Limited Sampling Reference Frame for MaskTrack R-CNN

Zhuang Li [1,2*] Leilei Cao[1] Hongbin Wang[1]
[1]Ant Group
[2]College of Electronic and Information Engineering, Tongji University

Abstract

With the great achievement for the computer vision tasks, e.g., image classification, object detection and segmentation, people are diving into more complex vision tasks. Video instance segmentation is a new task which includes detection, segmentation and tracking of instances simultaneously in a video. Occluded Video Instance Segmentation (OVIS) is used for this task, and it includes many heavily occluded scenes. Besides, there is a long range for the length of videos in this dataset. In order to track instances in videos with different lengths, we make some improvements based on MaskTrack R-CNN. Based on these optimizations, a refinement model can be well used to detect and segment instances, which acquires a better track accuracy in long videos. Furthermore, we apply Stochastic Weights Averaging training strategy to get a better result. Finally, The proposed method can achieve the mAP score of 28.9 for the validation set and 32.2 for the test set on the OVIS dataset.

1. Introduction

Video Instance Segmentation (VIS) is first proposed in [11]. It's a new task to detect, segment and track instances in a video simultaneously. This task extends the instance segmentation on a static image to multi frames in a video. Furthermore, tracking instances among different frames is also required. The evaluation metrics in [11] give a detailed description, $\mathbf{O}^i_{t_1 \ldots t_T}$ and $\tilde{\mathbf{O}}^j_{t_1 \ldots t_T}$ denotes an ground truth and the predicted instance in a video. It evaluates the mask accuracy in a frame, in the meanwhile, tracking failed will result in a low VIS IoU.

$$IoU(i,j) = \frac{\sum_{t=1}^{T} |\mathbf{O}^i_t \cap \tilde{\mathbf{O}}^j_t|}{\sum_{t=1}^{T} |\mathbf{O}^i_t \cup \tilde{\mathbf{O}}^j_t|} \tag{1}$$

YouTube-VIS [11] which built upon the YouTube-VOS [10] gives a new large-scale benchmark. There are 2883 videos, 40 categories and 4883 instances in this datasets. All the videos are no more than 6 seconds. Compared to this, Occluded Video Instance Segmentation (OVIS) [6] collects more data with heavily occluded scenes. It only has 901 videos and 25 categories but with 5223 instances and 296k masks. Moreover, the length of videos have a long range between 5s to 60s. Based on the definition of mBOR in [6], which used to evaluate the degree of occlusion, the mBOR value of OVIS is 0.22 however it only gets 0.07 in YouTube-VIS. In conclusion, this is a difficult multi-task learning problem. but it's obvious that OVIS gives more challenge on this task.

In general, methods dealing with these problems can be divided into two parts. Tracking-based methods [9, 11, 5] and transformer-based methods [8, 3]. The tracking based methods have 4 branches to classify, detect, segment and track instances simultaneously. MaskTrack R-CNN [11] adds a track branch over the Mask R-CNN [2]. Specifically, the tracking branch is used to track instances in the current frame based on instances saved in the memory queue. As a result, methods like these can be applied to videos with any lengths. VisTR [8] extends DETR [1] to the VIS task, which first introduced the transformer [7] architecture into the computer vision task. Tracking results are parallelly obtained from the box and mask results, this is due to it predicts a fixed-size sequence of N predictions. Networks like that need a fixed number of input images once a time. Padding will be needed when a video is shorter than the fixed length. However, if a video with length longer than the fixed length, it has to be split into several clips with shorter length. Besides, post-processing will be needed to link the results from different clips.

In the OVIS Dataset, The number of frames in videos are widely distributed. The longest videos even have up to 500 frames. As for our proposed method, we extend MaskTrack R-CNN with a new reference frames sampling technique and an enhanced backbone network. This model can be applied to videos of any length without any post-processing. Finally we adopt a training strategy from [12], and get 0.89 AP and 3.0 AP improvement for val and test set.

*Interns at Ant Group.

