Dual Embedding Learning for Video Instance Segmentation

Qianyu Feng, Zongxin Yang, Peike Li, Yunchao Wei, Yi Yang
ReLER, Centre for Artificial Intelligence, University of Technology Sydney
{qianyu.feng, zongxin.yang, peike.li}@student.uts.edu.au, {yunchao.wei, yi.yang}@uts.edu.au

Abstract

In this paper, we propose a novel framework to generate high-quality segmentation results in a two-stage style, aiming at video instance segmentation task which requires simultaneous detection, segmentation and tracking of instances. To address this multi-task efficiently, we opt to first select high-quality detection proposals in each frame. The categories of the proposals are calibrated with the global context of video. Then, each selected proposal is extended temporally by a bi-directional Instance-Pixel Dual-Tracker (IPDT) which synchronizes the tracking on both instance-level and pixel-level. The instance-level module concentrates on distinguishing the target instance from other objects while the pixel-level module focuses more on the local feature of the instance. Our proposed method achieved a competitive result of mAP 45.0% on the Youtube-VOS dataset, ranking the 3rd in Track 2 of the 2nd Large-scale Video Object Segmentation Challenge.

1. Introduction

Video object segmentation is a fundamental task in computer vision, which extends image object segmentation into video aiming to assign each frame pixel with a foreground category or background. Recently, video instance segmentation proposed by Yang et al. [14] sets a more challenging task. Beyond the challenges existing in the semi-supervised video object segmentation, the video instance segmentation brings some new ones, including 1) the ground-truth masks of any instances or any other side-information are not given at the beginning; 2) the low quality of videos caused by various factors such as low resolution and motion blur often prevents the new instances from being detected with the state of the art detectors [9, 6]; 3) the intersection of two independent instances or unexpected occlusion will further increase the confusion of the model, especially for the case when one instance is disappearing or reappearing; 4) the identification of new instances during the tracking of existed ones.

To tackle the above-mentioned challenges, we propose a novel approach with a bi-directional tracker, named Instance-Pixel Dual-Tracker (IPDT), on the top of detectors. The proposed IPDT devotes to the tracking of the selected object candidates temporally together with the instance-level and the pixel-level embedding learning. First, each detected proposal is regarded equally as an object candidate in the video. With the calibration of the predicted category, the key instance candidates can be better filtered to be targeted on. Second, with high-quality detected masks for reference, the instances can be better tracked in frames when the instances are blurry or occluded. Otherwise, the tracking results will be inferior if the quality of the reference mask is of low-quality. Third, from the instance level, the proposed IPDT can better capture the representations of the target instance and its context from the instance granularity. Simultaneously, IPDT also tracks the target instance on the pixel-level which focuses more on the local representation of the local parts of the instance. Our proposed IPDT achieved mAP of 45.0% in the track 2 of the 2nd Large-Scale Video Object Segmentation Challenge.

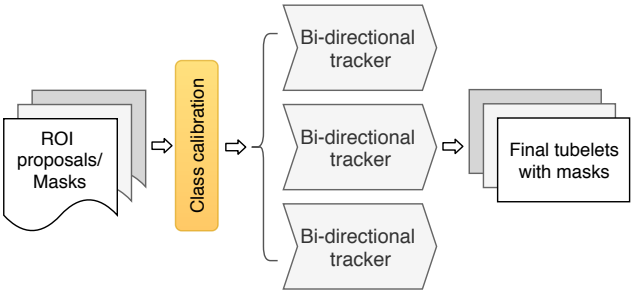

Figure 1. An overview of our framework. The proposals generated by Mask R-CNN [6] are first calibrated on the class prediction and then fed into the proposed bi-directional tracker IPDT. The final tubelets are filtered with post-processing.

