Multi-Constrained Tracking Approach for Video Instance Segmentation

Fang Gao
School of Electrical Engineering,
Guangxi University
Nanning, China
fgao@gxu.edu.cn

Yan Jin*
School of Electrical Engineering,
Guangxi University
Nanning, China
2112401017@st.gxu.edu.cn

Wenjie Wu
School of Electrical Engineering,
Guangxi University
Nanning, China
2212392119@st.gxu.edu.cn

Lei Shi
School of Electrical Engineering, Guangxi University
Nanning, China
sl20000929@163.com

Shengheng Ma
Guangxi Ctech Alpha Technology Co., LTD.
Nanning, China
zkma@ctunite.com

ABSTRACT

Video Instance Segmentation (VIS) is an emerging computer vision task that involves performing object detection, classification, segmentation, and instance association simultaneously within videos. The primary objective of VIS is to identify and segment individual object instances present in video frames. In this report, we propose an online approach that integrates object interaction relationships, allowing for learning spatial interactions between objects to facilitate association. Additionally, addressing the challenge of significant object instance variations in low-frame-rate video sequences, we endeavor to introduce a deformable compensation module to dynamically adjust attention regions. The method we proposed achieved an AP of 52.5 on the YouTubeVIS 2023. We trust that the straightforward yet impactful attributes of our method have the potential to contribute positively to future research endeavors.

1 INTRODUCTION

Video instance segmentation, a complex task involving the simultaneous detection, segmentation, and tracking of object instances within videos, has garnered significant attention post-2019 [1], [2]. This surge in interest can be attributed to its multifaceted challenges and wide-ranging applications, spanning video comprehension, editing, autonomous driving, and augmented reality.

Methods for Video Instance Segmentation (VIS) can be broadly categorized into online, where frames are processed individually, and offline, where the entire video is considered as a whole. Online methods, exemplified by [3]–[5], handled each frame independently, executing object detection, segmentation, and tracking in tandem, while optimizing outcomes across frames. In contrast, offline techniques such as [6]–[9] generated instance sequences for the entirety of the video content.

The majority of online VIS methods extended image-level instance segmentation by incorporating tracking mechanisms for association. Leveraging contrastive learning enhanced their ability to associate instances, ensuring better performance. The fundamental principle underlying this improvement was the establishment of similarity between instances of the same kind across frames, while simultaneously differentiating between instances that might look alike, thus enhancing temporal consistency and overall accuracy.

Recent advancements have converged tracking and segmentation in the realm of VIS. VisTR [7], for instance, introduced an end-to-end approach that predicted trajectories and segmentation masks simultaneously. IFC [9] introduced memory tokens to alleviate computation burdens, while Mask2Former-VIS [10] extended a transformer-based image segmentation model to this context. Our contribution lies in the integration of tracking and segmentation, resulting in a highly competitive performance that secured a remarkable third-place ranking in the 5th Large-scale Video Object Segmentation Challenge. Further insights can be found in [11], which provides detailed explanations and experimental results on prominent datasets like YouTubeVIS and OVIS.

2 Method

Our proposed approach, as depicted in Figure 1, comprises three essential functional components: the segmenter, the Dual-Constraint Tracker, and the Interactive Information Module. To implement the framework, we adopt Mask2Former as our segmenter. The Dual-Constraint Tracker, outlined in section 2.1, encompasses the Abrupt Information Filtering Block and the Deformation Compensation Block. The specifics of the Interactive Information Module are detailed in section 2.2.

