Direct Association of Object Queries for Video Instance Segmentation

Sukjun Hwang[1*] Miran Heo[1*] Seoung Wug Oh[2] Joon-Young Lee[2] Seon Joo Kim[1]

[1]Yonsei University [2]Adobe Research

Abstract

Recent Transformer-based offline Video Instance Segmentation (VIS) studies have shown that localizing the information in Transformer layers is more effective than attending to the entire spatio-temporal feature volume. From this observation, we hypothesize that explicit use of object-oriented information on spatial scenes can be a strong solution for understanding the context of the entire sequence. Thus, we introduce a new paradigm for offline VIS that learns to integrate decoded object queries from independent frames. Specifically, we propose a simple module that can be easily built on top of an off-the-shelf Transformer-based image instance segmentation model. Leaving the frame-level model to distill the rich knowledge of the spatial scene into its object queries, the proposed module directly associates and identifies the given potential objects by building temporal interactions in between. With a Swin-L backbone, our proposed method sets a record of 50.7 AP which ranks the 3^{rd} place in Track 2-Video Instance Segmentation of the 4^{th} Large-scale Video Object Segmentation Challenge.

1. Introduction

Video Instance Segmentation (VIS) is the task of predicting both mask trajectories and object classes for each object belonging to a set of predefined categories. Early-stage methods divide the problem into two components, segmentation and association, following the paradigm of *tracking-by-detection*. Leaving detection of each frame to the off-the-shelf image instance segmentation model, they focus on associating the independent detection [11].

However, their heavy reliance on image detector and context-limited architecture restricts their performance on video-specific challenges such as motion blur and occlusion. To alleviate such limitations, clip-level VIS methods take short video clip as input and predict object tracklets within a given local window. Subsequently, they match adjacent tracklets deploying its richer contextual information. While these approaches mitigate above issues, post-processing like NMS still remains, and induce error propagation.

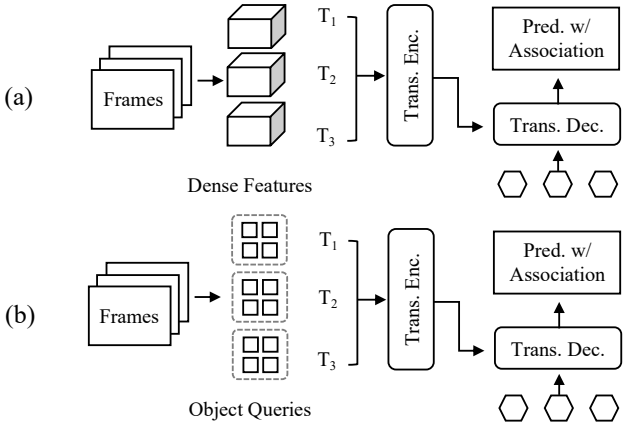

Figure 1. (a) Existing transformer-based VIS methods jointly track and segment instances in an end-to-end manner by employing dense spatio-temporal features. (b) On the other hand, we propose a new paradigm that directly leverages object queries for VIS.

With the motivation of relieving the locality and eliminating the heuristics, VisTR [8] made the first attempt to design an end-to-end model which jointly predicts object trajectories with corresponding segmentation masks. Specifically, VisTR follows the bipartite matching loss approach from DETR [1], and their set-based nature successfully eliminates the post-processing. Also, they process a whole video sequence at once by modeling long-range dependencies utilizing a whole pixel-level feature sequences as an input to the transformer encoder.

To further reduce the heavy computation of VisTR, IFC [5] adopts memory tokens to the transformer encoder. They prove that the efficient communication between context-representative tokens is sufficient rather than the dense self-attention of the spatio-temporal pixel-level feature. Recent Mask2Former-VIS [2] extends a strong transformer-based universal image segmentation model [3]

* Both authors contributed equally to this work.