2. Related Work

Video Object Tracking. The task of video object tracking employs a rectangular bounding box as an initialise target template to estimate its position in the subsequent frames. Recently some fully-convolutional siamese ap-

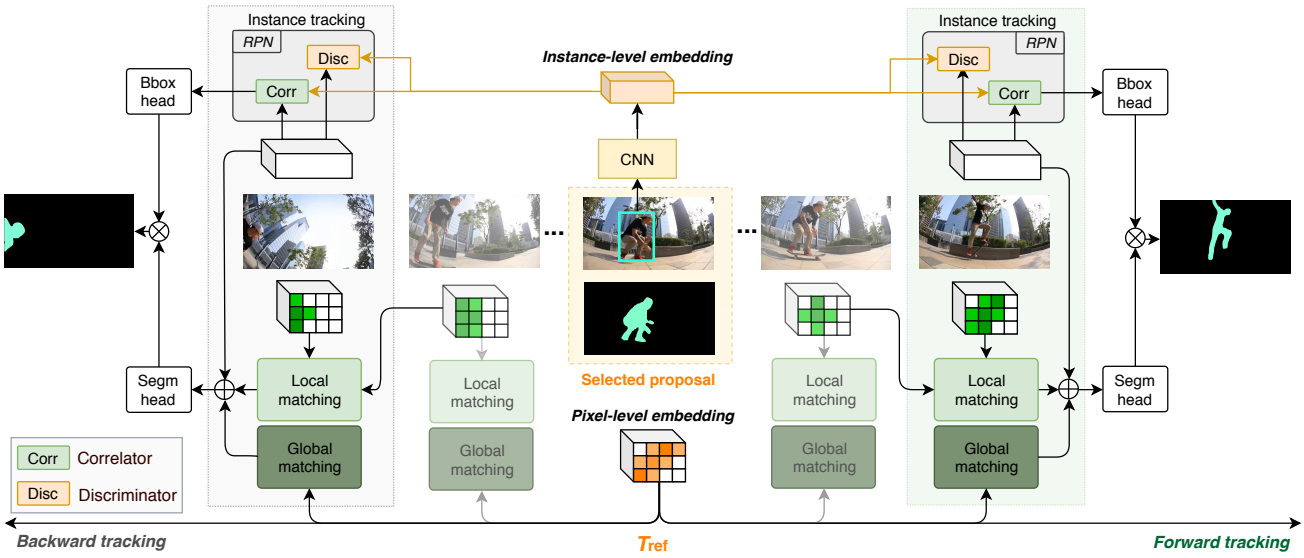

Figure 2. Details of the bi-directional Instance-Pixel Dual-Tracker (IPDT).

proaches [1, 5, 7, 16] learn a similarity function on pairs of video frames. Inspired by these work, we track the target object from instance-level with a siamese-style network.

Video Object Segmentation. Most approaches [2, 11, 8] for video object segmentation task rely on fine-tuning the first-frame ground truth. Recently some other works [4, 10] avoid the fine-tuning process to achieve better run-time performance. Our work is mostly inspired by FEELVOS [10], where they learned a pixel-wise embedding and employed a matching mechanism to produce accurate segmentation. However, the pixel-wise embedding tends to be hard to discriminate the instance from each other. In our approach, we combine both the pixel-level and the instance-level embedding feature to produce more accurate instance segmentation results.

3. Approach

3.1. Framework Overview

The main framework of our solution is illustrated in Fig. 1. We propose a novel bi-directional tracker named IPDT, which is built upon the state of the art instance segmentation model, *i.e.*, Mask R-CNN [6]. First, we use Mask R-CNN to harvest a set of object candidates. The category of each candidate is calibrated by the global context of the entire video to avoid false positives. Since there are many proposals related to the same object across different frames, we further filter the reluctant ones to reduce the computation cost. Then, we apply the proposed IPDT to extend the selected proposals in both forward & backward directions. The tracker not only aims to locate the detected instance in the adjacent frames on the instance level, but also learn the local embedding from the pixel level inspired by [10]. We demonstrate the details of our IPDT in Fig. 2.

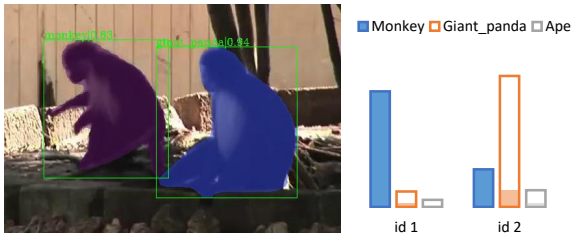

Figure 3. An example of class calibration. In this video, "monkey" is a reliable category while "giant panda", "ape" are not reliable based on the average scores of all the object candidates in the video. Thus, the detection scores of the unreliable classes ("giant panda", "ape") are re-weighted with factor β.

