Decoupling Video Instance Segmentation into two sub-tasks based on Propose-Reduce paradigm

Feng Zhu[1,2] Zongxin Yang[1,2] Wenguan Wang[3] Yunchao Wei[1] Yi Yang[1]
[1]University of Technology Sydney
[2]Baidu Research
[3]ETH Zurich
{v_zhufeng01}@baidu.com

Abstract

Video Instance Segmentation (VIS) is a complex task, which consists of detecting, segmenting and tracking objects in a video. In this work, we introduce a new pipeline for VIS task, which decouples VIS into two sub-tasks including image instance segmentation and semi-supervised video object segmentation based on a Propose-Reduce paradigm. We first perform image level instance segmentation on each frame of a video and then select multiple key frame to conduct video object segmentation in order to generate different video instance segmentation proposals. Finally, we reduce the redundant proposals to obtain the final result. This strategy could benefit from two well studied tasks, and achieve competitive results in the 2021 YouTube VIS challenge.

1. Introduction

Video instance segmentation(VIS) [14] has recently attracted great attention as a new task of video understanding. This task aims to segment multiple object instances of the predefined categories in a video. Different from image-level instance segmentation, VIS is more challenge as it further requires tracking objects across frames in addition to frame-level object detection and segmentation. On the other hand, videos contains temporal information compared to image-level instance segmentation, which is beneficial for object detection and segmentation.

In recent studies, various frameworks and methods are proposed to tackle this problem, and these frameworks could be categorized into three paradigms according to [7]:1) 'Track-by-Detect', 2)'Clip-Match', 3)'Propose-Reduce'. The Propose-Reduce framework has shown better performance compared with other two frameworks. Therefore, we adopt the Propose-Reduce framework as the basic framework of our method. However, different from Seq Mask R-CNN proposed in [7], we split the VIS task into two sub-tasks: image-level instance segmentation and semi-supervised video object segmentation [11, 12]. We first acquire strong instance segmentation results by utilizing state-of-the-art instance segmentation models and data techniques, and then select several key frames to perform video object segmentation to produce sequence proposals using state-of-the-art video object segmentation methods, finally we apply the NMS algorithm to the instance proposals to achieve the final result.

2. Related Work

Video instance segmentation is highly related to several tasks such as image instance segmentation and semi-supervised video object segmentation. In this section, we provide a brief overview of recent studies in video instance segmentation and related fields.

Image Instance Segmentation The classical two stage architecture of Mask R-CNN[5] is predominate in image instance segmentation field. Cascade R-CNN[2] propose a multi-stage detection architecture to address the problem of training object proposals with low quality. Hybrid Task Cascade[4] integrates semantic features to further improve the performance of instance segmentation. DetectoRS[13] propose Recursive Feature Pyramid and Switchable Atrous Convolution to enhance instance segmentation significantly. Besides, various strong backbone networks are proposed to boost object detection and instance segmentation. ResNest[17] presents a modularized architecture that combines the channel-wise attention strategy with multipath network layout. Swin-Transformer[9] proposes a hierarchical transformer with Shifted Window based Self-Attention.

Semi-supervised video Object Segmentation Semi-supervised video object segmentation[11, 12] targets at segmenting given objects with the annotated first frame. Many previous approaches for semi-supervised video object segmentation task rely on fine-tuning the first frame at test time.

