1st Place Solution for YouTubeVIS Long Video Challenge

Cong Wei[1]*, Yong Liu[1,2]*, Jixiang Sun[1]*, Yitong Wang[2], Yansong Tang[1], Yujiu Yang[1]
[1] Shenzhen International Graduate School, Tsinghua University
[2]ByteDance Inc.
{liu-yong20,weic22,sun-jx22}@mails.tsinghua.edu.cn,
wangjingshen@bytedance.com,
{tang.yansong,yang.yujiu}@sz.tsinghua.edu.cn,

Abstract

Video Instance Segmentation (VIS) is an important vision task that aims to simultaneously perform classification, tracking, and segmentation in videos. To solve VIS in scenes involving long video sequences, we believe that the segmentation performance of single frame is of great importance. Besides, due to the computational limits, offline methods usually require hand-designed clip matching while online methods have inherent advantage in handling long video sequences. Thus, we take online method as baseline to process these long videos. In this report, we will describe how we can further improve the performance of the state-of-the-art online methods. With different model ensemble, the proposed method finally obtains 42.9 AP on the YouTube-VIS 2022 long video test set and was ranked first place in the YouTube-VIS Long Video Challenge.

1. Introduction

Video instance segmentation (VIS) [21] is one of the most fundamental tasks of computer vision and the extension of image instance segmentation. It aims at simultaneously classifying, tracking, and segmenting objects in videos. Due to its wide applications in video editing, autonomous driving, and augmented reality, VIS has attracted great attention in recent years.

With the development of deep learning, there have been many excellent works focusing on video instance segmentation. Generally speaking, current methods can be divided into two categories: offline methods and online methods. Offline methods [1, 2, 6, 8, 11, 18, 19] take the whole video as input and output video instances simultaneously. In contrast, online methods [3,5,7,9,10,20,22] take each frame as input and perform classification and segmentation frame by frame. Finally, the instances of different frames are linked by the design of tracking head.

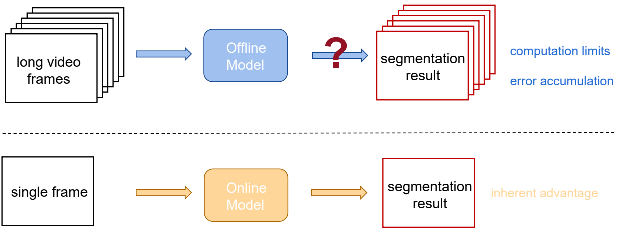

Figure 1. Illustration of offline methods and online methods. The offline methods take the whole video as input and make predictions simultaneously. But for long video task, they may show inability due to computational limits. On the contrary, the online methods have inherent advantage by processing each frame independently.

Since taking the entire video as input, the offline method can utilize the rich context information of the whole video sequences. However, for difficult scenes like long videos, the advantage of aggregating inter-frame information of the offline method may vanish due to the the computational limits. Besides, with the video sequences growing longer, it may lead to more noise while taking too many frames into consideration to aggregate inter-frame information. Fig. 1 illustrates this problem and the comparison between the online methods and offline methods. Thus, we take online methods to process the challenging videos in YouTube-VIS 2022 dataset. Specifically, we adopt the IDOL [20], the state-of-the-art online models, as our baseline. The core idea of IDOL is processing each frame independently and learning more discriminative instance embeddings through contrastive learning. This approach allows queries of each frame to contain sufficient intra-frame instances information and introduces no more information from other frames when updating queries of each frame, which is beneficial to process long video sequences.

Since the difficulty of associating different frames for long video instance segmentation, it's drastically important

*Equal Contribution.
This work was done during Yong Liu's internship at ByteDance Inc.

