Occluded Video Instance Segmentation with Set Prediction Approach

Heechul Bae, Soonyong Song, Junhee Park
Industry and IoT Intelligence Research Department
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, South Korea
{hessed, soony, juni}@etri.re.kr

Abstract

Occluded Video Instance Segmentation (OVIS) is a multi-task problem performing detection, segmentation, and tracking simultaneously under severe occlusions. We propose an extended model for the OVIS task based on the real-time one-stage instance segmentation method. The proposed model was applied to the OVIS dataset hold by the ICCV 2021 - Occluded Video Instance Segmentation Workshop 2021. We also show that the occlusions can be handled efficiently through one-stage approaches.

1. Introduction

Video instance segmentation tasks aim to classify, segment, and track individual instances' pixel-level segmentation masks. The slightly and heavily occluded objects make tracking and understanding the video scene and frame difficult. On the OVIS validation set, the highest mean AP in some cases of severe occlusions is only 15.4[11].

The 2021 Occluded Video Instance Segmentation (OVIS) challenge [11] encourages the development of accurate models for motion and occluded objects understanding in video and provides a new large-scale occluded benchmark dataset. Occluded video instance segmentation requires simultaneous detection, segmentation, and tracking under heavy occlusions for video scenes understanding.

Recent approaches for video instance segmentation tasks, such as MaskTrack R-CNN[15] and MaskProp[1], were based on two-stage methods like Mask R-CNN[7]. Two-stage models typically prioritize performance over simplicity and speed. YOLACT++[3] is a one-stage instance segmentation method that performs two parallel subtasks independently (prototype masks, instance mask coefficients). The YOLACT++ is one of the earliest instance segmentation methods for real-time using simple and parallel fully-convolutional structures. Many recent studies such as STMask[9], SipMask[4], and TensorMask[6] are approaching single-stage methods for real-time application with proper accuracy.

Most of the one-stage instance segmentation tasks are based on anchor-free object detection, such as CenterNet[16], CenterMask[8], and FCOS[14]. However, YOLACT++[3] and STMask[9] are anchor-based object detection methods to produce class confidences and bounding box regression coefficients for each anchor. Especially, YOLACT++[3] also improve the traditional sequential NMS(Non-Maximum Suppression) approach and introduce Fast-NMS for real-time instance segmentation. NMS was used at first in edge detection approaches[13]. It has been an essential part of object detection models. NMS and many of its variants, such as Soft-NMS[2], have received enormous attention in the research. In many detection approaches[7], [10], [12], NMS is usually conducted sequentially. Especially, It is necessary to study the effective and accurate bounding box in occlusion situations. One of the limitations of the anchor box technique is that when objects of similar shapes overlap, it is still challenging to derive a proper anchor box within one same grid. Moreover, it is also difficult to detect and segment small objects on the edge of frames.

This paper proposes an extended model based on the real-time instance segmentation method, which is then applied to a severe occlusions-affected OVIS benchmark dataset. We show how occlusions can be handled efficiently using a one-stage approach with a set prediction approach, such as YOLACT++[3] and STMask[9]. We present a set prediction approach for efficient bounding box selection.

The paper is structured as follows. In Section 2, we summarize our overall methods, including dataset, metrics, proposed model, and experimental setup. The results and discussion are addressed in Section 3. In Section 4, we finally present our conclusions.

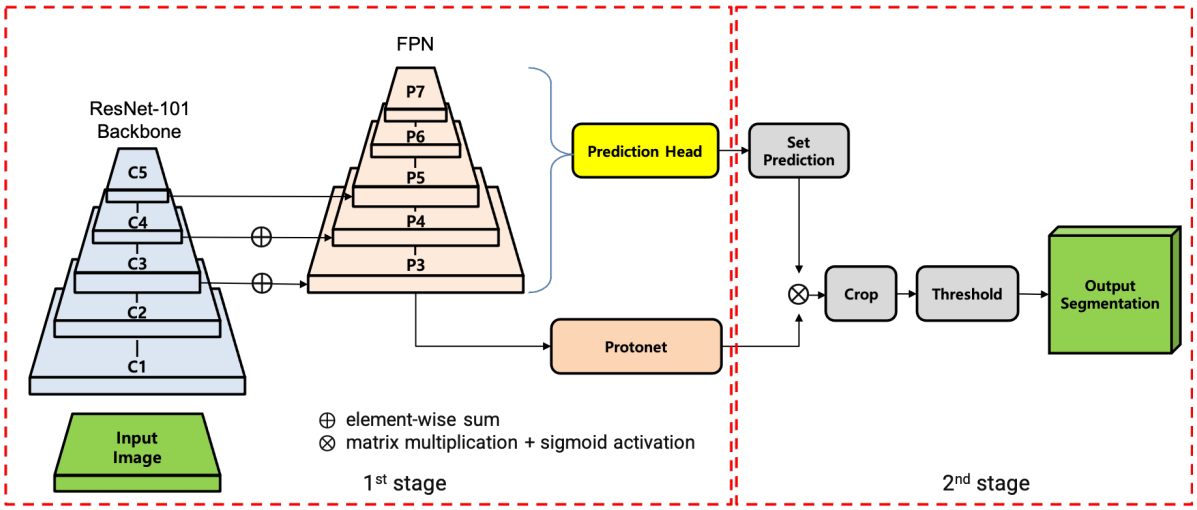

Figure 1. The proposed architecture is based on YOLACT++[3]. It has a ResNet-101 backbone and Feature Pyramid Network