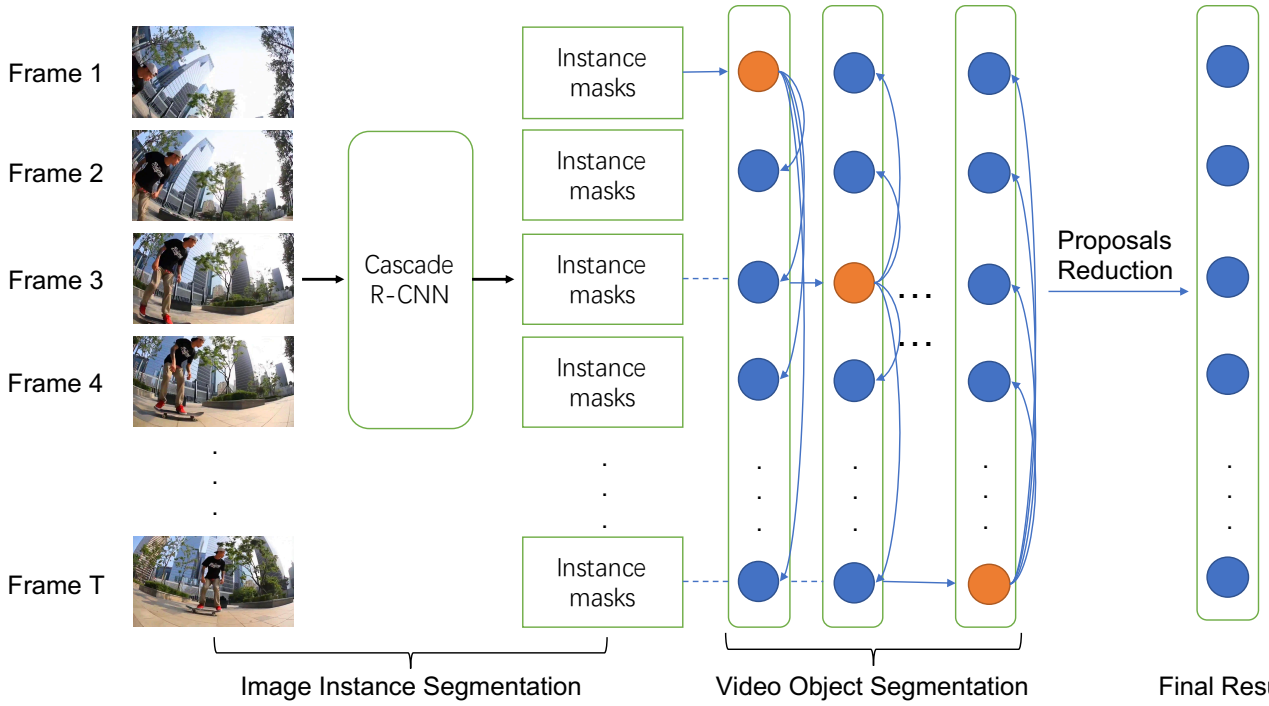

Figure 1. The framework of our method.

Some recent works propose methods without fine-tuning to achieve a better run-time. STM[10] leverages a memory network to perform long-term propagation. CFBI[15, 16] utilizes the feature embedding from the target foreground object and its corresponding background collaboratively.

Video Instance Segmentation There are three paradigms to solve the video instance segmentation problem:1) 'Track-by-Detect'; 2) 'Clip-Match'; 3) 'Propose-Reduce'. 'Track-by-Detect' paradigm directly adds a track head to to existing image instance segmentation model, such as Mask Track R-CNN[14] and SipMask[3]. 'Clip-Match' paradigm first splits the whole video into several overlapped clips. With clips, it generates multiple incomplete sequences via mask propagation and merges them by matching. MaskProp[1] achieves high performance using this paradigm. 'Propose-Reduce' paradigm[7] selects multiple frames as key frames to generate multiple sequence proposals for the whole video and then conducts sequence proposals reduction using NMS method with classification score and IoU.

3. Method

Inspired by the paradigm of Propose-Reduce, our method consists of two parts: 1) sequence proposals generation; 2) sequence proposals reduction. Considering efficiency and convenience, we further divide the sequence proposals generation process into image instance segmentation and semi-supervised video object segmentation. Thus, our method contains three parts:1)image instance segmentation; 2) semi-supervised video object segmentation(VOS); 3) sequence proposals reduction. The framework of our proposal is illustrated in Figure1. Given a video, we first use image instance segmentation model to obtain the mask of objects in each frame. After getting the segmentation result of each frame, we select several key frame and produce multiple sequence proposals via VOS model based on the prediction of each key frame separately. Finally, we utilize the traditional NMS method to reduce the redundant sequence proposals and merge them into the final result for VIS.

3.1. Image instance segmentation

The performance of image instance segmentation is crucial in our method as it is the basis of following stages. Thus, we adopt a series of strategies to improve the quality of image-level instance segmentation.