3.2. Class Calibration

To better exploit the video context, the categories of object candidates are calibrated base on the fact that most objects in a video appear more than one single frame. We first calculate the average score of each class over all the object candidates. Classes with average score higher than a pre-defined threshold th_{cc} are regarded as reliable categories which show up in the video. The other classes are with lower confidences and the score probabilities will be re-weighted with a factor β, which can significantly reduce a large number of false positives. An example is shown in Fig. 3. After the class calibration, we filter the object candidates with overlaps spatially and temporally. That is to say, if the IoU between two proposals is higher than a threshold th_{IoU}, they are regarded as the same instance and the candidate with higher detection score will be selected.

3.3. Bi-directional Instance-Pixel Dual-Tracker

Instance-level Embedding Learning. We first extract the RoI feature of the selected instance for reference. The instance-level tracking module adopts a region proposal network (RPN) [9] to generate proposals in a Siamese-style and aims to track the reference instance with bounding boxes. The Siamese network consists of two branches: 1) a discriminator to predict if the RPN proposal is the same target with the reference instance; 2) a correlator to refine the instance proposal. The extracted feature of the reference instance and the current frame are fed into the mentioned two branches: a discriminator and a correlator. The discriminator first embeds the instance-level features with convolution layers. Then the embedded feature of the reference instance is used to calculate the correlation with the embedded feature of the current frame. A classification layer is cascaded to predict whether the RPN proposal is similar to the reference instance. In the meanwhile, the correlator also calculates the correlation between the reference instance with the current frame and adopts a regression layer to predict the refined bounding box of the object candidate.

Pixel-level Embedding Learning. The pixel-level tracking module is inspired by FEELVOS [10]. There exist two types of embedding matching in this module: one matches the pixel embedding in the current frame globally with the reference instance, the other matches the current pixel embedding with the instance prediction in the previous frame. More detailly, the extracted features are fed into an embedding layer with the same stride. For each reference instance, a distance map by globally matching the embedding vectors of the current frame to the embedding vectors of the instance in the reference object candidate. Simultaneously, the predictions of the previous frame are adopted to compute the local distance map by matching the current embeddings vectors of that in the previous frame. The predictions of the previous frame are also used as auxiliary information as MaskTrack [3]. Finally, the distance maps of the local matching and global matching, the previous frame prediction and the backbone features are concatenated together and fed into a dynamic segmentation head to predict the probability distribution over all the pixels of the reference instance. More details could be found in [15].

Based on the dual-embedding learning and tracking, the instance-level box and pixel-level mask can be obtained in the current frame. We multiply the predicted mask with the box (the pixel value is 1 inside the box). After the bi-directional tracking with IPDT, each selected object candidate is extended to a tubelet with masks. Then we calculate the IoU of any two of the tubelets, the tubelet with a higher score is selected as a final output result if the IoU averaged over the frame number surpasses a threshold th_{seq}. The most frequent class is selected to be the final class of the tubelet and the final score is calculated with the average of re-weighted scores over all the frames with valid segmentation results.

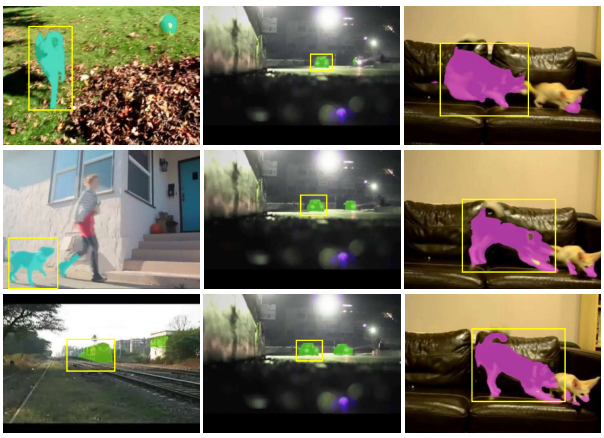

Figure 4. Some failure cases with only pixel-level tracking.

