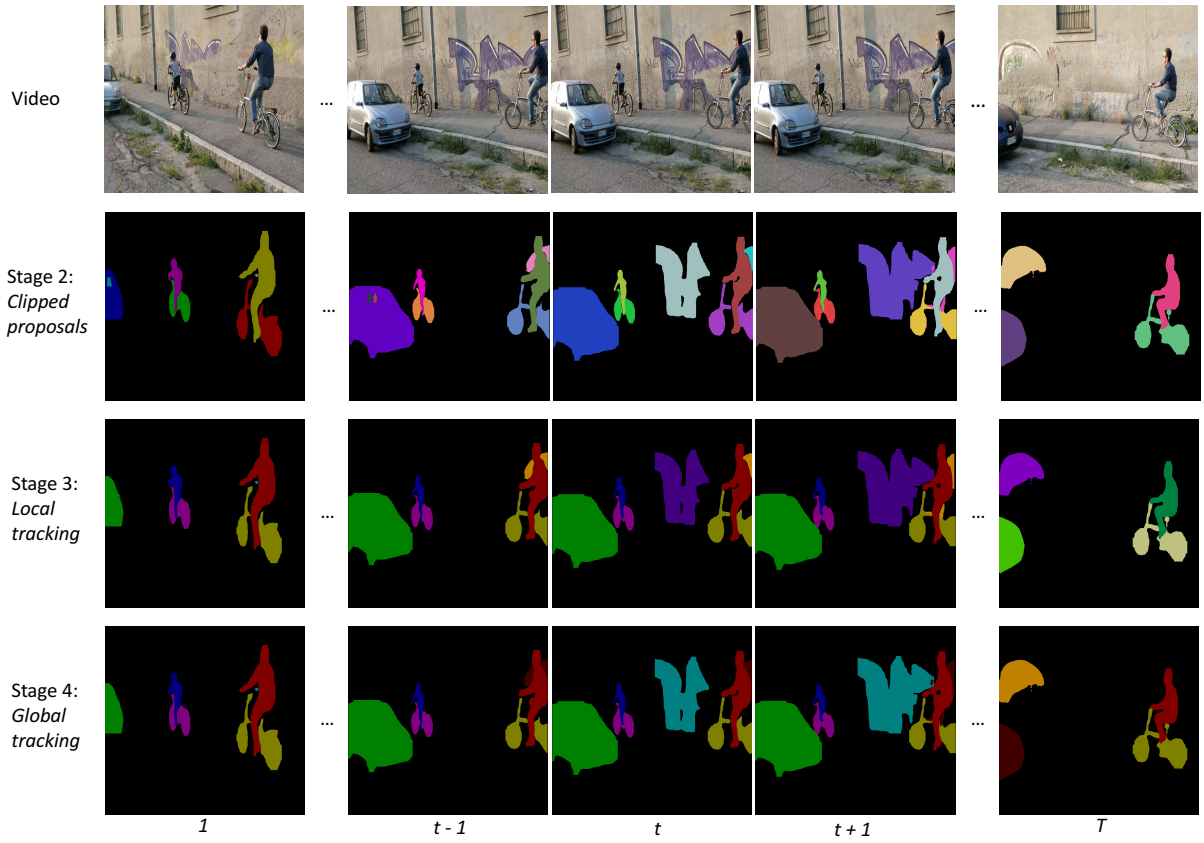

Figure 1. UnOVOST overview. Stage 1 generates object mask proposals. Stage 2 sub-selects and clips proposals to be non-overlapping. Stage 3 generates temporally local tracklets. Stage 4 merges tracklets into long-term tracks. Stage 5 selects the final set of tracks.

background, both stationary and in motion. Our approach is able to detect many objects from a large number of categories. This stage also shows that the already available detectors are able to segment many object categories without video or dataset specific fine-tuning. After generating object proposal masks, in the second stage, an algorithm is applied on overlapping proposal masks to sub-select and clip a set of non-overlapping object proposals in the video frame. In the third stage, the object proposal masks in successive frames are connected using optical-flow warping to generate tracklets that provide temporally consistent mask identities throughout short sections of the video sequence. However, these tracklets might not ensure long-term consistency throughout the video sequence, i.e. the tracklets can begin and terminate any time in the video. Therefore, we present a merging algorithm in the fourth stage to merge the tracklets for long-term consistency of object tracks. In the final stage, the most salient object tracks are selected by considering the temporal length of tracks and the detection confidence score. This last step is only necessary because the maximum number of object tracks per sequence is limited to 20 in the DAVIS 2019 Unsupervised VOS Challenge. We observe that UnOVOST produces more than 20 object tracks for many sequences which makes our method quite generic. An overview of the stages of our algorithm can be seen in Figure 1. In the following, the implementation details of each of the stages are explained.

