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

Video Segmentation by Detection for the 2019 Unsupervised DAVIS Challenge

Zhao Yang*
University of Oxford
zhao.yang@eng.ox.ac.uk

Qiang Wang*
CASIA, INTELLIMIND LTD
qiang.wang@nlpr.ia.ac.cn

Song Bai
University of Oxford
songbai.site@gmail.com

Weiming Hu
CASIA
wmhu@nlpr.ia.ac.cn

Philip H.S. Torr
University of Oxford
philip.torr@eng.ox.ac.uk

Abstract

In this work, we present a new framework, video segmentation by detection (VSD), for tackling the problem of unsupervised video multi-object segmentation. Our model employs an object detector for automatic target discovery and a set of single-object trackers for the simultaneous tracking of all targets.

While addressing the object re-identification problem, we observe that many of the objects of interest in the dataset are humans or human centric such as bicycles. As such, following a design philosophy that special purpose algorithms will always be better than general purpose ones, we explore whether we can leverage the rich existing research efforts on re-identifying humans to improve the results or exploit the spatial relations of human-centric objects to humans. The proposed method achieves the highest $\mathcal{J}$-Mean of 0.535 and an overall second place in the unsupervised track of the 2019 DAVIS Challenge.

1. Introduction

The 2019 DAVIS Challenge [3] sets out the task of *unsupervised video multi-object segmentation*, which requires the segmentation of all objects that are likely to capture human attention in a video, without receiving any hint what or where they are. Unsupervised refers to zero human interaction at inference time, as opposed to initializing target objects with masks.

In addition to difficulties such as occlusions, large deformation, cluttered scenes, *etc.*, the unsupervised multi-object setting requires the automatic detection of target objects. We discover that the naïve use of an object detector as initialization does not perform well. The main issues lie in a lack of a (re-)detection mechanism for lost/new targets and the frequent identity switches that happen when trajectories overlap, which we believe is a consequence of not jointly considering trajectories of all target objects.

We consider video object segmentation (VOS) from a tracking perspective, and borrow insights from multi-object tracking [7] and *tracking by detection* [16]. Our underlying idea is to associate detections across frames into a set of object trajectories via a data association algorithm. To this end, we present a novel framework, video segmentation by detection (VSD), which exploits state-of-the-art single object trackers for online visual tracking and detection association. Importantly, as occlusions happen frequently, objects re-entering a scene can often be falsely assigned new identities. As such, we develop an offline trajectory merging strategy for addressing object identity re-association.

Many objects of interest in the 2019 DAVIS Challenge are humans or human centric such as bicycles. Observing this and following a design philosophy that a special purpose algorithm is better than a general purpose one, we leverage existing research efforts on person re-identification for merging human trajectories, and in order to re-associate objects that are not people but are human centric, we match their identities based on their spatial relations to humans, which works well for instance with classes like horse and bicycle. Our method achieves the highest $\mathcal{J}$-Mean of 0.535 and an overall second place in the 2019 DAVIS Challenge.

2. Related Work

Object detection provides a natural mechanism for automatic target discovery for our task, as it seeks to classify and localize all objects in a given image. We employ the state-of-the-art two-stage detector Mask-RCNN [6], which is one of the top detectors and can output detection masks, an important feature that enhances the segmentation accuracy of our proposed model.

Visual object tracking focuses on tracking a single object with initialization given as a bounding box. We adopt

*Equal contribution. Order determined by a coin flip.