Figure 5. A failure example where objects are largely overlapped.

4. Experiments

Implement details. We first finetune the Mask R-CNN with Resnext-101 [13] as backbone on the train split of Youtube-VOS dataset. Empirically, we set the detection score threshold as 0.2, cause some categories are hard to detect in a dusky environment, *e.g.*, torture, lizard. We choose the re-weight factor β as 0.1 in the experiments. The parameters of RPN network in the instance-level tracking follows the implementation in SiamRPN [16]. DeepLabv3+ [3] is adopted as backbone to extract features with a stride of 4.

As shown in Table 1, we obtain very competitive results. Our proposed IPDT achieved 45.0, 63.6, 50.2, 44.7, 50.3 in terms of mAP, AP50, AP75, AR1 and AR10, respectively. Furthermore, the contribution of our proposed IPDT as shown in Table 2. Only with the instance-level tracking, DeepSORT [12] is used to link the detection results of Mask R-CNN, which could only reach mAP 25.7. Only tracking on pixel-level based on FEELVOS [10], mAP score reaches 37.6. Some failure cases are shown in Fig. 4, which reveal that the pixel-level tracker lacks the concept at instance-level. However, it may fail when the objects are largely overlapped, *e.g.*, as the case in Fig. 5. Particularly, our proposed IPDT improves the mAP to 45.0 which surpasses pixel-level tracking by 7.4. Our IPDT can successfully track and segment the key instances.

#	Team	mAP	AP50	AP75	AR1	AR10
1	Jono	46.7 (1)	69.7 (1)	50.9 (1)	46.2 (1)	53.7 (2)
2	foolwood	45.7 (2)	67.4 (3)	49.0 (3)	43.5 (5)	50.7 (4)
3	**ReLER_VIS (ours)**	**45.0 (3)**	**63.6 (5)**	**50.2 (2)**	**44.7 (3)**	**50.3 (5)**
4	linhj	44.9 (4)	66.5 (4)	48.6 (5)	45.3 (2)	53.8 (1)
5	mingmingdiii	44.4 (5)	68.4 (2)	48.7 (4)	43.6 (4)	50.8 (3)
6	xiAaonice	40.0 (6)	57.8 (9)	44.9 (6)	39.6 (9)	45.2 (9)
7	guwop	40.0 (7)	60.8 (7)	43.9 (8)	41.2 (7)	49.1 (6)

Table 1. Ranking results in the 2nd Large-scale Video Object Segmentation Challenge - Track 2.

Method	mAP	AP50	AP75	AR1	AR10
DeepSORT [12]	25.7	39.2	27.2	26.4	30.6
FEELVOS [10]	37.6	54.7	41.2	36.8	42.4
Ours	**45.0**	**63.6**	**50.2**	**44.7**	**50.3**

Table 2. The performance of our methods in the challenge. DeepSORT is used to simply track the detection results of Mask R-CNN. FEELVOS tracks the selected object candidates on pixel-level.

5. Conclusion and Future Work

In this paper, we propose a novel approach IPDT for video instance segmentation task. Throughout the experiment results, our proposed approach is proved to be efficient and achieves a competitive result among the leading rank submissions. In the future, we plan to incorporate instance-level tracking to generate mask directly and also embed the Mask R-CNN head into the framework to be able to be trained in an end-to-end manner.

References

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision (ECCV)*, pages 850–865, 2016.

[2] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 221–230, 2017.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[4] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7415–7424, 2018.

[5] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3038–3046, 2017.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[7] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8971–8980, 2018.

[8] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *European Conference on Computer Vision (ECCV)*, pages 90–105, 2018.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[10] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9481–9490, 2019.

[11] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[12] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, 2017.

[13] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500, 2017.

[14] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[15] Zongxin Yang, Peike Li, Qianyu Feng, Yunchao Wei, and Yi Yang. Going deeper into embedding learning for video object segmentation. In *IEEE International Conference on Computer Vision Workshop*, 2019.

[16] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *European Conference on Computer Vision (ECCV)*, pages 101–117, 2018.