2.1 Dual-Constraint Tracker

The Dual-Constraint Tracker consists of two components: Abrupt Information Filtering and Deformation Compensation. Addressing abrupt information and low frame rate issues in the

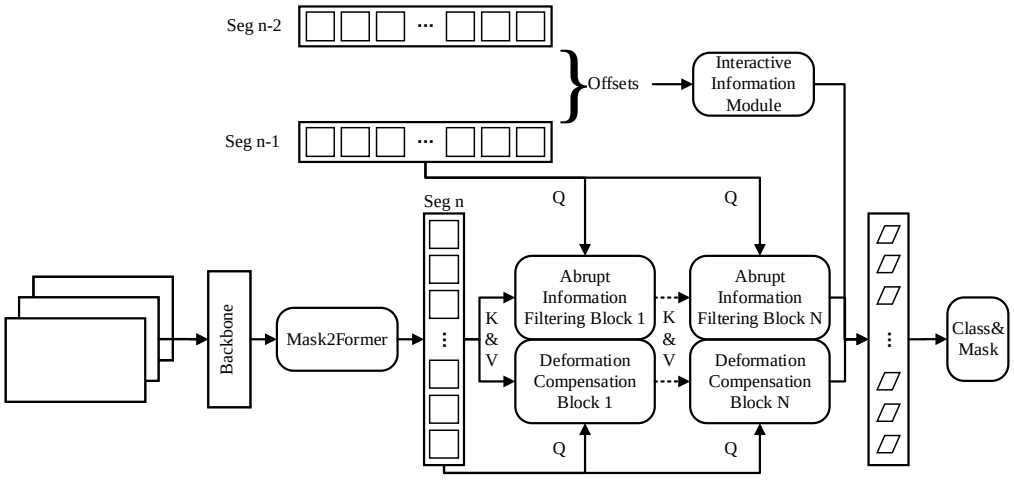

Figure 2. Overview of our proposed framework.

tracking process, we employ these two modules to salvage frame rate.

Figure 2 illustrates the network structure of the Dual-Constraint Tracker. It enforces constraints on segmentation results from two perspectives: abrupt information filtering and the introduction of deformable attention. The tracker takes the instance output from the segmenter and combines it with interactive information to output the tracking result for the current frame. On one hand, abrupt information filtering effectively utilizes similarity between instance representations in neighboring frames, enhancing the tracker's discriminative ability. On the other hand, deformation compensation [12] better adapts to drastic changes caused by object motion in low frame rate videos, capturing object correlations between different frames more accurately. Ultimately, this improves tracking stability and accuracy.

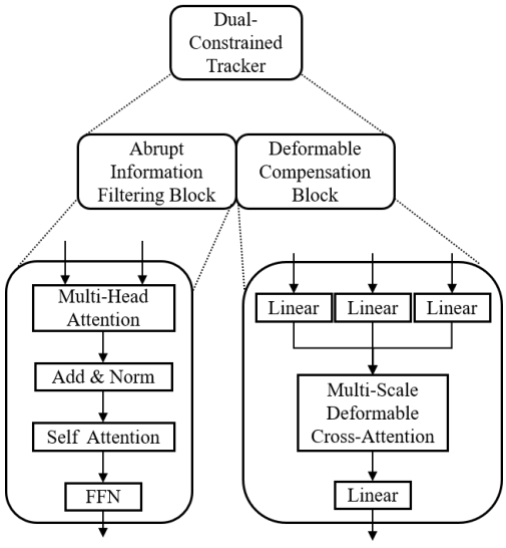

Figure 1. The Dual-Constraint Tracker.

The Abrupt Information Filtering Block, depicted in Figure 2, takes key (K) and value (V) from the current frame's segmentation result, along with query (Q) from the previous frame's segmentation result. It processes these through multi-head attention, self-attention, and feed-forward neural networks (FFN), ultimately outputting instance queries. The Deformation Compensation Block, as illustrated in Figure 2, also acquires Q, K and V from instances. It's worth noting that Q is obtained directly from the current frame. To simplify deformation information extraction and reduce memory consumption, a linear layer transitions to a multi-scale deformable cross-attention mechanism, which finally outputs deformation information.

2.2 Interactive Information Module

In current VIS algorithms, other specific targeted datasets often exhibit a certain frame rate advantage. However, when confronted with the YouTubeVIS 2023 test, algorithms incorporating spatiotemporal information struggle to leverage their full potential in matching trajectories of low frame rate targets. Hence, in similar low frame rate scenarios with multiple objects, we aim to model interactions between targets. In this manner, while performing tracking association, we introduce more comprehensive instance interactions, thereby providing richer information for instance association.