Extra training data The YouTube-VIS dataset consists of 3,859 high-resolution YouTube videos, a 40-category label set and 232k high-quality instance masks. In order to improve data diversity, we create a new training dataset combining YouTube-VIS, COCO[8] and Open Images[6] datasets. We obtain the subset of COCO and Open Images from classes which overlap with the YouTube-VIS classes. Besides, we exclude part of samples from the subset of COCO and Open Images to improve class balance and training efficiency, because there are too many samples of human and car in Open Images and COCO.

Table 1. Final ranking results on the 3rd Large-scale Video Object Segmentation Challenge - Track 2: Video Instance Segmentation.

#	Team	mAP	AP50	AP75	AR1	AR10
1	tuantng	0.541 (1)	0.742 (2)	0.616 (1)	0.433 (2)	0.589 (2)
2	eastonssy	0.523 (2)	0.767 (1)	0.577 (2)	0.439 (1)	0.570 (3)
3	vidit98	0.491 (3)	0.681 (5)	0.545 (3)	0.410 (5)	0.550 (4)
4	linhj	0.478 (4)	0.693 (3)	0.527 (5)	0.422 (3)	0.591 (1)
5	hongsong.wang	0.476 (5)	0.684 (4)	0.529 (4)	0.414 (4)	0.546 (5)
6	gb7	0.473 (6)	0.665 (6)	0.511 (7)	0.405 (6)	0.516 (7)
7	**zfonemore**	**0.461 (7)**	**0.644 (8)**	**0.510 (8)**	**0.383 (8)**	**0.506 (8)**
8	DeepBlueAI	0.460 (8)	0.646 (7)	0.520 (6)	0.387 (7)	0.542 (6)

Data augmentation We employ conventional data augmentation techniques including random scaling and flip.

Better model We adopt the Cascade Mask R-CNN with strong backbone including ResNest200 and Swin Transformer. Following [7], our training process consists of two stages:main-training and fine-tuning. In the main-training, we train the COCO pre-trained model for 12 epochs on the mixed training dataset. In the fine-tuning, the model is further trained on the YouTube-VIS dataset for 4 epochs.

3.2. Semi-supervised video object segmentation

After image-level instance segmentation, we uniformly select K key frames. With the image-level instance segmentation masks of each key frame, we utilize the VOS model CFBI to propagate mask from key frame to other frames in a video. After K times propagation, we could get K video instance segmentation proposals.

3.3. Instance proposals reduction

After the first two stages, we obtain redundant instance proposals because the same instance may be detected in different key frames. To reduce the redundant proposals, we apply the NMS method to the proposals with the classification score and IoU and the reduction process is the same as [7]. After getting non-repetitive instance proposals, we further associate the instances in different frames with sequences IoU defined in [7].

4. Experiments

Implementation details. We choose Cascade R-CNN with a strong backbone ResNest200 as the image instance segmentation model. For the VOS task, we adopt the state-of-the-art VOS model CFBI with ResNet101-DeepLabV3+ as the backbone. The input size is downsampled to 640x360 for inference. The detection score threshold is set to 0.3 and the IoU threshold is set to 0.5. The number of key frames is set to 6.

Results on the YouTube-Video Instance Segmentation challenge. Our proposed approach achieves competitive performance on the 2nd YouTube-VIS challenge. Table 1 shows the final ranking results on the YouTube-Video Instance Segmentation challenge test set. Our method achieve a 0.461 mAP which ranks the 7th in the final leaderboard.

5. Conclusions

In this paper, we decouple the original VIS task into two existing sub-tasks including image instance segmentation and VOS. This strategy makes the VIS task benefit from the progress of two existing tasks. In the future, we will explore more flexible framework to combine these two sub-tasks in a end-to-end manner.

References

[1] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 1

[3] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. *Proc. European Conference on Computer Vision*, 2020. 2

[4] 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, 2019. 1

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 1

[6] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 2

[7] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm, 2021. 1, 2, 3

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin trans-

former: Hierarchical vision transformer using shifted windows, 2021. 1

[10] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[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. 1

[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. 1

[13] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution, 2020. 1

[14] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 2

[15] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, pages 332–348. Springer, 2020. 2

[16] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2

[17] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 1