Robust and Consistent Online Video Instance Segmentation via Instance Mask Propagation

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

[1]Yonsei University
[2]Adobe Research

Abstract

Recent advancements in online Video Instance Segmentation methods show notable performance improvements across benchmarks. However, the leading methods in the tracking-by-detection paradigm often result in temporally inconsistent predictions at both *instance-level* and *pixel-level* that lead to visually unsatisfactory outcomes. To address these challenges, we propose RoCoVIS, a simple yet effective approach that integrates segmentation and tracking to provide consistent online VIS. Our approach is an end-to-end sequential learning where object queries are propagated through mask predictions, improving the accuracy of temporal instance mapping at the pixel level. Additionally, we propose a new label assignment criterion in harmony with our approach. We also examine the limitations and challenges presented by the current standard evaluation protocol (AP) and suggest adopting additional metrics, Tube-Boundary AP and AP^{Pool}. RoCoVIS demonstrates superior performance on challenging VIS benchmarks with a Swin-L backbone and shows competitive results when employing a ResNet-50 backbone. By employing Tube-Boundary AP and AP^{Pool} as metrics to measure mask accuracy and consistency, RoCoVIS outperforms its counterpart, GenVIS, on the HQ-YTVIS and VIPSeg.

Introduction

Video Instance Segmentation (VIS) (Yang, Fan, and Xu 2019) is a fundamental task in computer vision that aims to segment object masks and classify them across the temporal dimension. Given the inherent challenges of tracking in the video domain, modern methods focus on bridging image segmentation models with novel association techniques. Such efforts have successfully driven major progress in the field, propelled by recent advancements in image segmentation models (Cheng et al. 2022; Li et al. 2023a).

Despite this progress, existing methods often produce visually unsatisfactory results in real-world applications. We attribute this discrepancy to two types of inconsistencies: 1) pixel-level inconsistency, which leads to low-quality mask predictions or temporal jittering of the masks, and 2) instance-level inconsistency, which results in redundant mask predictions, *i.e.*, false positives, or ID switching.

We argue that there are two main causes of the problem. First, we identify popular VIS architectures that take the decoupled approach, commonly known as the "tracking-by-detection" paradigm (Fig. 1 (a)). These models generate plausible proposals independently for each frame and mostly focus on grouping (Heo et al. 2022b), propagation (Heo et al. 2023; Li et al. 2023b; Hannan et al. 2023; Zhang et al. 2023), and matching (Huang, Yu, and Anandkumar 2022; Wu et al. 2022b; Ying et al. 2023). However, this method has clear limitations in enhancing the mask quality and the mask consistency across frames, as it heavily depends on the post-association of temporally discretized outputs. On the other hand, some approaches attempt to explicitly incorporate spatio-temporal information to enhance both segmentation and tracking quality, *i.e.*, "joint detection-and-tracking", illustrated in Fig. 1 (b). For instance, VisTR (Wang et al. 2020) and IFC (Hwang et al. 2021) employ Transformer-based architecture that leverages self-attention mechanisms to aggregate spatio-temporal features. While pixel-level feature fusion aids in enhancing prediction quality, these methods also have two major limitations: i) they rely solely on dense correlations without considering object context, and ii) since fusion is confined to a temporal window, inconsistencies remain across different snippets.

Second, we identify the benchmarks' metrics as not encouraging the resolution of the problem. The standard metric in most VIS benchmarks, Average Precision (AP), counts true positives using segment-level Intersection Over Union (IoU), thus overlooking the precision of mask boundaries' quality. Moreover, as extensively discussed in (Dave et al. 2021), the AP implementation in current benchmarks can be gamed and ignores cross-category confidence calibration due to the post-ranking. This results in encouraging redundant mask generations while overlooking the importance of confidence calibration, producing visually unsatisfactory results. We contend that ensuring confidence-calibrated and consistent predictions is crucial for online VIS, considering applications such as autonomous driving and real-time video editing in live streaming.

In this paper, we propose a new class of *online joint detection-and-tracking* model, **RoCoVIS**, that addresses the inconsistency problems (Fig. 1 (c)). Based upon the prior query propagation architecture (Heo et al. 2023), we explore how to efficiently use the temporal context of objects at a

[*]Work partly conducted during an internship at Adobe.