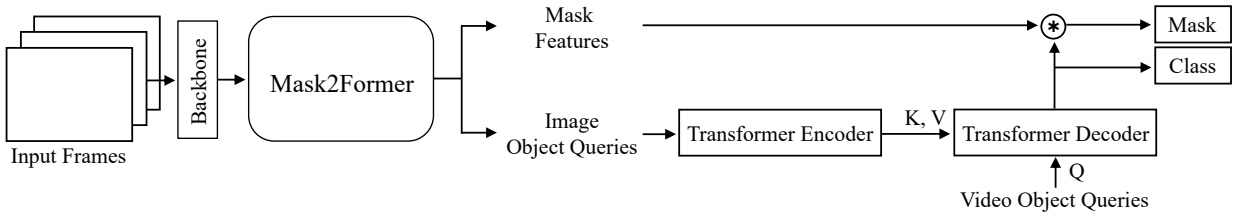

Figure 2. Overview of our framework.

to the VIS task. Taking advantage of its segmentation-oriented representation, Mask2Former-VIS records compelling performance on benchmarks without bells and whistles. This *jointly-track-and-segmentation* paradigm has achieved the highest performance in the VIS community to date (Figure 1. (a)).

As such, current VIS community has explored the above two paradigms separately. In this paper, we integrate both and propose a new paradigm of jointly learning trajectories and masks while having tracking-by-detection philosophy. Our key motivation is that if we take the object-oriented information rather than dense scene-based one, the transformer can sufficiently capture object relationship (Figure 1. (b)). To the end, we propose a simple module that can be easily built on top of the off-the-shelf image-level segmentation model. Given the decoded object queries from the off-the-shelf image-level segmentation model, the proposed module directly decodes the learnable clip-level queries taking frame-level object queries as key and value.

Even with this simple and intuitive structure, the proposed model shows competitive performance of 50.7 AP, and ranks the 3^{rd} place in Track 2-Video Instance Segmentation of the 4^{th} Large-scale Video Object Segmentation Challenge. More concrete explanation of our methodology and experimental results on other datasets such as YouTube-VIS 2019 & 2021, and OVIS can be found in [4].

2. Method

The overview of proposed method is illustrated in Figure 2. Our framework consists of two main modules: a frame-level detector and a video-level object associator.

2.1. Frame-level detector

We adopt Mask2Former [3] for the frame-level detector. Given input clip of T frames , we use two predictions from Mask2Former that are frame-level object queries and mask features. Specifically, frame-level object queries holds object-centric information of each frame, and mask features from the pixel decoder are exploited to predicts final mask tracklets.

2.2. Video-level associator

Video-level associator directly takes only two types of features as input that are independently generated by frame-level detector for entire clip: mask features and frame-level object queries. By directly constructing temporal interactions between frame-level object queries that encapsulate rich object-aware knowledge in spatial scenes, our framework yields mask trajectories with corresponding categories in an end-to-end manner. Our video-level associator follows the standard Transformer encoder and decoder architectures suggested in DETR [1].

Transformer encoder. Although the frame-level detector distills its spatial knowledge in its frame-level object queries, there is a lack of temporal interaction between different frames. Thus we make a temporal connection by stacking layers of Transformer encoder. However, a naive self-attention over the whole object queries is inefficient when processing long videos. Inspired by [7], we adopt window-based self-attention layers that shift along the temporal dimension.

Transformer decoder. After object queries in different frames undergo temporal interaction through the Transformer encoder layers, we stack Transformer decoder layers that performs cross-attention between learnable video-level object queries and the frame-level object queries. The use of object queries corresponding to the instance identity of the clip is similar to previous studies [2, 5], but the explicit use of object-oriented features rather than pixel-level features as keys and values is completely different.

Output heads. Having the decoded video-level object queries, we use two output heads for the final predictions like [5]: the class head and the mask head. The class head is a single linear classifier, which directly predicts class probabilities of each video-level object query. And the mask head dynamically generates mask embeddings per a video-level object query, which corresponds to the tracklet of an instance.

