1st Place Solution for YouTubeVOS Challenge 2022: Video Instance Segmentation

Junfeng Wu[1*] Xiang Bai[1] Yi Jiang[2] Qihao Liu[3*] Zehuan Yuan[2] Song Bai[2]

[1]Huazhong University of Science and Technology
[2]ByteDance [3]Johns Hopkins University

Abstract

Video Instance Segmentation (VIS) is an emerging vision task that aims to simultaneously perform detection, classification, segmentation, and association of object instances in videos. In this report, we propose an online method based on contrastive learning that is able to learn more discriminative instance embeddings for association and fully exploit history information for stability. The proposed method obtains 53.4 AP on the YouTube-VIS 2022 dataset and was ranked first place in the video instance segmentation track of the 2022 Large-scale Video Object Segmentation Challenge. Moreover, the proposed method outperforms previous state-of-the-art methods on the YouTube-VIS 2019 dataset in a fair comparison. We hope the simplicity and effectiveness of our method could benefit further research.

1. Introduction

Video instance segmentation aims at detecting, segmenting, and tracking object instances simultaneously in a given video. It has attracted considerable attention after first defined [19] in 2019 due to the huge challenge and the wide applications in video understanding, video editing, autonomous driving, augmented reality, etc. Current VIS methods can be categorized as online or offline methods. Online methods [3,5,8,9,19,20] take as input a video frame by frame, detecting and segmenting objects per frame while tracking instances and optimizing results across frames. Offline methods [1,2,7,10,17,18], in contrast, take the whole video as input and generate the instance sequence of the entire video (or video clip) with a single step.

Most online VIS methods are built upon image-level instance segmentation with an additional tracking head to associate instances across the video. To improve the association performance, we take advantage of contrastive learning and propose an online model for video instance segmentation. The key idea is to ensure, in the embedding space, the similarity of the same instance across frames and the difference of different instances in all frames, even for instances that belong to the same category and have very similar appearances. It provides more discriminative instance features with better temporal consistency, which guarantees more accurate association results.

Our method achieves an overall first place in the YouTube-VIS Challenge 2022, with the score 57.6 AP on the public validation set, and 53.4 AP on the private test set. We conduct more experiments on the YouTube-VIS 2019 [19] dataset, on which the proposed method outperforms all previous state-of-the-art methods in a fair comparison. We believe the simplicity and effectiveness of our method shall benefit further research.

2. Related Work

Online Video Instance Segmentation. Most previous online VIS methods follows the tracking-by-detection paradigm, extending image instance segmentation models with a tracking branch. The baseline method MaskTrack R-CNN [19] is built upon Mask R-CNN [6] and introduces a tracking head to associate each instance in the video. SipMask [3] follows the similar pipeline based on the one-stage FCOS [15]. CompFeat [5] proposes a comprehensive feature aggregation approach to refine features with temporal and spatial context information, and a new tracking module to enhance the local discriminative power of features with local and global correlation maps. STMask [9] proposes a spatial feature calibration to predict bounding boxes and extract features from them for frame-level instance segmentation, then aggregate temporal information from adjacent frames and track instances across frames. CrossVIS [20] proposes a new learning scheme that uses the instance feature in the current frame to pixel-wisely localize the same instance in other frames. Recently, PCAN [8] proposes a cross-attention network to retrieve rich information from past frames. Compared with the offline models, the online

* Work done during an internship at ByteDance.

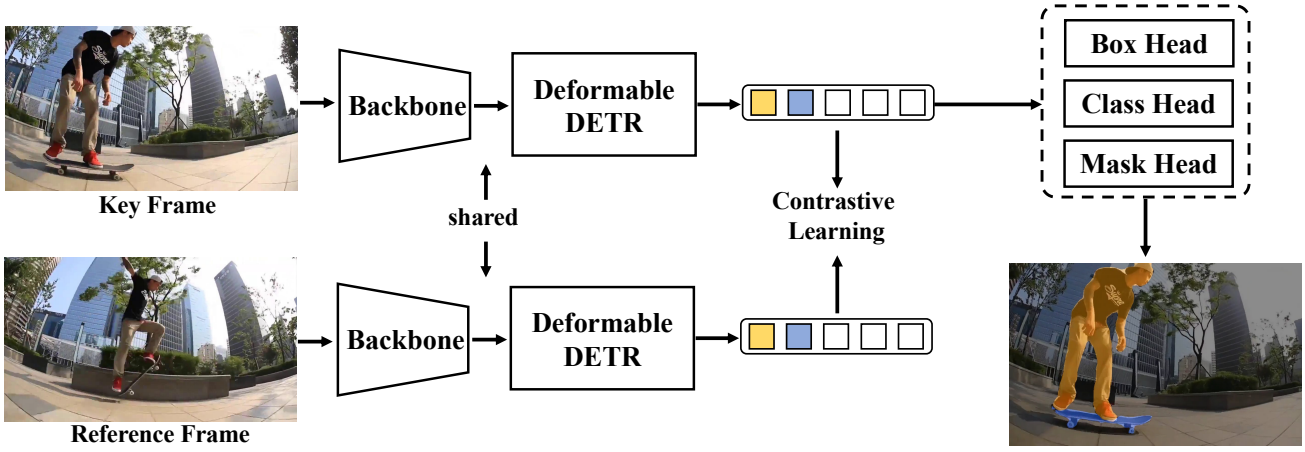

Figure 1. The training pipeline of our method. Given a key frame and a reference frame, the shared-weight backbone and transformer predict the instance embeddings on them respectively. The embeddings on the key frame are used to predict masks, boxes, and categories, while the embeddings on the reference frame are selected as positive and negative embeddings for contrastive learning.

models have a wider range of application scenarios.