As shown in Figure 3, we retain the offsets of the first two frames' instances, inputting them into sparse multi-head attention to capture element relationships [13]. Subsequently, an attention

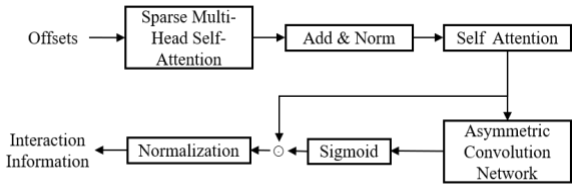

Figure 3. The Interactive Information Module.

Table 1. Final leaderboard in the YouTubeVIS Challenge 2023.

Method	mAP	mAP_S	AP50_S	AP75_S	AR1_S	AR10_S	mAP_L	AP50_L	AP75_L	AR1_L	AR10_L
zhangtao-whu	56.0	62.4	82.9	69.2	50.1	67.8	49.7	71.2	51.8	37.0	54.8
KainingYing	53.0	59.5	81.1	64.3	48.1	65.6	46.4	71.0	47.5	36.5	52.6
GXU(ours)	**52.5**	**58.2**	**80.3**	**63.7**	**48.0**	**63.8**	**46.7**	**68.2**	**48.9**	**38.2**	**53.7**
guojuan	52.3	58.3	80.0	63.7	48.2	64.0	46.3	67.0	48.9	38.1	53.1
jmy	52.1	58.3	79.9	63.8	48.2	64.2	45.9	67.5	49.0	38.4	52.5

mechanism captures instance interactions. After applying asymmetric convolution and Sigmoid filtering, only high-response values are retained from the interaction information. Element-wise multiplication is performed with the original interaction information, yielding the final interaction information.

During instance association, we introduce deformation and interaction information to achieve more precise matching, resulting in improved accuracy. By combining deformation and interaction cues, we enhance the capability of VIS algorithms to handle low frame rate videos with multiple targets, elevating the tracking performance significantly.

3 EXPERIMENTS

3.1 Implementation Details

Building upon an existing segmentation framework, we tackle tracking tasks. Our approach employs Mask2Former as the segmentation component, with the Swin-L [14] backbone network. We maintain the same hyperparameters as the original settings, making slight adjustments to some parameters based on GPU performance. Training and testing are conducted on four A100 GPUs, with batch size set at 8 and a shift window size of 12, considering GPU limitations. Additionally, our model incorporates 200 video-level object queries. This framework allows us to seamlessly integrate segmentation capabilities with tracking tasks, leveraging the strengths of both for improved performance.

Training. Our framework employs Mask2Former as the segmentation component, utilizing its pretrained model. Subsequently, we conduct training on our complete framework using both the COCO dataset and the YouTubeVIS 2021 training set. The loss function is derived from Mask2Former. We employ the AdamW [15] optimizer, setting the initial learning rate to 1e-4. The training process spans 40,000 iterations, with a learning rate decay of 0.1 at 4,000 iterations. To enhance the training dataset, we employ a multi-scale training schedule with resolutions of [360, 480, 600]. This comprehensive training approach capitalizes on the strengths of Mask2Former and efficiently integrates segmentation into the tracking framework.

3.2 Dataset

Similar to previous editions of the competition, the YouTubeVIS 2023 [16] includes the dataset from YouTubeVIS 2022, with the addition of 117 new test video sequences. This iterative approach ensures continuity while introducing fresh data for evaluation, thereby fostering the ongoing advancement of video instance segmentation research.

YouTubeVIS 2019, an extension of the YouTubeVOS [17], stands as a pioneering contribution to video instance segmentation. Released in 2021, it encompasses 3,859 high-resolution YouTube videos and includes category labels for 40 common objects. With a substantial 8,171 distinct video instances and 232,000 precise manual annotations, the dataset enables diverse and accurate training and evaluation. It is partitioned into 2,985 training, 421 validation, and 453 test videos, offering a comprehensive resource for advancing video instance segmentation research.