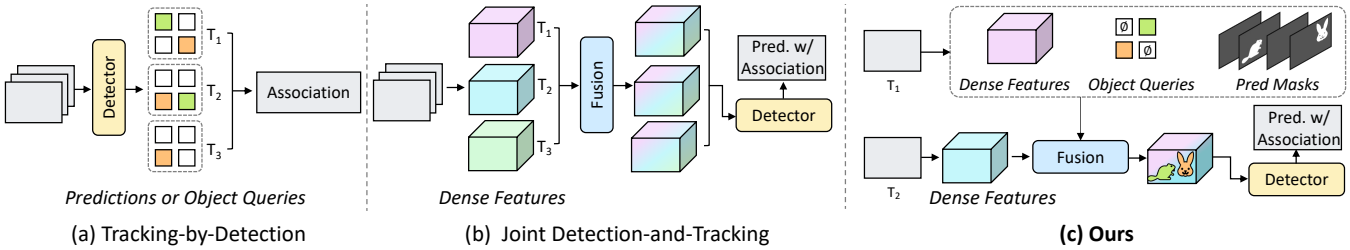

Figure 1: **Comparison of VIS paradigms with our approach.** (a) Prevalent methods associate output predictions or object queries from frame-level detectors. (b) Offline models calibrate dense features for joint detection and tracking. (c) Our method fuses image features with tracked information using mask predictions.

dense, pixel-level scale to improve both the accuracy and consistency of mask predictions. Our approach includes *direct query decoding* and *instance mask propagation*, forming the core of our new architecture. This involves integrating the query decoding directly with dense pixel features to merge tracking and mask prediction seamlessly. We also refine the pixel decoding process, which precedes query decoding, by incorporating object-aware sparse embeddings into corresponding pixels through mask predictions. Furthermore, we introduce a new *label assignment criterion* designed to complement our architecture effectively.

To measure the consistencies and mask quality more accurately beyond the standard metric, we incorporate two additional metrics into our evaluation. For a more precise assessment of mask quality, we utilize Tube-Boundary AP (Cheng et al. 2021b; Ke et al. 2022b), which emphasizes the accuracy of pixels near the boundary contours. To reward calibrated predictions and encourage robust and consistent predictions, we use AP^{Pool} (Dave et al. 2021), which aggregates detections from all classes to form a unified precision-recall curve. This ensures that true positives across all classes outrank any false positives within individual classes. It effectively reduces the bias towards generating redundant predictions and highlights precise predictions.

We validate the general performance of RoCoVIS on popular VIS benchmarks (YouTube-VIS (Yang, Fan, and Xu 2019) and OVIS (Qi et al. 2021)). Based on the standard metric (AP), our method demonstrates competitive performance compared to off-the-shelf approaches with ResNet-50, while showing state-of-the-art performance with large-scale Swin-L backbone, underscoring the effectiveness of end-to-end learning in sequence modeling. To further demonstrate our contributions, we employ two additional benchmarks, HQ-YTVIS (Ke et al. 2022b) and VIPSeg (Miao et al. 2022), which are video segmentation benchmarks noted for their high-quality mask annotations. Through newly employed measures, Tube-Boundary AP and AP^{Pool}, our method achieves significant improvements over the current state-of-the-art models.

To summarize, our contributions are as follows:

- We investigate two temporal inconsistency issues within existing tracking-by-detection VIS methods that lead to visually unsatisfactory results.
- We introduce RoCoVIS, a simple yet scalable strategy, that simultaneously improves segmentation and tracking through temporal modeling in an end-to-end manner.
- We propose instance mask propagation that incorporates object sequence information into dense features, aligned with a new label assignment criterion.
- We also highlight overlooked issues in standard metrics and adopt additional metrics to assess consistency and mask quality more accurately.