Object Mask Proposal Generation. Object proposal masks are generated using a Mask R-CNN [7] implementation by [28] with a ResNet101 [8] backbone trained on COCO [15]. This network produces masks, bounding boxes, object categories and confidence scores for object proposals as outputs. In our algorithm, we benefit from only mask proposals and confidence scores. We extract proposal masks with a confidence score greater than 0.1.

Proposal Sub-Selection and Clipping. We implement an algorithm to prevent overlapping proposals in each frame. All proposal masks in a frame are compared against one another using their intersection over union (IoU) to detect overlaps. If the IoU between two proposal masks is higher than a 0.2 threshold, then the proposal mask with higher confidence score is held and the other proposal mask is clipped. In this way, we get more accurate and consistent pixel masks for the next steps of our algorithm.

Tracklet Generation. To connect the object proposal masks temporally, first each proposal mask in a frame is warped to the next frame using optical flow calculated with PWCNet [21]. Then, we create a complete bipartite

			Ours	Oxford-CASIA	SK T-Brain
U17 T-C	$\mathcal{J}\&\mathcal{F}$	Mean	**56.4**	56.2	51.6
	$\mathcal{J}$	Mean	53.4	**53.5**	48.7
		Recall	60.9	**61.3**	55.1
		Decay	1.5	**-2.1**	4.0
	$\mathcal{F}$	Mean	**59.4**	59.0	54.5
		Recall	**64.1**	63.2	59.4
		Decay	5.8	**0.1**	7.7
U17 T-D	$\mathcal{J}\&\mathcal{F}$	Mean	**58.0**	56.5	54.2
	$\mathcal{J}$	Mean	**54.0**	51.7	50.0
		Recall	**62.9**	59.9	58.9
		Decay	**3.5**	21.7	8.4
	$\mathcal{F}$	Mean	**62.0**	61.4	58.3
		Recall	**66.6**	65.7	62.1
		Decay	**6.6**	15.7	11.4
U17 Val	$\mathcal{J}\&\mathcal{F}$	Mean	67.0	-	-
	$\mathcal{J}$	Mean	65.6	-	-
		Recall	75.5	-	-
		Decay	0.3	-	-
	$\mathcal{F}$	Mean	68.4	-	-
		Recall	75.9	-	-
		Decay	3.7	-	-

Table 1. Our results compared to the top two other competitors in the 2019 Unsupervised DAVIS VOS Challenge on the three DAVIS 2017 Unsupervised benchmark datasets: the `test-challenge` set (U17 T-C), the `test-dev` set (U17 T-D), and the `val` set (U17 Val).

	U17 Val	U17 T-D	U17 T-C
Mask R-CNN	0.74	0.78	0.77
Optical Flow	0.10	0.14	0.12
ReID	0.10	0.15	0.11
UnOVOST Tracking	0.08	0.07	0.06
Total	1.02	1.15	1.06

Table 2. Runtime analysis of UnOVOST on the DAVIS 2017 Unsupervised `val`, `test-dev` and `test-challenge` datasets. Times are seconds per frame.

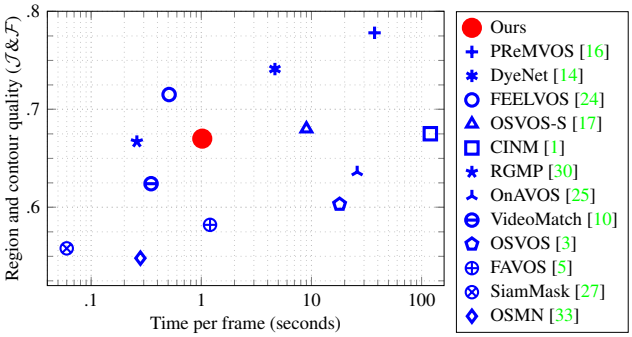

Figure 2. Quality versus timing plot comparing UnOVOST to state-of-the-art semi-supervised methods on DAVIS17 `val`. All methods other than ours are "semi-supervised" and use the given first-frame ground-truth. Our methods obtains similar results while working in an "unsupervised" manner without using any given information about which objects should be tracked.