2. Methods

2.1. Dataset

OVIS dataset is made up of 296k high-quality instance masks from 25 commonly seen semantic categories, where severe object occlusions usually happen at a total of 901 videos and 5,223 unique instances. The dataset is available at http://songbai.site/ovis. Distinctive properties of the OVIS dataset are severe occlusions, long videos, and crowded scenes to reserve enough motion and occlusion scenarios [11]. They[11] explain video collection, annotation, and statistics of the OVIS dataset well in section 3. Statistics of the dataset can be summarized from [11] as follows.

- The average video and instance duration of OVIS are 12.77s and 10.05s respectively. Video length is generally between 5s and 60s. Thus, the length of the video varies from 15 to 500 frames for each video item. Each video has about 69.5 frames (see Table 1) .
- Video resolution is also different, and generally 1920 x 1080.
- The OVIS dataset has 5.80 instances in each video and 4.72 objects in each frame.

In addition to severe occlusions, the long videos and crowded scenes properties could be taken into consideration. The OVIS dataset can be summarized in Table 1.

2.2. Metrics

The 2021 Occluded Video Instance Segmentation(OVIS) challenge provides a standard evaluation methodology. Two evaluation metrics are chosen to assess the performance of the OVIS task: Average Precision(AP) and Average Recall(AR). The performance is mainly measured by AP, defined as the area under the precision-recall (PR) curve.

Dataset	video items	min length	max length	total frames	average frames
training	607	15	500	42,149	69.4
validation	140	15	292	8,784	62.7
testing	154	21	309	11,708	76.0
total	901	15	500	62,641	69.5

Table 1. The OVIS 2021 dataset

Data loading and evaluation functionalities for video instance segmentation are provided at https://github.com/youtubevos/cocoapi. It is built based on COCO API designed for the MSCOCO dataset (http://cocodataset.org/). Thus, COCO-style annotated data can be adapted for OVIS using COCO API.

2.3. Proposed model

Network architecture Our model is based on YOLACT++[3] and STMask[9], as depicted in Figure 1. The network includes a ResNet-101 with deformable convolution layers (interval=3) backbone and a Feature Pyramid Network to extract features with different image sizes.

Tracking method Our tracking method is inspired and used from STMask[9]. To extract features using anchors with different image sizes, STMask[9] uses a new bounding box regression branch replacing single 3x3 convolution with three aspect ratios of convolutions, 3x3, 3x5, 5x3, respectively. We also use the bounding box regression branch and temporal fusion module by STMask[9].

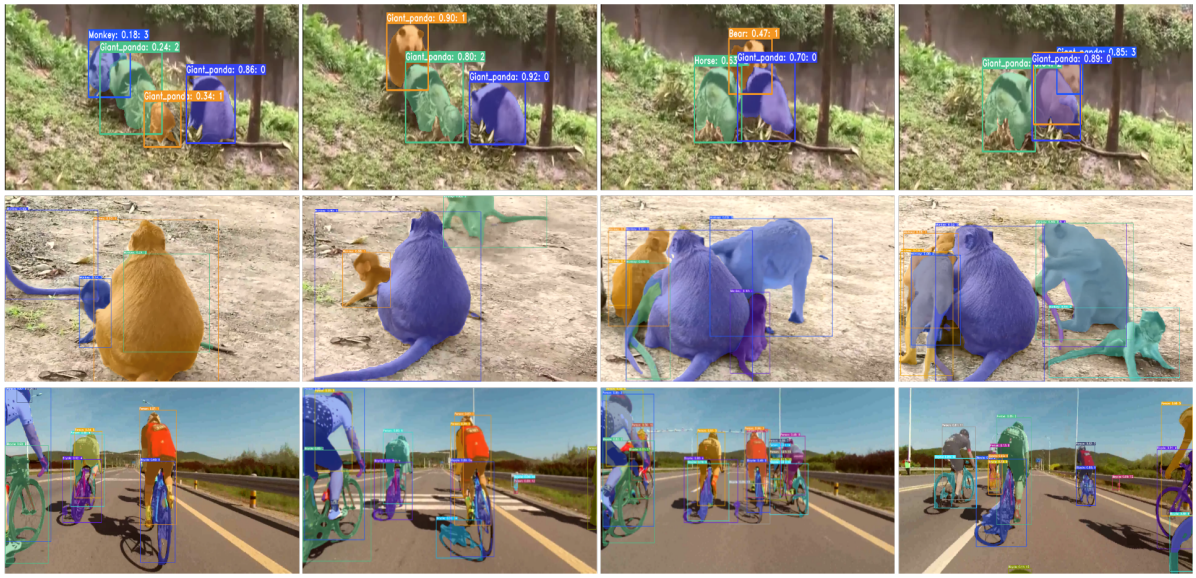

Figure 2. Results on OVIS's validation set. Each row is a video.