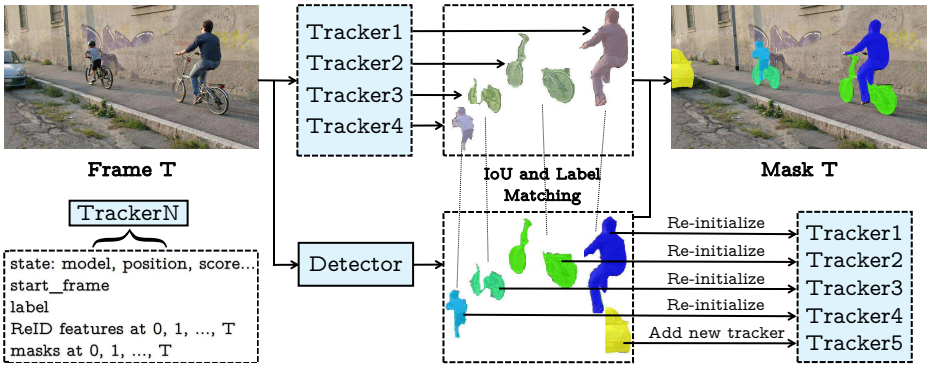

Figure 1. Online inference at time step T. The detector predicts all objects in the current frame, each of which is matched to a tracking mask predicted by a tracker, by comparing their semantic labels and thresholding the intersection-over-union. Matched detection masks are used to re-initialize the corresponding trackers, and unmatched detection masks that satisfy a saliency requirement (explained in Section 3) are used to initialize new trackers. The final output consists of a set of non-overlapping detection masks and tracking masks.

the popular approach of Siamese networks [2, 13, 17] and SiamMask [13] in particular, for joint online tracking and segmentation. For other popular methods such as the correlation filters, we refer readers to [4, 5]. The underlying idea of a Siamese tracker is to locate an object via dense cross correlations between features of a target template and features of a search region, encoded by two symmetric networks with shared weights. As cross correlation encodes similarity, the location of the object can be interpolated from the position of the maximum response.

A drawback of Siamese trackers is that they tend to gradually lose the target and drift away as time progresses (despite many recent attempts [2, 17] at improvements). In this work, we remedy this issue by re-initializing each online tracker at every time step, with matched detection masks from the object detector. As trackers are re-calibrated at every step, tracking errors do not accumulate and the drifting problem is alleviated significantly.

Multi-object tracking is addressed most commonly via the tracking-by-detection approach, which first detects objects independently in each frame, and then links these detections into a set of trajectories across frames. Popular models for linking detections include graph-based optimization [16, 11], binary classification or representation learning via deep learning [8, 14], reinforcement learning [10, 15], *etc*. We explore a separate path that solely relies on state-of-the-art single-object trackers for optimal online tracking, which are also empowered to segment.

Video object segmentation not only tracks individual objects, but precisely delineates them with masks. Due to space limits, we focus on the unsupervised setting defined in Section 1. Previously, this problem is largely cast as foreground-background segmentation [9, 12], which implicitly avoids the predicament of automatic target discovery. Under the new multi-object setting, we analyze the key issues and explore a new direction for this problem, which is formally introduced in the following section.

3. Method

Our overall video segmentation by detection (VSD) framework is composed of a segmentation-by-detection phase and a trajectory merging phase. The former is specifically designed to exploit the power of single-object trackers for optimal tracking performance, as well as to output masks instead of bounding boxes for the purpose of object segmentation. On the other hand, trajectory merging focuses on the major challenge of correctly re-associating a misidentified "new" object to its original identity, and achieves this by merging trajectories via person re-identification (ReID) [1] and a subsequent relation matching scheme that we develop and term as relation association.

Algorithm 1 Video Segmentation by Detection

// Hyperparameters θ_1 and θ_2 and function $salient()$
// are explained in the text.
Initialize an empty list of trackers $tr\_list \leftarrow \emptyset$
for frames $T = 0, 1, ..., M$ **do**
  $D \leftarrow$ Detect objects in T with threshold θ_1
  $TR \leftarrow$ Update trackers in tr_list with threshold θ_2
  DM,$TRM \leftarrow$ Match masks in D with masks in TR
  Use DM to re-initialize trackers in tr_list
  Initialize an empty set of new targets $DN \leftarrow \emptyset$
  for $x \in D \setminus DM$ **do**
    if $salient(label(x))$ **then**
      Add x to DN
      Initialize a new tracker tr with x
      Append tr to tr_list
    end if
  end for
  Output masks $DM \cup DN \cup (TR \setminus TRM)$
end for

Segmentation by detection. Figure 1 schematically illustrates this process, and Algorithm 1 details the algo-

rithmic steps. We adopt Mask-RCNN [6] as the object detector and SiamMask [13] as the single-object tracker, both of which are state-of-the-art models capable of segmentation. Mask-RCNN is responsible for target discovery and tracker (re-)initialization at each step, and replaces the tracking masks of SiamMask with detection masks as output, since single-object trackers, including SiamMask, commonly suffer prediction drifts down stream. As an online tracker, SiamMask is responsible for tracking an object from one frame to the next. In cases of successful detections in both frames, it essentially links the two detections into a trajectory, and in cases of any missed detection, its tracking mask is used as the output. At each time step, the detector and trackers output detection and tracking masks, which are matched by IoU and category labels. Then each existing tracker is re-initialize with the detection matched to its tracking mask and a new tracker is added for any new target detection.