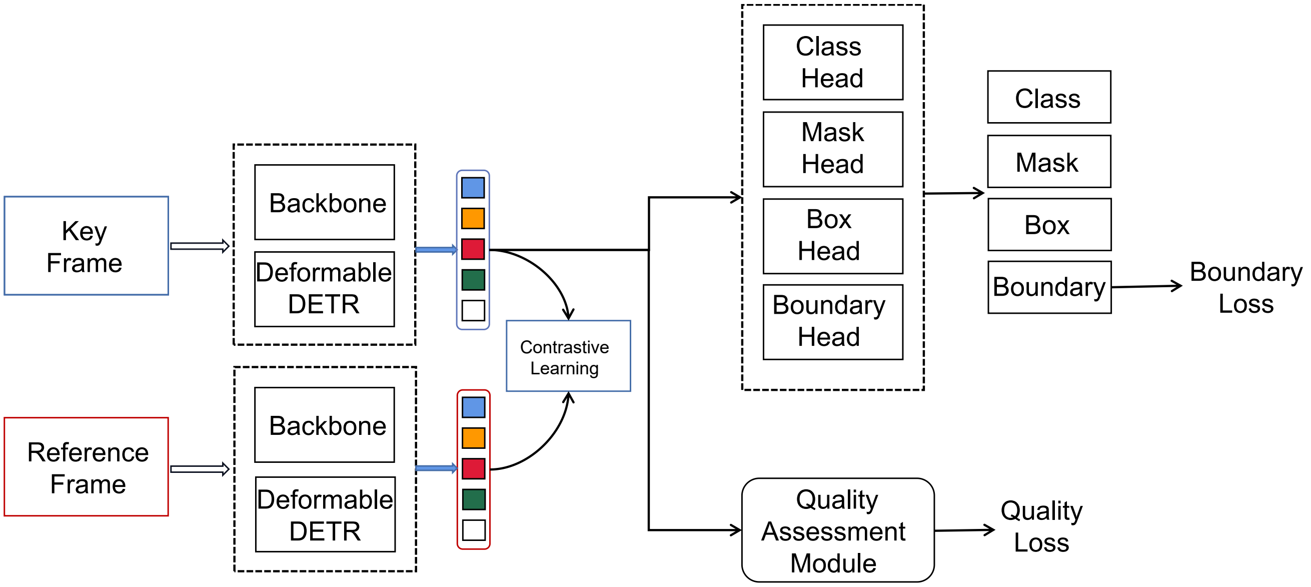

Figure 2. Overview of our training pipeline. The network takes key frame and reference frame as input. Two parallel backbone and deformable DETR produce the feature embedding for contrastive learning. The boundary head and quality assessment module are added to better distinguish different instances and make the training process more robust.

to improve the representation ability of each frame's query. In other words, the single frame segmentation performance is of great importance especially for the long videos. To this end, we introduce some modules to improve the single frame processing part based on IDOL [20] . First of all, it is important to correctly identify the object boundaries in each frame especially when video sequences become much longer. Because incorrect boundary information may lead to error accumulation and do harm to association of different frames, which has a significant impact on the segmentation and classification. In this case, focusing more on the object boundaries helps the model to better distinguish the different instances and make better performance in tracking the long videos. Besides, inspired by the [14, 15], we apply the quality assessment module only at the training stage. The quality assessment module can improve the robustness of the training process by making the model predict its own accuracy.

Thanks to the superior performance of online methods [20] and the above improvements, we achieved the first place in Long Video Instance Segmentation Track of the 4th Large-scale Video Object Segmentation Challenge with the score of 42.9 AP on the final test set.

2. Method

2.1. Overview

The overview of our training framework is illustrated in Fig. 2. Following recent work [20], we first take a key frame and a reference frame as input. They are passed into a share-weighted backbone which extracts their feature maps. Then, the feature maps are passed to the Deformable DETR [23]module along with additional fixed positional encodings [4] and N learnable object queries to predict the instance embedding for contrastive learning. After that, the feature embedding originated from the key frame is used to predict classes and masks. Besides, we additionally add the boundary head to learn the boundaries of different instances and make the model utilise more discriminative information through a boundary loss. Following [14], we pass the the instance embedding of key frame to quality assessment module to obtain a segmentation quality score, which is used to calculate the quality loss. This module is only conducted during training. Finally, we adopt the same memory bank-based instance association strategy as [20] during inference.