Related Works

Video Instance Segmentation

In this work, we categorize existing approaches into two paradigms: 1) tracking-by-detection, and 2) joint detection-and-tracking. **Tracking-by-Detection** has recently emerged as the dominant approach, attributed to its superior performance in benchmarks. Most conventional methods in this paradigm adopt object detection models to predict frame-level outputs, *i.g.*, masks and categories. They then train an additional tracking module using two adjacent frames to yield discriminative re-identification features (Yang, Fan, and Xu 2019; Cao et al. 2020; Yang et al. 2021; Liu et al. 2021; Wu et al. 2022b; Ying et al. 2023) using contrastive learning algorithm such as (Pang et al. 2021). These methods achieve state-of-the-art scores on most benchmark datasets. However, during inference, they rely on post-processing algorithms with heuristics such as NMS and rule-based trajectory initialization. While MinVIS (Huang, Yu, and Anandkumar 2022) eliminates post-processing, it struggles with the association of complex trajectories.

Recently, GenVIS introduces a high-performance tracking-by-detection approach through **object query propagation**. The key strategy involves: 1) taking full proposals from frozen segmentation models in given frames, 2) propagating the output video-level tracking query as an input to the subsequent frames, and 3) learning various tracking scenarios during training leveraging multiple frames. Subsequent works, such as (Li et al. 2023b; Hannan et al. 2023; Zhang et al. 2023), also show strong performance on benchmarks. While these approaches are favored for mAP metric, they inherently lack in mask-quality as they depend on frame-independent proposals.

Joint Detection-and-Tracking approaches were initially introduced using Transformer-based offline architectures, benefiting from their capability to model sequences in an end-to-end learning framework. With numerous approaches embracing this paradigm, we further categorize them based on whether their encoder module processes in-

puts through spatio-temporal aggregation or independently for each frame. For instance, VisTR (Wang et al. 2020) and IFC (Hwang et al. 2021) extend the DETR (Carion et al. 2020) framework to the temporal axis, computing pixel correlations within a specified window and then predicting clip-level mask tracklets. On the other hand, models like Mask2Former-VIS (Cheng et al. 2021a), SeqFormer (Wu et al. 2022a), and VITA (Heo et al. 2022b) embed spatial information in a frame-independent manner, while leveraging the temporal information using the clip-level decoder module. Concretely, VITA is in between both aforementioned paradigms as it collects frame-level decoded object queries to build a temporal association.

Mask Propagation

Semi-supervised Video Object Segmentation (VOS) aims to predict the sequence of object masks in a video, given the mask for the target object in the annotated frames. As the target object is defined with a binary mask (not with categories), most architectures run in a class-agnostic manner. Fundamentally, this task formulation has led to separate development paths between VOS and VIS. Specifically, VOS has developed to focus on modeling spatio-temporal pixel-level correlation to capture the appearance of the target object (Oh et al. 2019; Cheng, Tai, and Tang 2021; Cheng and Schwing 2022; Seong, Hyun, and Kim 2020; Seong et al. 2021; Yang, Wei, and Yang 2021; Yang and Yang 2022), while VIS has advanced alongside detection models that emphasize high-level semantics and object locations.

Although the two communities have developed differently, we find a clear consistency between the recent query propagation approach in VIS and the standard VOS methods. In the query propagation VIS approach, output object queries, which implicitly encapsulate corresponding instance *masks*, can be directly used to mine future mask outputs. Consequently, from the standpoint of VOS, the masks predicted from historical frames can be viewed as target conditions for subsequent frames. As a result, our approach can be interpreted as identifying a sub-problem of VOS, within the VIS framework.

High-quality Segmentation

Compared to the advancement in methods for high-quality image segmentation, encompassing both instance segmentation (Kirillov et al. 2020; Ke et al. 2022a; Wen et al. 2023) and class-agnostic segmentation (Ke et al. 2023; Qi et al. 2023), the progression within the VIS field appears relatively limited. VMT (Ke et al. 2022b) addresses this gap by introducing a top-down approach. Starting with initial coarse proposals derived from an established offline VIS model (Wu et al. 2022a), it refines the mask output by incorporating an extra module that models temporal coherence.