YouTubeVIS 2023, the pioneering large-scale dataset for video instance segmentation, builds upon the original YouTubeVOS [17]. In its 2023 version, the dataset introduces extended validation and testing subsets with long videos, enhancing its scope and complexity. This extension includes 71 long videos in validation and 89 in the test set, featuring 259 additional video instances in validation and 268 in the test set. Additionally, there are 33,597 more high-quality masks. This version distinguishes itself from the 2022 version by offering a more intricate long video subset in the test set, underscoring its evolution and advancement in facilitating video instance segmentation research.

3.3 Evaluation Metric

For the evaluation of video instance segmentation, we have adapted standard image instance segmentation metrics. These metrics include Average Precision (AP) and Average Recall (AR). AP measures the area under the precision-recall curve, considering confidence scores for instance categorization. It is averaged across various IoU thresholds. AR reflects the highest recall achieved for a fixed number of segmented instances per video. Both metrics are assessed per category and then averaged across categories, providing comprehensive insights into segmentation performance while accounting for instance variations and category characteristics.

3.4 Results

As depicted in Table 1, we present the final rankings and scores of the top five participants in the YouTubeVIS 2023 competition. Our approach achieved an impressive AP score of 52.5. While we initially tested our algorithm on the YouTubeVIS 2019 dataset during the early phases of the competition, we made incremental adjustments to the model as the competition progressed. Looking ahead, we plan to conduct further experiments in subsequent work

to finalize and refine our model, considering the insights gained from the competition's progression.

4 CONCLUSION

We extended the Mask2Former framework and conducted further theoretical analysis of current VIS algorithms using the YouTubeVIS dataset. Addressing challenges posed by low frame rate videos, we introduced the Dual-Constraint Tracker and Interaction Compensation Module. By leveraging multiple instance information for association, our approach achieves higher precision in video instance segmentation, resulting in a third-place ranking in this competition. Furthermore, we intend to enhance and broaden our current model in our upcoming work, resulting in a comprehensive culmination of our endeavors.

REFERENCES

[1] B. Cheng *et al.*, "Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12475–12485.

[2] A. Athar, S. Mahadevan, A. Osep, L. Leal-Taixé, and B. Leibe, "Stem-seg: Spatio-temporal embeddings for instance segmentation in videos," in *Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part XI 16*, 2020, pp. 158–177.

[3] L. Ke, X. Li, M. Danelljan, Y.-W. Tai, C.-K. Tang, and F. Yu, "Prototypical cross-attention networks for multiple object tracking and segmentation," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 1192–1203, 2021.

[4] S. Yang *et al.*, "Crossover learning for fast online video instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8043–8052.

[5] M. Li, S. Li, L. Li, and L. Zhang, "Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11215–11224.

[6] T. Zhang *et al.*, "DVIS: Decoupled Video Instance Segmentation Framework," *arXiv Prepr. arXiv2306.03413*, 2023.

[7] Y. Wang *et al.*, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8741–8750.

[8] H. Lin, R. Wu, S. Liu, J. Lu, and J. Jia, "Video instance segmentation with a propose-reduce paradigm," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1739–1748.

[9] S. Hwang, M. Heo, S. W. Oh, and S. J. Kim, "Video instance segmentation using inter-frame communication transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 13352–13363, 2021.

[10] B. Cheng, A. Choudhuri, I. Misra, A. Kirillov, R. Girdhar, and A. G. Schwing, "Mask2former for video instance segmentation," *arXiv Prepr. arXiv2112.10764*, 2021.

[11] M. Heo, S. Hwang, S. W. Oh, J.-Y. Lee, and S. J. Kim, "Vita: Video instance segmentation via object token association," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 23109–23120, 2022.

[12] J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection".

[13] Z. Qin, S. Zhou, L. Wang, J. Duan, G. Hua, and W. Tang, "MotionTrack: Learning Robust Short-term and Long-term Motions for Multi-Object Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17939–17948.

[14] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[15] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv Prepr. arXiv1711.05101*, 2017.

[16] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5188–5197.

[17] N. Xu *et al.*, "Youtube-vos: Sequence-to-sequence video object segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 585–601.