2.2. Boundary Branch

The previous VIS methods mainly predict masks and classes of a frame and additionally add the bounding boxes to track the different instances. But for long video scenarios, it's difficult to perform accurate frame matching for each instance, which may lead to error accumulation. So it would be meaningful to identify the target instances if the model could pay more attention to the shape or boundary of the objects. So we add the boundary branch to make the model pay more attention to the instances boundaries to learn stronger representation and thus provide more guidance for mask and class prediction.

Given a key frame $X \in R^{3\times H\times W}$of a video, H and W are the height and width of the input frame, a CNN-based backbone extract the feature maps, which are passed to the Deformable DETR module along with additional fixed positional encodings and N learnable object queries. In that case, the object queries are transformed into instance embeddings $E \in R^{N \times C}$. The Boundary Head performs several convolution layers on the embeddings E similar to the mask head and predicts the boundaries for each instance.

Boundary Ground Truth. Following [13], we predict boundaries of each frame in the manner of pixel-level classification. Besides, in view that only the ground truth of the mask is available in the VIS dataset, we obtain the boundaries using the Laplacian operator and binarize them with a threshold T_b into binary maps.

Boundary Loss. Following [13], we use dice loss [17]and binary cross-entropy to calculate the total boundary loss. Dice loss measures the overlap between prediction and ground truth and is often used to resolve category imbalance problem. Therefore, it's naturally compatible with boundary prediction because the points of boundaries are usually much less than points of non-boundary. The boundary loss L_b is formulated as:

$$\mathcal{L}_b = \mathcal{L}_{Dice} + \mathcal{L}_{BCE}, \tag{1}$$

The dice loss L_{Dice} is formulated as:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_i p_i q_i}{\sum_i (p_i)^2 + \sum_i (q_i)^2 + \epsilon}. \tag{2}$$

where p and q denote the predictions and ground-truth, i denotes the i-th pixel and ϵ is a smooth term.

2.3. Quality Assessment Module

For long video instance segmentation, it's difficult to maintain mask quality, which may lead to error accumulation as frame sequences grow much longer. Therefore, it's significant to improve segmentation quality and the robustness of the training process by making the model predict its own accuracy.

Specifically, we apply the Quality Assessment Module(QAM) following [14] for our VIS model and improve the mask quality through the quality loss. Given the instance embeddings $E \in R^{N \times C}$ originated from the key frame, QAM takes two FFN like layers to predict the mask quality score S for the key frame. Besides, we calculate maskIoU between the predicted mask and ground truth as the target value of the quality score. The process is as follows:

$$V_i = maskIoU(M_i, GT_i) \tag{3}$$

$$\mathcal{L}_q = \frac{1}{N}\sum_{i=1}^{N}(S_i - V_i)^2 \tag{4}$$

where S_i represents the quality score of the segmentation mask for i-th object, M_i indicates the predicted mask, GT_i is the ground truth and N indicates the total number of instances.

3. Experiment

3.1. Implementation Details

We took the Swin Transformer-Large [16] as backbone for our model taking IDOL [20] as baseline, the training setting is generally same as initial IDOL. We used AdamW optimizer with initial learning rate of 1e-4. Note that we did not perform pre-training on COCO dataset [12] but initialized the model by the pre-trained weights of IDOL directly. To train the proposed modules and finetune the IDOL part, we randomly cropped the image from COCO twice to generate the pseudo training videos. Then, we train our model on the pseudo video set and the YouTube-VIS 2022 train set for 175000 and 40000 iterations with batch size of 8, respectively. For training data augmentation, we performed multi-scale training scales and resized the shortest side to [320, 352, 392, 416, 448, 480, 512, 544, 576, 608, 640]. All models are trained on 8 80GB A100 GPUs. During inference, the input videos are resized with the short size of 720 pixels in default.