Method

We present RoCoVIS, a joint detection-and-tracking framework designed for online Video Instance Segmentation. We begin by revisiting the previous query propagation VIS architecture and introducing *direct query decoding* (Fig. 2 (b)), a simple yet effective skeleton of our method. Next, we introduce the core component, *instance propagation with masks* (gray colored region in Fig. 2 (c)). Lastly, we detail the training and inference processes for our framework.

Rethinking Query Propagation for VIS

Preliminary. We take GenVIS (Heo et al. 2023) as a representative method among the query-propagation methods (He et al. 2022; Zhang et al. 2023; Li et al. 2023b). GenVIS (Heo et al. 2023) is built upon a query-based Transformer architecture for segmentation (Cheng et al. 2022) with an additional query propagation module (Heo et al. 2022b). As illustrated in Fig. 2 (a), it consists of five main components: (1) backbone, (2) pixel decoder, (3) Transformer decoder, (4) object encoder, and (5) object decoder. It employs two types of object queries: frame-level and video-level. The frame-level query engages in generating frame-independent predictions through the first three modules. Specifically, the **pixel decoder** progressively upscales the **backbone** features into high-resolution per-pixel embeddings. The **Transformer decoder** subsequently processes a set of frame-level object queries referring to the per-pixel embeddings. Then, the video-level query is propagated through the **object decoder**, which associates the frame-level object queries encoded by the **object encoder**. Finally, each updated video-level query predicts the corresponding mask along with its category label.

Note that this decoupled propagation allows for freezing the frame-level detector (the first three modules), and thus enjoys parameter-efficient training across multiple frames. However, this method also leads to a disconnection between the propagated query and the pixel embeddings, which limits the opportunity for refining the detection quality.

Direct Query Decoding. To address the problem, we propose to replace the object decoder with direct query decoding onto pixel features. This simple change establishes a foundation for the direct incorporation of previous outcomes into current predictions. Specifically, we assign a unique instance ID to each object query, which is then propagated through the Transformer decoder to preserve the identity. Formally, at the t^{th} frame of the video, the output query embeddings from the previous frame, denoted as $\mathbf{Q_{t-1}}$, are used as inputs to the Transformer decoder to obtain $\mathbf{Q_t}$.

$$\mathbf{Q_t} = \mathcal{D}(\mathbf{Q_{t-1}}, \mathcal{P}(\mathbf{F}_t)), \tag{1}$$

where $\mathcal{P}$ is the pixel decoder that performs the feature encoding process on the current frame features $\mathbf{F}_t$ and $\mathcal{D}$ denotes the Transformer decoder.

It is important to highlight that in our approach, the object queries are directly propagated to the Transformer decoder. Our design choice allows the propagated object queries to refer directly to the pixel-level embeddings — rather than C-dimensional frame-level object queries — of the current frame. Such interaction enables the output from previous frames to enhance the mask predictions.

We empirically find that this formulation (Eq. (1)) shows improved results in mask quality (see GenVIS *vs* Direct