The parameters θ_1 and θ_2 in Algorithm 1 refer to confidence thresholds for the predictions of Mask-RCNN and SiamMask respectively, which are necessary for removing false positive detections and poor tracking masks. The function $salient$ in Algorithm 1 refers to identifying the semantic class of an object as salient in the given video, defined as appearing in more than a fraction of the total frames. This class saliency requirement prevents the tracking of insignificant objects.

Trajectory merging. If an object reappears after occlusion or being out of view (which indeed happens frequently), it is often assigned a new identity by the detector. Thus, we design the offline trajectory merging phase to match identities that actually belong to the same object and merge their trajectories.

As motivated in Section 1, we start with "person" objects via a person ReID module. In the previous phase, each person tracker appends a ReID feature vector at each frame, which is extracted using a pretrained ReID model. Then while offline, we first identify all frames at which new person identities are created, which we call the entry points and denote the group of identities created at points $[1, 2, ...m]$ as $[P_1, P_2, ...P_m]$ accordingly, assuming m entry points without loss of generality. Then we progressively match identities and merge trajectories between two adjacent groups P_t and P_{t+1} for $t = 1, ..., m$. At each step, we first compute the cosine similarity between each pair of persons using the extracted ReID features, encoded into a matrix of size $|P_t| \times |P_{t+1}|$. To find suitable matches, we iterate through each row and identify entries that are column-wise and row-wise maximums and surpassing a similarity threshold. The coordinates of the entries are then used to retrieve the matched identities, the trajectories of which are merged.

After merging trajectories of all humans, targeting human-centric objects as motivated before, we mine stable and distinguishable relations between a person and a non-person object, *i.e.*, spatial adjacency, and merge a pair of trajectories if they share the same relation to the same person. Specifically, for a given video and relation, the algorithm first constructs a person-objects dictionary, which links a person p to a set of objects O he/she is "related" to, by examining that such relation holds for enough fraction of the total frames. Then, trajectories of all elements in O for each p are merged. This step depends crucially on the quality of the mined relation. As an example, one relation may be that an object is below and spatially adjacent to the person. Such objects are fairly common and a subset of examples may include a bicycle, a motorcycle, a horse, a cow, or a kayak on which a person is "situated."

4. Experiments

We evaluate the performance of the proposed video segmentation by detection (VSD) by participating in the 2019 Unsupervised DAVIS Challenge [3], and achieve an official second place with very competitive results. The challenge is split into two phases, test-dev and test-challenge, each of which is composed of 30 challenging videos. Test-challenge is the official challenge stage from which final ranks are determined.

Rank	Team	Global Mean	$\mathcal{J}$-Mean	$\mathcal{F}$-Mean
1	RWTH Vision	0.564	0.534	0.594
2	**Oxford-CASIA**	0.562	0.535	0.590
3	SK T-Brain	0.516	0.487	0.545
4	UVOS-test	0.504	0.475	0.533
5	RWTH Vision 2	0.481	0.460	0.503
6	ZX_VIP	0.471	0.435	0.506
7	VIG	0.448	0.422	0.474
8	UOC-UPC-BSC	0.412	0.379	0.444

Table 1. Results in the test-challenge phase.

Rank	Team	Global Mean	$\mathcal{J}$-Mean	$\mathcal{F}$-Mean
1	RWTH Vision	0.585	0.545	0.626
2	**Oxford-CASIA**	0.565	0.517	0.614
3	SK T-Brain	0.542	0.500	0.583
4	RWTH Vision 2	0.517	0.480	0.554
5	ZX_VIP	0.491	0.445	0.537
6	ZX_VIP	0.491	0.443	0.538
7	UVOS-test	0.451	0.413	0.490
8	UOC-UPC-BSC	0.413	0.350	0.476

Table 2. Results in the test-dev phase.

Tables 1 and 2 summarize results from the test-challenge and test-dev phases respectively. As shown in Table 1, our method achieves the highest region similarity $\mathcal{J}$-Mean in all participants, while only falls behind the first place by 0.002 in Global Mean. We note that our model consistently performs well across the two sets, which shows that the proposed model is robust against variances in data distribution. For detailed accounts on the evaluation metrics and the competition, we refer readers to [3].

