An Empirical Study of Detection-Based Video Instance Segmentation

Qiang Wang[1], Yi He[2], Xiaoyun Yang[2], Zhao Yang[3], Philip H.S. Torr[3]
[1]CASIA [2]Intellimind Ltd [3]University of Oxford
qiang.wang@nlpr.ia.ac.cn yi.he@intellimind.ai xiaoyun.yang@intellimind.ai
zhao.yang@eng.ox.ac.uk philip.torr@eng.ox.ac.uk

Abstract

Video instance segmentation (VIS) is a composite task that requires the joint detection, tracking, and segmentation of objects in a video. In this work, we introduce a complete framework for VIS, which integrates the strengths of instance segmentation and general object tracking in addressing the unique challenges of VIS. In developing the framework, we investigate effective ways of coordinating the two components for maximum benefits while thoroughly investigate their separate contributions. Our approach improves over the official baseline by an absolute 14.4% *in mAP and achieves the second place in the 2019 YouTube-VIS challenge.*

1. Introduction

In this work we consider the recently proposed task of video instance segmentation [24], which aims at the simultaneous detection, tracking, and segmentation of objects present in a video. Given a video that records a set of moving objects, an algorithm must predict the category of each object while precisely delineate the object at each frame. Under such formulation, the sequence of instances of the same object across a video's frames is considered as a *video instance*, a counterpart of the notion of *instance* from the image domain. Therefore when tackling video instance segmentation, it is sensible to exploit the existing results in image instance segmentation, or, object detection. The crucial challenge remains of designing a system that can effectively establish consistent object tracks using image-level detections and resolve any classification inconsistencies that arise along the track of an object.

Thus in designing our solution, first and foremost we aim to obtain strong object detection results, which we achieve by employing state-of-the-art object detection models and data augmentation techniques. In order to associate individual detections from each frame into consistent object tracks, we draw inspirations from a common methodology in multiple-object tracking, tracking-by-detection, and adopt a Siamese object tracker for establishing object correspondences across frames. Due to challenges such as deformation and occlusion, object detection may fail to classify an object consistently during the course of its track, causing ambiguity when predicting its category. To address this issue, we develop a re-classification method for object tracks by leveraging an image classifier, which significantly improves the performance of our overall model.

2. Related Work

Video instance segmentation [24] overlaps with many existing tasks in the areas of detection, tracking, and segmentation of objects in videos. In this section, we provide a brief overview on each of the related tasks.

Visual object tracking [14] aims to localize an object captured in a video through its sequence of frames, with object's initial location given in the first frame. Recent advances in this area include the Siamese networks [3] and correlation filter-based methods [10, 22]. In this work, we exploit the efficiency of the light and robust Siamese tracker, SiamMask [23], for fast tracking and accurate segmentation of objects in videos.

Multiple-object tracking [15] aims to localize an arbitrary number of objects throughout a video. As no initialization of object locations is given at inference time, most methods employ object detection algorithms for target object proposals. The top-performing class of tracking-by-detection methods [21, 28] link individual detections from each frame into object tracks via association.

Video object segmentation [19] seeks to delineate the objects present in a video and falls under one of supervised, semi-supervised, and unsupervised settings depending on the amount of ground-truth annotations available at inference time. Many recent advances have been made in this challenging area (*e.g*., [1, 5, 17, 18, 25, 26]).

Object detection [12] algorithms locate and classify objects in images, and often form the basis of state-of-the-art solutions in the aforementioned tasks. In our work, we take advantage of the newest developments by employing the state-of-the-art multi-stage object detector, HTC [8], which