Method	mAP	mAP_S	AP_{50}	AP_{75}	AR_1	AR_{10}	mAP_L	AP_{50}	AP_{75}	AR_1	AR_{10}
SakuraT	53.4	59.6	80.7	65.2	48.1	66.5	47.2	67.5	48.6	37.2	60.9
peiyunc1989	50.9	57.1	79.6	61.7	46.8	61.7	44.8	70.5	43.7	32.7	50.2
mahogany (ours)	**50.7**	**57.1**	**77.5**	**61.7**	**46.8**	**61.4**	**44.2**	**63.6**	**46.1**	**38.7**	**48.9**
tianxingye	48.8	55.2	75.2	61.8	46.6	61.0	42.5	63.4	47.9	35.0	48.1
Ach	48.1	54.3	77.0	57.3	45.7	61.5	41.9	65.3	41.2	36.9	47.2
zhangtao-whu	48.0	53.0	73.6	59.0	45.4	58.4	43.1	67.1	46.6	36.0	48.2
JZ	48.0	54.3	77.0	57.1	46.0	61.5	41.7	64.7	41.6	36.8	47.3
M2F	47.1	53.1	73.5	59.2	45.3	58.4	41.2	63.3	43.8	33.9	45.4
anwesac2	46.6	52.9	74.7	58.0	45.0	60.7	40.2	61.4	43.6	33.3	49.0
justin-zk	46.5	52.8	75.7	56.8	45.6	59.3	40.3	64.1	40.0	35.8	45.4

Table 1. Final leaderboard in the YouTube-VIS Challenge 2022. We list the top 10 participants. Our results are highlighted in **bold**.

3. Experiments

3.1. Implementation Detail

Our proposed method is implemented on top of `detectron2` [10]. We use Swin-L [7] for backbone network, and all hyper-parameters regarding the frame-level detector (Mask2Former) are equal to the defaults of [3]. By default, our Transformer encoder of video-level associator is composed of three layers with the shifted-window size of 6, and Transformer decoder employs six layers with 100 video-level object queries.

Training. Having video-level associator built on top of Mask2Former [3], we first train our model on the COCO [6] dataset following [3]. The model is then fine-tuned using pseudo-videos generated from the images [6] and the YouTube-VIS 2021 `train` set simultaneously. We follow SeqFormer [9] to generate pseudo-videos from a single image, and we use loss functions same as [4].

Inference. During inference, each frame is resized to a shorter edge size of 448 pixels. We divide the input video into short clips with overlapping frames, and then stitch the predicted tracklets sequentially. The size of clip windows and stride is 18 and 3, respectively. The matching algorithm will be explained in Section 3.4.

3.2. Dataset

The dataset for YouTube-VIS Challenge 2022 includes both YouTube-VIS 2021 and additionally introduced YouTube-VIS 2022 dataset. And both of them uses same 40-category label set.

YouTube-VIS 2021 consists of 3,859 high-resolution videos total which includes 2,985 videos for training, 421 videos for validation, and 453 videos for testing.

YouTube-VIS 2022 tackles the scenarios of long and highly complicated sequences. In detail, it consists of 71 additional long videos in `valid` set and 50 additional long videos in `test` set.

3.3. Evaluation Metric

YouTube-VIS Challenge 2022 follows the standard evaluation metric [11]. Unlike previous challenges, the final mAP is computed by the average of the results for the YouTube-VIS 2021 and YouTube-VIS 2022 datasets. In this paper, we denote $mAP = (mAP_S + mAP_L)/2$, where mAP_S and mAP_L is the result of YouTube-VIS 2021 and YouTube-VIS 2022, respectively.

3.4. Clip Matching

Similar to previous per-clip VIS methods [5], we use a post-processing to matching and classifying each tracklet. During inference, we use window and stride sizes of 18 and 3, respectively. Then, similar to [11], we calculate matching scores of each clip-wise predictions and use the costs for matching.

3.5. Results

In Table 1, we list the results of the top 10 participants of the final leaderboard in YouTube-VIS Challenge 2022. Our proposed method achieved 50.7 AP which ranks the 3^{rd} place in Track 2. Note that we do not use separate models and inference algorithms for testing YouTube-VIS 2021 and YouTube-VIS 2022.

4. Conclusion

We introduced a novel framework for offline Video Instance Segmentation (VIS). Unlike existing Transformer-based VIS methodologies, our proposed method helps object-focused video understanding by directly deploying object queries that independently decoded by image-based instance segmentation model. With a Swin-L backbone, we achieved competitive performance of 50.7 AP which ranks the 3^{rd} place in Track 2 of the 4^{th} Large-Scale Video Object 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 *ECCV*, 2020. 1, 2

[2] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 1, 2

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1, 2, 3

[4] 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. 2, 3

[5] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. In *NeurIPS*, 2021. 1, 2, 3

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

[7] 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. 2, 3

[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 *CVPR*, 2020. 1

[9] 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. 3

[10] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 3

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