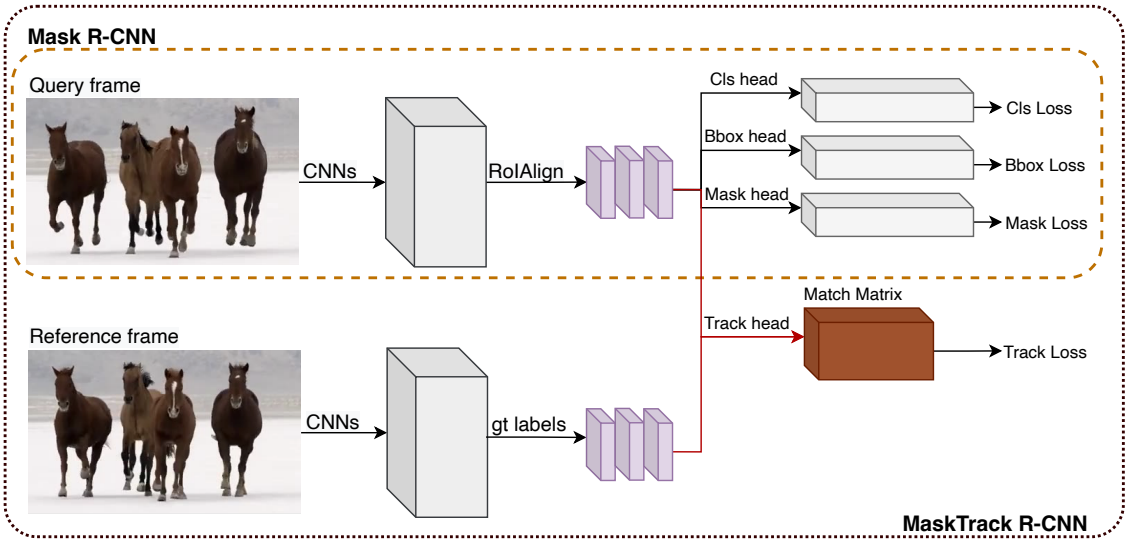

Figure 1. Mask R-CNN and MaskTrack R-CNN in our approach.

2. Approach

Our approach is developed based on the MaskTrack R-CNN [11]. The structure is illustrated in Figure 1. For a complete gradient computation, two frames named query frame and reference frame are needed. For the classification loss, bbox regression loss and mask loss, only query frame is used. As for the reference frame, we only extract its features with instances in the its label, together with the region features of interest in the query frame, they are used to get the track loss. These four kinds of loss are put together to get the gradient information.

2.1. Sampling Strategy

The inference pipeline and training pipeline of MaskTrack R-CNN are different. When the model used to evaluate a video, the first frame is fed into the model to predict instances on the frame. In addition to that, all the instances features are saved in the memory queue to be matched in the future. After that, the instance results of the next frame are predicted in the same way. Based on the current instances features and features saved in the memory queue, the tracking branch are used to track the instances. In the meanwhile, instances that have already been appeared, the features of which are used to update the features in the memory queue.

When training MaskTrack R-CNN in the YouTube-VIS dataset, the reference frame are sampled randomly from the same video as the query frame, as illustrated in Figure 2(a). It works well cause the max length of all videos is only 36 frames, which as a result, the difference between query and reference frame instances is not going to have long distances in the feature space. However, when it comes to the OVIS, the length of all videos has hundreds of frames. The model will be trained with two frames which have long time interval, if sampling reference frame randomly. As for the same instances in these two frames, they differ a lot in feature space. It's quite hard for model to match them and leads unhealthy gradients. On the other hand, we update the instance features timely in the inference pipeline, instead of using the old features. So it's not necessary to train track branch with long time interval frames.

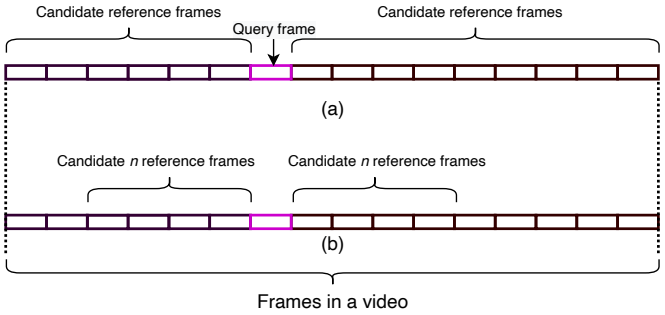

Figure 2. Reference frame sampling strategy in [11] and our method.