graph whose nodes are the proposal masks in the successive frames and whose edge costs are IoU scores between the warped proposal mask and the proposal mask in the next frame. We then generate tracklets by solving a bipartite graph matching problem maximizing the total IoU score, where each object proposal mask is matched with only one proposal mask in the next frame. In order to solve the matching problem, a greedy algorithm is used that only considers pairs with a flow-warped IoU greater than 0.03. If a proposal is not matched, this ends a tracklet. It is possible that tracklets may span only a single frame.

Merging Tracklets. Our merging algorithm uses nearest neighbour matching of proposals' ReID embedding vectors. The ReID embedding vector for each object proposal mask is calculated using a network based on a triplet-loss proposed in [18]. This network is a wide ResNet variant [29] trained on COCO [15] and YouTube-VOS [32]. It is trained using the batch-hard loss with a soft-plus margin [9]. This network extracts a ReID embedding feature vector for each proposal. The ReID vectors of the proposals in a tracklet are averaged to obtain a tracklet ReID embedding vector used as a similarity metric. We implement a merging algorithm to merge tracklets that are similar to each other in terms of the average ReID vector. Our merging algorithm starts by finding a list of compatible tracklets for each tracklet. Tracklet A is compatible with tracklet B if A ends before the start of B or vise versa. We define the distance between tracklets as the L2 distance of their ReID feature vectors. For each tracklet, the preceeding tracklet with the lowest L2 distance is linked with this tracklet. In this way, each tracklet is only linked with one preceding tracklet.

However, a tracklet should be thought as a node with the two links, the before-link and the after-link. Our linking procedure assures that the before-link of a tracklet is connected with only one tracklet, but a tracklet can be the nearest compatible tracklet for multiple successor tracklets resulting in multiple after-links. This link structure is represented in a forest with at least one tree whose root corresponds to one of the tracklets with the earliest starting time and each path in the tree is a possible object track. Therefore, a selection strategy is needed to determine the best possible object tracks.

We apply a greedy recursive track selection strategy. Given a tree, we first consider only the paths from the root node to the leaf nodes. Each of these paths are scored with the sum of temporal gaps between the successive tracklets; this is the number of frames between the finishing time of the tracklet, the parent node, and the starting time of the next tracklet, the child node. For each tree, we select the path with the lowest sum of temporal gaps as the best object track, and add this to the set of final object tracks. We then create a new set of trees by removing all nodes belong-

ing to this path from the tree, resulting in possibly multiple child-trees being created. If a child-tree is created that is only a single path, this is automatically added to the list of final tracks. We apply this greedy path selection recursively until there are no more trees. In this way, we select a set of object tracks which do not include any overlapping tracklets and have long-term temporal consistency.

Final Tracks Selection. A saliency score $S_{sal,i}$ is calculated for each track using each tracklet t_j in the track i:

$$S_{sal,i} = \sum_j temp(t_j)\ conf(t_j) \tag{1}$$

where $temp(t_j)$ is the temporal length of tracklet j and $conf(t_j)$ is the average of the confidence scores of proposal masks in tracklet t_j. Finally, if there are more than 20 tracks, the 20 tracks with the highest saliency score are selected as the final object tracks.

4. Experiments

We evaluate our algorithm on the DAVIS 2017 Unsupervised benchmark datasets [4]. The DAVIS 2017 Unsupervised datasets contain the video sequences for unsupervised multi-object segmentation in the same way as the DAVIS 2017 Semi-Supervised dataset [20] extends the DAVIS 2016 [19] to multiple objects. The DAVIS 2017 Unsupervised `val` set has the same sequences as the DAVIS 2017 Semi-Supervised `val` set. However, the `test-dev` and `test-challenge` sets contain new videos introduced for the DAVIS 2019 Unsupervised VOS Challenge. All of the DAVIS 2017 Unsupervised datasets include multiple objects per video sequence. Table 1 shows our results on three DAVIS 2017 Unsupervised benchmarks. The $\mathcal{J}\&\mathcal{F}$ metrics are used in the evaluation, more details on these metrics can be found in [20]. We also perform a runtime analysis for our algorithm, which can be seen in Table 2. Our algorithm gives better results on the `test-dev` and `test-challenge` datasets than all other competitors in the DAVIS 2019 Unsupervised VOS Challenge for the $\mathcal{J}\&\mathcal{F}$ metric which is the mean of the $\mathcal{J}$ and $\mathcal{F}$ scores. In Figure 2, we compare UnOVOST for both its speed an accuracy on the DAVIS17 `val` benchmark to state-of-the-art semi-supervised VOS methods. Even though UnOVOST is an unsupervised method, without using any given information about which objects are to be tracked, it is still able to perform comparatively to many state-of-the-art methods that do use the first frame as input.