Offline Video Instance Segmentation. Offline methods for VIS take the whole video as input and predict the instance sequence of the entire video or video clip with a single step. MaskProp [2] and Propose-Reduce [10] perform mask propagation in a video clip to improve mask and association. However, the propagation process is time-consuming, which limits its application. STEm-Seg [1] models a video clip as a single 3D spatial-temporal volume and enables inference procedure based on clustering. Recently, VisTR [17] adopts the transformer [16] to VIS and models the instance queries for the whole video. However, it learns an embedding for each instance of each frame, which makes it hard to apply to longer videos and more complex scenes. SeqFormer [18] dynamically allocates spatial attention on each frame and aggregates spatial-temporal features of the same instance to learn a video-level instance embedding, which greatly improves the performance on VIS.

3. Method

3.1. Instance Segmentation

We propose a contrastive learning method to extract more discriminative features for instance association. Previous online VIS models [19, 20] utilize additional association head upon on instance segmentation models [6, 14]. Following the state-of-the-art method [18], we take DeformableDETR [21] with dynamic mask head [14] as our instance segmentation pipeline in this report. Our method can be coupled with other instance segmentation methods with minor modifications.

Given an input frame $x \in R^{3\times H\times W}$ of a video, a CNN backbone extracts multi-scale feature maps. The Deformable DETR module takes the feature maps with additional fixed positional encodings [4] and N learnable object queries as input. The object queries are first transformed into output embeddings $E \in R^{N\times C}$ by the transformer decoder. After that, they are decoded into box coordinates, class labels, and instance masks following SeqFormer [18] as shown in Fig 1. Then we calculate pair-wise matching cost which takes into account both the class prediction and the similarity of predicted and ground truth boxes. Finally, the whole model is optimized with a multi-task loss function

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1\mathcal{L}_{box} + \lambda_1\mathcal{L}_{mask} + \lambda_2\mathcal{L}_{embed}, \quad (1)$$

$\mathcal{L}_{embed}$ is the contrastive loss described in the next section.

3.2. Contrastive Learning between Frames

More discriminative feature embeddings can help distinguish instances on different frames, thereby improving the quality of cross-frame association. To this end, we introduce contrastive learning between frames to make the embedding of the same object instance closer in the embedding space, and the embedding of different object instances farther away. Object queries are used to query the features of instances from each frame in our instance segmentation pipeline. Therefore, the output embeddings can be regarded as features of different instances. We employ an extra light-weighted FFN as a contrastive head to decode the contrastive embeddings from the instance features.

Given a key frame for instance segmentation training, we select a reference frame from the temporal neighborhood. The instances appearing on the key frame may have different positions and appearances on the reference frame, but their contrastive embeddings should be as close as possible in embedding space. For each instance in the key frame,

we send the output embedding with the lowest cost to the contrastive head and get the contrastive embedding **v**. If the same instance appears on the reference frame, we select positive and negative samples for it according to the cost with predictions. The contrastive loss function for a positive pair of examples is defined as follows:

$$\mathcal{L}_{embed} = \log[1 + \sum_{\mathbf{k}^+} \sum_{\mathbf{k}^-} \exp(\mathbf{v} \cdot \mathbf{k}^- - \mathbf{v} \cdot \mathbf{k}^+)]. \quad (2)$$

where $\mathbf{k}^+$ and $\mathbf{k}^-$ are positive and negative feature embeddings from the reference frame, respectively.

3.3. Instance Association

Given a test video, we initialize an empty memory bank for it and perform instance segmentation on each frame sequentially in an online scheme. Assume there are N instances predicted by the model with N contrastive embeddings, and M instances in the memory bank. We compute similarity score f between predicted instance i and memory instance j , and search for the best assignment for instance i by:

$$\hat{j} = \arg\max \mathbf{f}(i, j), \forall j \in \{1, 2, ..., M\}. \quad (3)$$

If $\mathbf{f}(i, \hat{j}) > 0.5$, we assign the instance i on current frame to the memory instance $\hat{j}$. For the prediction without an assignment but has a high class score, we start a new instance ID in the memory bank.