Based on this, we proposed the new sampling strategy with limited reference frame sampling range, as denoted in Figure 2(b). The sampling range of reference frame is limited to the range of n frames before and after the query frame. With a suitable n, loss of track branch return to the normal level. And this is also more in line with the situation at the inference pipeline. In our experiment, we use $n = 5$ to conduct training pipeline.

Table 1. Final results on OVIS dataset, MaskTrack R-CNN* denotes training with the proposed sampling strategy and Swin Large Transformer backbone network.

Methods	OVIS validation set					OVIS test set				
	AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
MaskTrack R-CNN	10.8	25.3	8.5	7.9	14.9	11.8	25.4	10.4	7.9	16.0
CSipMask	14.3	29.9	12.5	9.6	19.3	14.5	31.1	13.5	9.0	19.4
CMaskTrack R-CNN	15.4	33.9	13.1	9.3	20.0	15.1	31.6	13.2	9.8	20.5
MaskTrack R-CNN*	**28.0**	**56.5**	**25.8**	**13.6**	**33.1**	-	-	-	-	-
MaskTrack R-CNN*+SWA	**28.9**	**56.3**	**26.8**	**13.5**	**34.0**	**32.2**	-	-	-	-

2.2. Backbone

Both the training and inference pipelines, Instance features are used to fed into the track head to match them. which means it has to be accurate for the instance segmentation in a single frame. Here, we use Swin Transformer [4] to extract features, which has been proved to have better performance in many experiments.

3. Experiments

3.1. Training Details

We follow the architecture of MaskTrack R-CNN in [11] and use Swin Large Transformer to extract features. And all the backbones are pretrained with ImageNet. As for sampling the reference frame, we use the strategy illustrated in Figure 2(b) with $n = 5$. During training, all the images are scaled to 960×540, and flip ratio is set to 0.5. No other data augmentation is applied. Besides, the max number of instances in a frame are set to 15.

3.2. Experimental Results

As shown in Table 1, Our proposed method finally achieves 28.9 mAP and 32.2 mAP on the OVIS validation and test set. And this method ranks first place in the 1st Occluded Video Instance Segmentation Challenge.

3.3. Ablation Study

In this section we study the contribution of our proposed method and how we achieve the final results as show in Table 2. The baseline is the original MaskTrack R-CNN and with ResNet50. All the results are evaluate on OVIS validation set.

4. Conclusion

In this paper, we propose a new sampling strategy for reference frame in MaskTrack R-CNN. With a powerful Swin Large Transformer backbone and SWA training strategy, our approach achieves mAP score of 28.9 on OVIS validation set and 32.2 on OVIS test set, ranking the first place in the 1st Occluded Video Instance Segmentation Challenge.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[3] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *arXiv preprint arXiv:2106.03299*, 2021. 1

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 3

[5] Thuy C Nguyen, Tuan N Tang, Nam LH Phan, Chuong H Nguyen, Masayuki Yamazaki, and Masao Yamanaka. 1st place solution for youtubevos challenge 2021: Video instance segmentation. *arXiv preprint arXiv:2106.06649*, 2021. 1

[6] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation. *arXiv preprint arXiv:2102.01558*, 2021. 1

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[8] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 1

[9] 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. IEEE, 2017. 1

[10] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 1

[11] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International*

Table 2. Ablation results on OVIS dataset

Methods	mAP	Boost
Baseline(ResNet50, random sampling strategy)	8.7	-
+ ResNet101 Backbone	11.2	2.5
+ limited range sampling strategy $n = 24$	13.1	1.9
+ limited range sampling strategy $n = 5$	14.1	1.0
+ Swin Small Window 7 Transformer Backbone	20.6	6.5
+ image scaled to 960×540	22.9	2.3
+ Swin Basic Window 7 Transformer Backbone	25.3	2.4
+ Swin Large Window 12 Transformer Backbone	28.0	2.7
+ SWA training strategy	28.9	0.9

Conference on Computer Vision, pages 5188–5197, 2019. 1, 2, 3

[12] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. Swa object detection. *arXiv preprint arXiv:2012.12645*, 2020. 1