5. Conclusion

In this paper, present the UnOVOST algorithm to tackle the unsupervised VOS task. This achieves state-of-art results and wins the DAVIS 2019 Unsupervised Video Object Segmentation Challenge.

Acknowledgements: This project has been funded, in parts, by ERC Consolidator Grant DeeViSe (ERC-2017-COG-773161) and by a Google Faculty Research Award. We would like to thank Paul Voigtlaender for helpful discussions.

References

[1] L. Bao, B. Wu, and W. Liu. CNN in MRF: video object segmentation via inference in a cnn-based higher-order spatio-temporal MRF. In *CVPR*, 2018. 3
[2] P. Bideau, A. RoyChowdhury, R. R. Menon, and E. Learned-Miller. The best of both worlds: Combining cnns and geometric constraints for hierarchichal motion segmentation. In *CVPR*, 2018. 1
[3] 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. 3
[4] S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019. 1, 4
[5] J. Cheng, Y.-H. Tsai, W.-C. Hung, S. Wang, and M.-H. Yang. Fast and accurate online video object segmentation via tracking parts. In *CVPR*, 2018. 3
[6] A. Dave, P. Tokmakov, and D. Ramanan. Towards segmenting anything that moves, 2019. 1
[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 2
[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016. 2
[9] A. Hermans*, L. Beyer*, and B. Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, 2017. 3
[10] Y.-T. Hu, J.-B. Huang, and A. G. Schwing. Videomatch: Matching based video object segmentation. In *ECCV*, 2018. 3
[11] S. D. Jain, B. Xiong, and K. Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. *CVPR*, 2017. 1
[12] S. Li, B. Seybold, A. Vorobyov, A. Fathi, Q. Huang, and C.-C. J. Kuo. Instance embedding transfer to unsupervised video object segmentation. *CVPR*, 2018. 1
[13] S. Li, B. Seybold, A. Vorobyov, X. Lei, and C.-C. Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *ECCV*, September 2018. 1
[14] X. Li and C. Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *ECCV*, 2018. 3
[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 3
[16] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018. 1, 3
[17] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. L. Taixé, and L. Van Gool. Video object segmentation without temporal information. *PAMI*, 2018. 3
[18] A. Ošep, P. Voigtlaender, J. Luiten, S. Breuers, and B. Leibe. Large-scale object mining for object discovery from unlabeled video. *ICRA*, 2019. 3
[19] 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 *CVPR*, 2016. 4
[20] 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. 4
[21] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. *arXiv preprint arXiv:1709.02371*, 2017. 2
[22] P. Tokmakov, K. Alahari, and C. Schmid. Learning motion patterns in videos. *CVPR*, 2017. 1
[23] P. Tokmakov, K. Alahari, and C. Schmid. Learning video object segmentation with visual memory. *ICCV*, 2017. 1
[24] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen. FEELVOS: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019. 1, 3
[25] P. Voigtlaender and B. Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *BMVC*, 2017. 1, 3
[26] P. Voigtlaender, J. Luiten, and B. Leibe. BoLTVOS: Box-Level Tracking for Video Object Segmentation. *arXiv:1904.04552*, 2019. 1
[27] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019. 3
[28] Y. Wu et al. Tensorpack. `github.com/tensorpack/`, 2016. 2
[29] Z. Wu, C. Shen, and A. van den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 2019. 3
[30] S. Wug Oh, J.-Y. Lee, K. Sunkavalli, and S. Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, 2018. 3
[31] C. Xie, Y. Xiang, Z. Harchaoui, and D. Fox. Object discovery in videos as foreground motion clustering, 2018. 1
[32] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, 2018. 3
[33] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos. Efficient video object segmentation via network modulation. In *CVPR*, 2018. 3