3.2. Comparison with Other Methods

In the YouTube-VIS Long Video Challenge, we rank the first place on the test set. The leaderboard is shown in Tab. 1. It can be seen that with the help of the above modules, our model achieved the 42.9 mAPL (1st) and 46.8 AP75L (1st), and surpasses others by 2.3 on mAPL(mAPL means the resulting mAP score of long videos which is same for the rest of metrics). The good performance above demonstrates that the recognition and detection effect can be significantly improved by focusing on object boundaries and improving the mask quality as well as the training robustness of the model.

Table 1. Comparison with other methods on the long video test set.

Method	mAPL	AP50L	AP75L	AR1L	AR10L
Ours	**42.9**	**60.7**	**46.8**	**35.0**	**51.4**
Man	40.6	58.4	43.4	34.0	53.0
sjx	40.6	60.7	42.2	32.6	49.0
ID4	40.2	61.1	41.7	32.6	55.0
SakuraT	38.6	57.8	39.1	34.1	53.7

Table 2. Ablation study of our applied modules on the long video test set.

Method	mAPL	AP50L	AP75L
Baseline	39.45	58.76	42.06
+Quality assessment	40.58	60.68	42.17
+Boundary branch	41.65	59.90	44.56
+Model ensemble	42.9	60.7	46.8

3.3. Ablation Study

In this section we analyze the effectiveness of our applied modules on the long video test set and the results are shown in Tab. 2. The baseline is the IDOL with Swin-L backbone. Integrated with the quality assessment module, our method achieved the score of 40.58 mAP. After applying the boundary branch and corresponding loss, the performance is improved to 41.65 mAP. Finally, by ensembling with the different combination of the above methods, the performance eventually reached 42.9 mAP, ranking first place in the YouTubeVIS Long Video Challenge.

4. Conclusion

In this work we introduce the boundary module and quality assessment module for online VIS framework. The proposed modules make the model utilise more discriminative information through the boundary learning and improve segmentation quality and the robustness of the training process. With the above modules and model ensemble, we achieved the first place in Long Video Instance Segmentation Track of the 4th Large-scale Video Object Segmentation Challenge with the score of 42.9 AP on the test set.

References

[1] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020. 1

[2] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*, 2020. 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. In *ECCV*, 2020. 1

[4] 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. 2

[5] Yang Fu, Linjie Yang, Ding Liu, Thomas S. Huang, and Humphrey Shi. Compfeat: Comprehensive feature aggregation for video instance segmentation. In *AAAI*, 2021. 1

[6] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *arXiv preprint arXiv:2206.04403*, 2022. 1

[7] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *arXiv preprint arXiv:2208.02245*, 2022. 1

[8] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. In *NeurIPS*, 2021. 1

[9] Lei Ke, Xia Li, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Prototypical cross-attention networks for multiple object tracking and segmentation. In *NeurIPS*, 2021. 1

[10] 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

[11] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. In *ICCV*, 2021. 1

[12] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 3

[13] Yong Liu, Ran Yu, Jiahao Wang, Xinyuan Zhao, Yitong Wang, Yansong Tang, and Yujiu Yang. Global spectral filter memory network for video object segmentation. *arXiv preprint arXiv:2210.05567*, 2022. 3

[14] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. *arXiv preprint arXiv:2207.07922*, 2022. 2, 3

[15] Yong Liu, Ran Yu, Xinyuan Zhao, and Yujiu Yang. Quality-aware and selective prior enhancement memory network for video object segmentation. In *CVPR Workshop*, 2021. 2

[16] 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. In *ICCV*, 2021. 3

[17] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 3

[18] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 1

[19] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. *arXiv preprint arXiv:2112.08275*, 2021. 1

[20] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. *arXiv preprint arXiv:2207.10661*, 2022. 1, 2, 3

[21] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 1

[22] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. In *ICCV*, 2021. 1

[23] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2