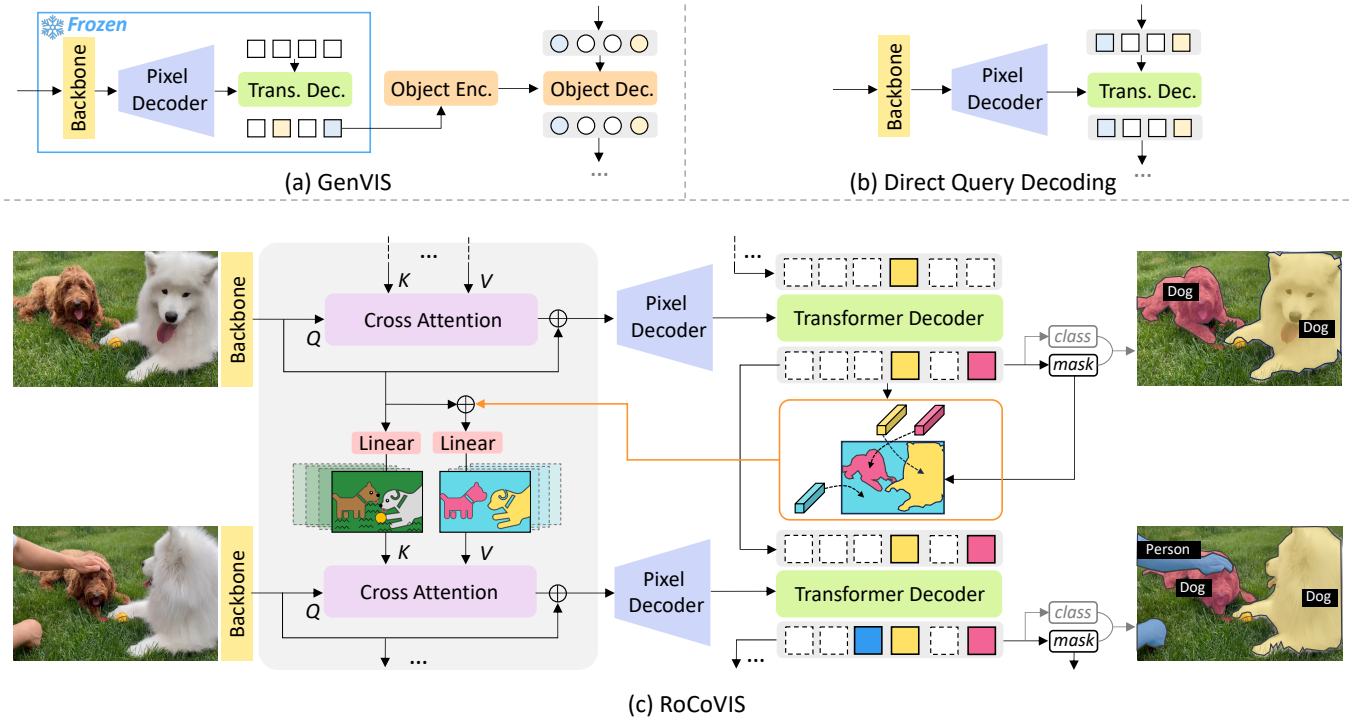

Figure 2: **Overview.** (a) While the previous query propagation approach employs a disjoint architecture between detection and tracking, (b) we introduce direct query decoding that allows the propagated object queries to interact directly with pixel embeddings. (c) Building on this, our RoCoVIS leverages mask predictions to encode object-aware feature correlations to achieve robust consistency.

Query Decoding in Tab. 2). However, we observe two limitations in the Transformer decoder that affect the consistency of object representation. First, the image features of the current frame denoted as $\mathbf{F}_t$, lack information from the previous frame, while the propagated object queries now directly reference the spatial information in the *cross-attention* layer. Indeed, the image features are already decoded by the pixel-decoder in a solely frame-independent manner, which is sub-optimal in terms of pixel-level consistency. Second, the interaction (where to find the new object?) relies exclusively on the *self-attention* within sparse object queries for the object queries that attempt to detect new objects (ideally mutually exclusive from previously appeared objects). This lacks the necessary spatial context for the effective identification of new objects, which is sub-optimal in terms of instance-level consistency.

Instance Propagation with Masks

Instance Mask Propagation. To address the aforementioned problems, we introduce *instance mask propagation*, calibrating the per-pixel embeddings of the current frame by utilizing object-aware information from previous sequences. Our approach involves efficient and effective feature calibration right at the beginning of the pixel decoder. We strategically map sparse object queries onto dense pixel embeddings by utilizing the predicted mask predictions, which form the essence of our work.

Given the output object query $\mathbf{Q_{t-1}} \in \mathbb{R}^{N\times C}$, we generate corresponding binary mask predictions $\mathbf{M_{t-1}} \in [0,1]^{N\times H\times W}$ by convolving the object query with the final per-pixel embedding produced by the pixel decoder (Cheng et al. 2022). We then encode the per-segment identity features ("object query") into their respective spatial regions, *i.e.*, we compute *spatial identity* $\mathbf{Z} = \mathbf{Q_{t-1}}^\top \cdot \mathbf{M_{t-1}}$ where $\mathbf{Z} \in \mathbb{R}^{C\times H\times W}$.