4. Experiments

4.1. Implementation Details

We use the same setting for Deformable DETR and the dynamic mask head following SeqFormer [18]. For the transformer, we use 6 encoder and 6 decoder layers of width 256 with bounding box refinement mechanism, and the number of object queries is set to 300. We use AdamW [13] optimizer with base learning rate of 2×10^{-4}. The models are first pre-trained on the MS COCO 2017 dataset for instance segmentation. After that, we randomly and independently crop the image from COCO twice to form a pseudo key-reference frame pair, which is used to pre-train the contrastive embedding of our models. Then, the models are trained on YouTube-VIS 2022 for 12000 iterations, the learning rate is decayed by a factor of 0.1 at the 6000 iterations. For data augmentation in training, we use multi-scale training scales at [320, 352, 392, 416, 448, 480, 512, 544, 576, 608, 640] for shortest side. During inference, the input frames are downscaled that the shortest side is at 480 pixels by default. For multi-scale testing, the shortest side is at [480, 640, 800]. The model is trained on 8 A100 GPUs, with 4 pairs of frames per GPU.

Team	mAP	mAP_S	AP50_S	mAP_L	AP50_L
Ours	**53.4**	**59.6**	**80.7**	**47.2**	67.5
peiyunc1989	50.9	57.1	79.6	44.8	**70.5**
mahogany	50.7	57.1	77.5	44.2	63.6
Ach	48.1	54.3	77.0	41.9	65.3
zhangtao-whu	48.0	53.0	73.6	43.1	67.1

Table 1. Comparison with other methods on the Youtube-VIS 2022 test set.

Method	mAP	AP50	AP75	AR1	AR10
MaskTrack R-CNN [19]	30.3	51.1	32.6	31.0	35.5
SipMask [3]	33.7	54.1	35.8	35.4	40.1
CompFeat [5]	35.3	56.0	38.6	33.1	40.3
CrossVIS [20]	36.3	56.8	38.9	35.6	40.7
STEm-Seg [1]	30.6	50.7	33.5	37.6	37.1
VisTR [17]	36.2	59.8	36.9	37.2	42.4
Propose-Reduce [10]	40.4	63.0	43.8	41.1	49.7
IFC [7]	42.8	65.8	46.8	43.8	51.2
SeqFormer [18]	45.1	66.9	50.5	**45.6**	54.6
Ours	**47.1**	**71.7**	**51.4**	44.7	**54.5**

Table 2. Comparison with other methods on the Youtube-VIS 2019 validation set. We use a ResNet-50 backbone and single scale testing for fair comparison. **Bold** represent the best metrics.

4.2. Main Results

We evaluate the performance of the proposed method by participating in the YouTube-VIS Challenge 2022 and achieve first place. As shown in Table 1, our method achieves 53.4 AP on the Youtube-VIS 2022 test set and surpasses others by a large margin. For a fair comparison with previous methods, we conduct experiments on the standard YouTube-VIS 2019 dataset, as shown in Table 2. With the same ResNet-50 backbone and single scale testing, we achieve 47.1 AP, surpassing the previous state-of-the-art method by 2.0 AP.

4.3. Ablation Study

In this section we study how we achieve the final results as show in Table 3. The baseline is with ResNet-50 backbone and single-scale testing. Integrated with the Swin Transformer backbone [11], our method achieves a much higher AP of 53.0. The pseudo key-reference frame pair pre-train improves the AP from 53.0 to 55.2. After that, we utilize multi-scale testing for further boosting performance. Different from image instance segmentation, the IoU computation is carried out in both spatial domain and temporal domain. Multi-scale testing can further improve the result from 55.2 to 56.6. Finally, by ensembling Swin-L and ConvNext-L [12] in the same way, we achieve the state-of-the-art with 57.6 AP. All the results are evaluate on Youtube-VIS 2022 validation set.

Method	mAP	ΔmAP	mAP_S	mAP_L
ResNet-50	42.4	-	45.5	39.2
Swin-L	53.0	10.6	57.6	48.4
+pseudo frame	55.2	2.2	59.7	50.7
+multi-scale	56.6	1.4	61.2	52.0
+multi-model	57.6	1.0	61.7	53.6

Table 3. Ablation study on the Youtube-VIS 2022 validation set.

5. Conclusions

In this work, we propose an online method to perform the VIS task based on contrastive learning that is able to learn discriminative instance embeddings for instance association. Our method achieves first place in the YouTube-VIS Challenge 2022, with 57.6 AP on the validation set, and 53.4 AP on the test set. In addition, the proposed method outperforms all previous state-of-the-art methods on the YouTube-VIS 2019 dataset in a fair comparison. We believe the simplicity and effectiveness of our method shall benefit further research.

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, 3

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

[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, 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[5] Yang Fu, Linjie Yang, Ding Liu, Thomas S Huang, and Humphrey Shi. Compfeat: Comprehensive feature aggregation for video instance segmentation. *arXiv preprint arXiv:2012.03400*, 2020. 1, 3

[6] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 2

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

[8] 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

[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

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

[11] 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

[12] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 3

[14] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020. 2

[15] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, 2019. 1

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[17] 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, 2, 3

[18] 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, 2, 3

[19] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 1, 2, 3

[20] 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, 2, 3

[21] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2