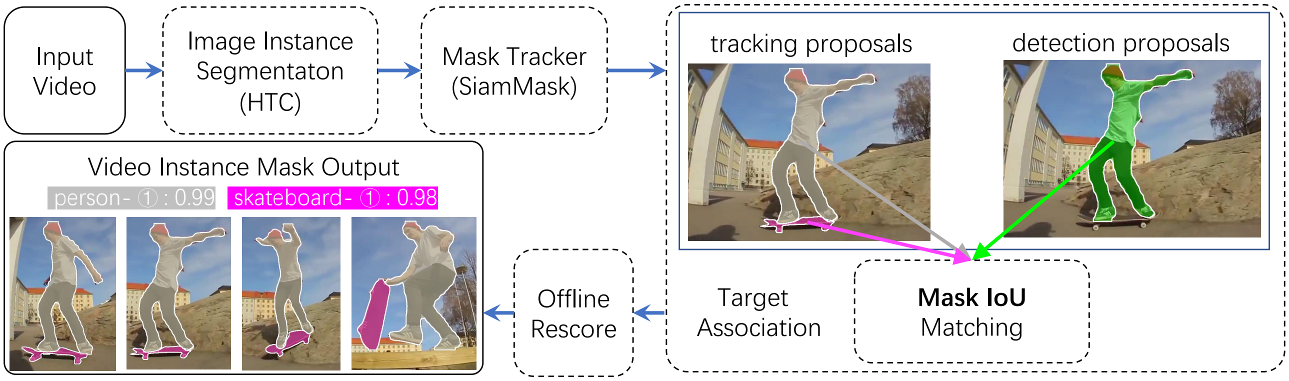

Figure 1: The proposed framework for video instance segmentation.

advances Mask R-CNN [13] and Cascade R-CNN [7] with a novel cascade structure of top-layer heads, and applying the recently proposed data augmentation technique, InstaBoost [11], to further improve our detection results.

3. Method

Our method draws inspirations from the paradigm of tracking-by-detection [28] for multiple-object tracking [15] and is an improvement over the recent proposal of video segmentation-by-detection [25] for unsupervised video object segmentation [6]. Figure 1 schematically illustrates our proposal. Given a new frame with detection masks from an object detector, an association module computes similarity scores between existing tracklets (maintained by SiamMask [23]) and these newly detected objects. By repeating this step at each frame, we obtain a set of object tracks. When classifying each object track (hence the object), we develop a special re-classification algorithm for improving the classification accuracy. Finally, our algorithm outputs the set of object tracks in masks, with a corresponding class label for each object.

3.1. Proposal generation

It is shown that improving detection accuracy significantly improves the performance of a multiple-object tracking algorithm [27]. As we also use detection results as target proposals, we employ a set of strategies to improve the quality of detections, including several off-the-shelf tricks, which are discussed below.

Better Detector. Instead of using Mask R-CNN for proposal generation, as is done in the offical baseline [24], we adopt HTC [8], with a ResNeXt101-DCN [9] backbone. As shown in Table 1, the quality of the detector contributes greatly to the final performance.

Auxiliary training data. While YouTube-VIS contains a total of $131{,}000$ object masks, the number of unique objects is $4{,}883$, averaging 122 objects for each category. In order to increase data diversity, we expand our training data to include a subset of COCO [16] and OpenImage [2] data sets.

Data augmentation. In addition to conventional data augmentation strategies (*e.g*., random scaling and cropping), we adopt InstaBoost [11] to further improve data efficiency.

Mirrored input. Rotating, mirroring, and multi-scaling of the input are some standard practices applied at test time for object detection. We find that only mirrored input improves the detection performance on YouTube-VIS, which may be attributed to the generally larger size of its objects.

3.2. Mask propagation

After proposal generation, we must associate the proposals of the same object across different frames into an object track. While many good choices of visual object trackers and video object segmentation models are available for establishing object correspondences across frames, we choose SiamMask for its efficiency and robustness agaist noisy initializations. In addition, compared to trackers based on motion models such as the Kalman filter [4], SiamMask provides more accurate state estimations, with a tracking score that can indicate when a target goes out of view and a segmentation mask that can substitute for a missing detection.

Our modified version of SiamMask is illustrated in Figure 2, which is composed of two original SiamMask models stacked in a cascade manner, which we refer to as stages one and two. In the first stage, the box branch of SiamMask generates an initial proposal of the object's location, which is used in the second stage for predicting a refined mask of the object. This tracking mask is used to find the object proposal that should be associated with the current object track. Specifically, we compute the IoU between this tracking mask and each new detection mask, and threshold at 0.7 when the predicted categories of the two are the same and 0.4 otherwise.