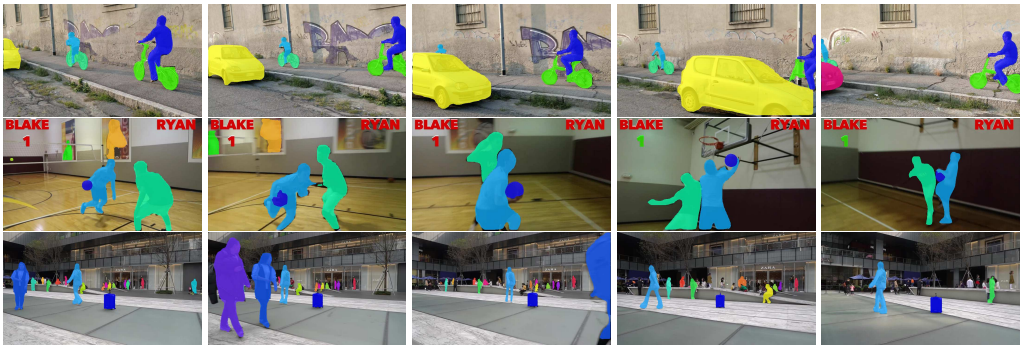

Figure 2. Qualitative analysis on several videos from the test-challenge set. From the top to the bottom are city-ride, dribbling, and luggage.

Model	$\mathcal{J}$	$\triangle_{\mathcal{J}}$	$\mathcal{F}$	$\triangle_{\mathcal{F}}$
VSD	0.517	0.00	0.614	0.00
No merging	0.484	-0.033	0.582	-0.032
No re-init	0.433	-0.084	0.516	-0.098
No SiamMask [13]	0.400	-0.117	0.488	-0.126

Table 3. Ablation study on the test-dev set. $\triangle_{\mathcal{J}}$ and $\triangle_{\mathcal{F}}$ denote, respectively, decrements in region similarity and contour accuracy.

Table 3 shows results of an ablation study on the proposed model. We incrementally remove components from our overall model, corresponding to rows in Table 3 from top to the bottom. As shown, while each step is quite important, re-initialization contributes the most, which highlights the importance and efficacy of re-calibration at each step in defending against online drifts incurred by the trackers.

5. Conclusion

In this work, we present the new framework of video segmentation by detection (VSD) for the automatic segmentation of video objects in an unsupervised manner. The efficacy of VSD is empirically demonstrated by competitive results in the 2019 DAVIS Challenge. As this is a new task, there are many areas for potential improvements within our framework, including the design of a non-human centric refinement module, a more robust identity re-association model, and jointly optimized detector and tracker. We leave them as future work.

Acknowledgements. This work was supported by the EPSRC grant Seebibyte EP/M013774/1, EPSRC/MURI grant EP/N019474/1, and Tencent. We would also like to acknowledge the Royal Academy of Engineering and Five AI.

References

[1] S. Bai, P. Tang, P. H. Torr, and L. J. Latecki. Re-ranking via metric fusion for object retrieval and person re-identification. In *CVPR*, 2019. 2

[2] L. Bo, W. Wei, W. Qiang, Z. Fangyi, X. Junliang, and Y. Junjie. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019. 2

[3] 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, 3

[4] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. Eco: Efficient convolution operators for tracking. In *CVPR*, 2017. 2

[5] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, 2019. 2

[6] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 3

[7] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, 2015. 1

[8] A. Sadeghian, A. Alahi, and S. Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *ICCV*, 2017. 2

[9] M. Siam, C. Jiang, S. Lu, L. Petrich, M. Gamal, M. Elhoseiny, and M. Jagersand. Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting. In *ICRA*, 2019. 2

[10] J. Supancic III and D. Ramanan. Tracking as online decision-making: Learning a policy from streaming videos with reinforcement learning. In *ICCV*, Oct 2017. 2

[11] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple people tracking by lifted multicut and person re-identification. In *CVPR*, 2017. 2

[12] 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. 2

[13] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019. 2, 3, 4

[14] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 2

[15] Y. Xiang, A. Alahi, and S. Savarese. Learning to track: Online multi-object tracking by decision making. In *ICCV*, 2015. 2

[16] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008. 1, 2

[17] Z. Zhu, Q. Wang, L. Bo, W. Wu, J. Yan, and W. Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, 2018. 2