Bounding box selection and inference YOLACT++ is an anchor-based one-stage model. We focus on improving anchor and bounding boxes to address the occlusion problem. To overcome the problem, we suggest a set prediction approach for an efficient bounding box selection. To fix and improve the sequential process of conventional NMS, we just followed parallel process from YOLACT++[3]. The main difference is to guide to find set matching using direct set prediction approach without duplicates.

In Figure 1, the first stage generates bounding box regression coefficients and class confidences for each anchor, and the second stage uses the coefficients and confidences for detection. Detection failure has emerged as a result of our experiments. To suppress duplicate detections, a set-based matching method was used. DETR[5] inspired and influenced our set prediction module.

2.4. Experimental setup

We start to train the proposed model using YOLACT++[3] pre-trained weights. The basic training setup is mainly following the original YOLACT++[3] and STMask[9]. We train the model for up to 250k iterations with a learning rate of 0.0001, the momentum of 0.9, weight decay of 0.0001, and batch size of 8. The training process takes about four days on two NVIDIA RTX 2080Ti for OVIS training datasets. The implementation version is with PyTorch 1.4 and Torchvision 0.5, and CUDA 10.1 with Python 3.7 on Ubuntu 18.04. The training and evaluations are performed with two NVIDIA RTX 2080Ti GPUs and i9-9900X CPU.

3. Results and discussion

The results of the OVIS 2021 validation dataset are shown in Table 2. The outcomes are denoted in bold. Our solution achieves 15.33 mAP on the validation dataset. We see that the result has a low level of accuracy but high quality and consistency in the masks. Figure 2 depicts some of our model's validation dataset prediction results. Three giant pandas heavily overlap each other in the first row of the Figure. The failure of tracking and segmenting three giant pandas is caused by light and heavy occlusion. There are some failures to detect a giant panda as other instances such as a horse, a monkey, and a bear. In the middle row of the Figure, there are several monkeys under severe occlusions. It looks good to detect and track the monkeys, but it fails to show exact bounding boxes and masks. In the last row of the Figure, bicycles and person are overlapped heavily with each other. The model sometimes fails and duplicates to track a person and segment a bicycle. Moreover, it is not correct to detect and segment small objects on the edge of frames.

We can observe that YOLACT++[3] and STMask[9] are a strong baseline on their own. We wished this post-processing modification was increasing the accuracies, but it rarely influenced. Moreover, it is still difficult to detect and segment small objects on the edge of frames with different classes instances. These problems represent valuable future research works. In the future, we also plan to use the set prediction approach to training in the end-to-end process, such as DETR[5].

Team	mAP	AP_{50}	AP_{75}	AR_1
Ach	28.04	56.47	25.75	13.55
huapohen	25.18	41.80	25.96	13.91
Ali2500	19.75	40.29	17.97	12.48
LI-Minghan	19.25	38.32	17.68	10.41
sabarim	18.38	39.40	16.25	11.41
taihengYe	15.42	33.89	13.06	9.26
qjy	15.42	33.89	13.06	9.26
hessed	15.33	33.79	12.48	8.87
JialeCao	13.22	29.39	11.04	8.47
HaichaoShi	13.07	26.65	11.81	8.88

Table 2. Results and comparison of top 10 teams in the OVIS Challenge 2021

4. Conclusion

In this paper, we looked at the occluded video instance segmentation task. The OVIS challenge introduces a new large-scale occluded benchmark dataset. The OVIS benchmark dataset's unique properties, such as severe occlusions, long videos, and crowded scenes, provide enough motion and occlusion scenarios in video instance segmentation tasks. Based on YOLACT++[3] and STMask[9], we demonstrated that occlusions could be handled efficiently using a one-stage approach. We were also successful in introducing a bounding box selection network by incorporating a set prediction approach. The results of the OVIS challenge 2021 show that occlusion is a significant barrier to understanding video scenes.

Acknowledgment

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [21ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling and evolving ways]

References

[1] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9739–9748, 2020. 1

[2] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 1

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact++: Better real-time instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 3, 4

[4] 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. 1

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020. 3

[6] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. 2019. 1

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 1

[8] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. 2020. 1

[9] Minghan Li, Shuai Li, Lida Li, and Lei Zhang. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In *CVPR*, 2021. 1, 2, 3, 4

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1

[11] 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, 2

[12] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 1

[13] Azriel Rosenfeld and Mark Thurston. Edge and curve detection for visual scene analysis. *IEEE Transactions on computers*, 100(5):562–569, 1971. 1

[14] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: A simple and strong anchor-free object detector. 2021. 1

[15] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. *CoRR*, abs/1905.04804, 2019. 1

[16] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 1