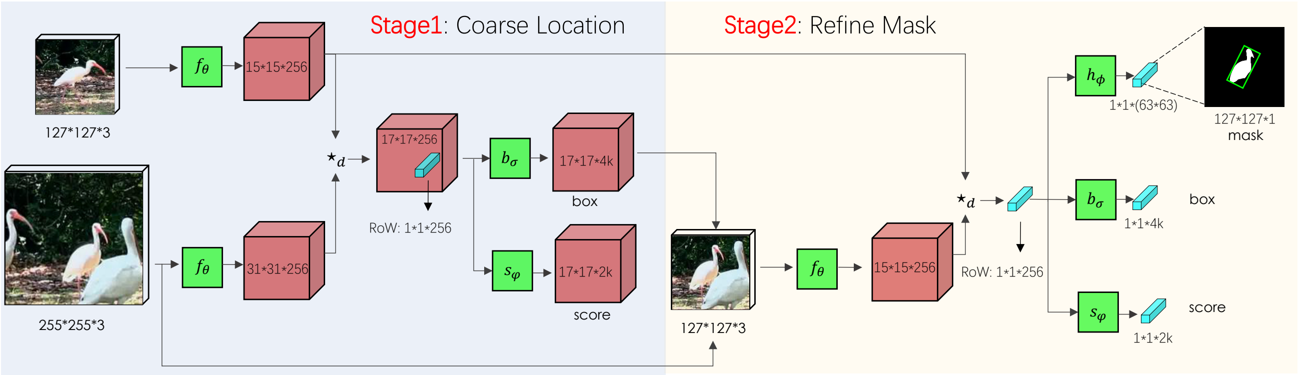

Figure 2: A modified two-stage variant of SiamMask, in which we stack two original SiamMask models in a cascade manner to provide more accurate localization and segmentation.

3.3. Re-classification of object tracks

After obtaining the set of object tracks, we still need to assign a class label to each of them, which represents the category of the object. A straightforward approach is to average the class probabilities from detections along the track and assign the label of the highest probability. Adopting this approach, however, we find that around 10% of the objects from the validation set are misclassified. And if an object is misclassified, even a perfect segmentation track still counts zero in the evaluation metric. In our case, we find that the classification accuracy of object tracks becomes the bottleneck in the whole pipeline.

Therefore, we introduce an offline post-processing step to predict the final class label of an object track. Specifically, we employ a state-of-the-art image classifer, HRNet [20], and feed it cropped image patches of the object along the track, to obtain a new track of classification probabilities for this object. We then average the probabilities across the track and assign the object the label of the highest probability. This step is only performed when the object is larger than a size threshold (set as 100 pixels) and the predictions along the track have been inconsistent.

4. Experiments

The YouTube-VIS data set comprises 2,883 high-resolution videos with annotated objects in 40 categories and a total of 131,000 masks. The evaluation metrics are average precision (AP) and average recall (AR). We refer readers to [24] for details.

To analyze quantitatively the importance of each of the components in our framework, we provide evaluation results after components ablation in Table 1. We can see that object detection and our re-classification strategy both play a vital role in achieving good performance, as we lose 9.4 and 4.4 absolute points in mAP without each. The contribution from auxiliary training data is significant as performance decreases by 4.3% without it. In addition, mirrored input, InstaBoost, and the improved structure of the tracker each contributes reasonbly to the final performance.

Model	mAP	$\triangle_{mAP}$	R10	$\triangle_{R10}$
MaskRCNN-R50-FPN	0.272	0.0	0.304	0.0
HTC-X101-DCN	0.310	+0.038	0.381	+0.077
+ COCO & OpenImage	0.353	+0.043	0.426	+0.045
+ Mirror + InstaBoost	0.366	+0.013	0.423	-0.003
+ two stage SiamMask	0.381	+0.015	0.439	+0.016
+ Re-classification	**0.425**	+0.044	**0.478**	+0.039

Table 1: Ablation studies on the validation set of YouTube-VIS. $\triangle_{mAP}$ and $\triangle_{AR10}$ denote, respectively, absolute improvements in mAP and AR@10.

Team	mAP	AP50	AP75	R1	R10
Jono	0.467	0.697	0.509	0.462	0.537
foolwood	**0.457**	**0.674**	**0.490**	**0.435**	**0.507**
bellejuillet	0.450	0.636	0.502	0.447	0.503
linhj	0.449	0.665	0.486	0.453	0.538
mingmingdiii	0.444	0.684	0.487	0.436	0.508
baseline [24]	0.313	0.503	0.338	0.335	0.369

Table 2: Performance comparison of different methods on the test set of YouTube-VIS. Our results are in bold.

Table 2 shows comparison with the state of the art with final results on the test set. In Figure 3, we further demonstrate visualizations of segmentation results on some challenging videos, which cover objects in different scales and undergoing large deformation or in fast motion.

5. Conclusion

In this paper, we present an integrated framework for video instance segmentation. With detailed analyses, we show how, existing research efforts for object detection and visual object tracking can be effectively utilized in this new

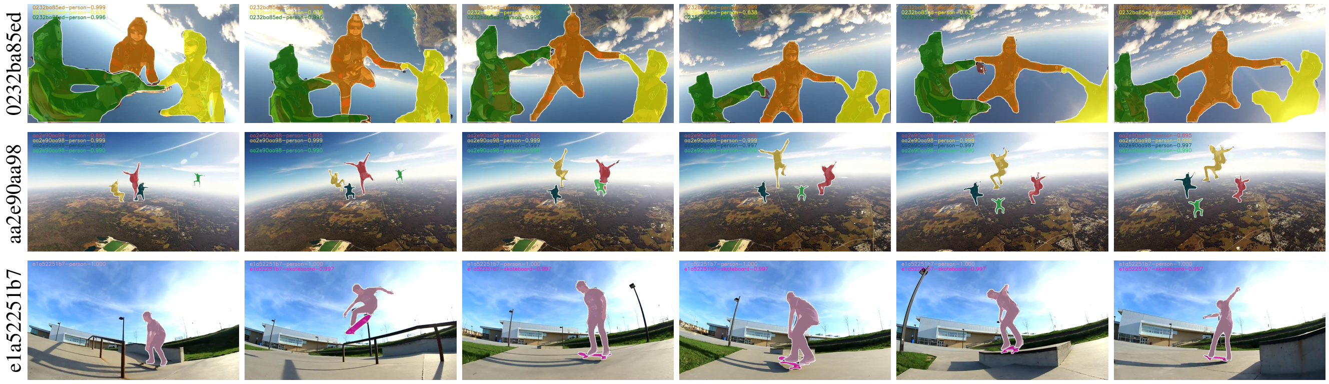

Figure 3: Qualitative results of our method on videos in Youtube-VIS.

task, and highlight contributions from a set of very practical techniques. Our overall system achieves state-of-the-art performance on the challenging data set of YouTube-VIS.

Acknowledgements. This work was supported by the ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1, EPSRC/MURI grant EP/N019474/1, Tencent, and Intellimind Ltd.

References

[1] Harkirat Singh Behl, Mohammad Najafi, Anurag Arnab, and Philip H. S. Torr. Meta learning deep visual words for fast video object segmentation. *arXiv:1812.01397*, 2018. 1

[2] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. *arXiv:1903.10830*, 2019. 2

[3] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, 2016. 1

[4] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *ICIP*, 2016. 2

[5] Ning Wang Shunfei Wang Xiaofeng Zhang Shaoli Liu Si Gao Kaidi Lu Diankai Zhang Lin Shen Yukang Wang Yongchao Xu Bofei Wang, Chengjian Zheng. Object-based spatial similarity for semi-supervised video object segmentation. *CVPR Workshops*, 2019. 1

[6] 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. 2

[7] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 2

[8] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 1, 2

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *arXiv:1703.06211*, 2017. 2

[10] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. Eco: Efficient convolution operators for tracking. In *CVPR*, 2017. 1

[11] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. *arXiv:1908.07801*, 2019. 2

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[13] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[14] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *TPAMI*, 2016. 1

[15] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942*, 2015. 1, 2

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[17] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, 2019. 1

[18] Kevis-Kokitsi Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *TPAMI*, 2018. 1

[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. 1

[20] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3

[21] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple people tracking by lifted multicut and person re-identification. In *CVPR*, 2017. 1

[22] Jack Valmadre, Luca Bertinetto, Joao Henriques, Andrea Vedaldi, and Philip H. S. Torr. End-to-end representation learning for correlation filter based tracking. In *CVPR*, 2017. 1

[23] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019. 1, 2

[24] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. *arXiv:1905.04804*, 2019. 1, 2, 3

[25] Zhao Yang, Qiang Wang, Song Bai, Weiming Hu, and Philip H.S. Torr. Video segmentation by detection for the 2019 unsupervised davis challenge. In *CVPR Workshops*, 2019. 1, 2

[26] Zhao Yang, Qiang Wang, Luca Bertinetto, Song Bai, Weiming Hu, and Philip H.S. Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, 2019. 1

[27] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan. Poi: Multiple object tracking with high performance detection and appearance feature. In *ECCV*, 2016. 2

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