To further encode the information of regions without any detected object, we employ a vector $\mathbf{B} \in \mathbb{R}^{1\times C}$ and replace it into pixels not assigned to any foreground object.

$$\mathbf{Z'}_{:,h,w} = \mathbf{Z}_{:,h,w} + \mathbf{B}\times 1\{\sum_{c=1}^{C} \mathbf{Z}_{c,h,w} = 0\}, \quad \forall (h,w). \quad (2)$$

After that, we build a cross-attention layer denoted as $\texttt{CA}(\textbf{Query}, \textbf{Key}, \textbf{Value})$. Taking the image features of the current frame $\mathbf{F}_t$ as **Query**, we feed previous image features as **Key** and **Value**, with the latter being augmented by the residual connection to the spatial identity features $\mathbf{Z'}_{t-1}$. This layer captures the correlations between the current and the previous image features, gathering the identity information carried by the **Value**. The refined image features for the current frame, denoted as $\mathbf{F'}_t$, are then obtained through the following computation:

$$\mathbf{F'}_t = \texttt{CA}(\mathbf{F}_t, \mathbf{F}_{t-1}, \mathbf{F}_{t-1} + \mathbf{Z'}_{t-1}). \quad (3)$$

Finally, the Eq. (1) is simply changed to the equation below with residual path:

$$\mathbf{Q_t} = \mathcal{D}(\mathbf{Q_{t-1}}, \mathcal{P}(\mathbf{F}_t + \mathbf{F'}_t)). \quad (4)$$

To mitigate the discrepancy between the first frame and the following frames, we set $\mathbf{F'}_0 = \texttt{CA}(\mathbf{F}_0, \mathbf{F}_0, \mathbf{F}_0 + \mathbf{E})$, where $\mathbf{E} \in \mathbb{R}^{C \times H \times W}$ is a tensor consisting of $\mathbf{B}$ at all pixel coordinates. In our implementation, we preserve **Key** and **Value** for recent time steps and concatenate them to enrich sequential information.

Training. For the matching criterion between predictions and ground truth, we adopt the Unified Video Label Assignment (UVLA) (Heo et al. 2023) with a minor modification, which we refer to as UVLA++. The alteration considers the tracking scenario where a certain object query disappears after it has been detected. In the original UVLA, such instances are treated as "occupied", with the corresponding target labels assigned as "no object". Conventionally, following the loss computation from (Cheng et al. 2022), only the classification loss is calculated, while the mask loss is ignored. However, our approach underscores the importance of using the output prediction mask in subsequent frames to suppress the propagation of redundant masks. Therefore, we impose a constraint to predict an empty mask for the queries corresponding to the state of a "disappeared" object, preventing any unnecessary mask propagation. Additionally, we find that limiting the propagation of prediction masks solely to matched object queries enhances performance. We simply suppress the masks of unmatched object queries by setting them to zero before computing the spatial identity features. This training strategy along with the instance mask propagation ensures both pixel-level (improved AP^B) and instance-level (improved AP^{Pool}) consistencies.

Inference. During inference, we refine the pixel probabilities using the class probabilities to dynamically adjust the valid outputs. Specifically, we first calculate the foreground probabilities for each segment by simply subtracting one from the class probability of the "no object" token ($\emptyset$). Then, we scale the output pixel probability by the foreground score before binarization.

Experiments

Datasets

YouTube-VIS. YouTube-VIS (Yang, Fan, and Xu 2019) is a standard benchmark dataset for VIS in three versions (2019/2021/2022). Each version is designed to segment objects from 40 predefined categories within videos. While the 2021 and 2022 datasets share an identical training set, the 2022 version uniquely adds 71 additional videos to the validation set of 2021.

OVIS. Occluded-VIS (OVIS) dataset (Qi et al. 2021) has been introduced to specifically address the challenging scenario of heavy occlusions between objects. Furthermore, the dataset presents challenging tracking scenarios in long videos, extending over hundreds of frames.

HQ-YTVIS. The mask annotations in the aforementioned VIS benchmarks are not sufficiently detailed to accurately measure the quality of mask predictions. To address this and benchmark high-quality VIS, HQ-YTVIS (Ke et al. 2022b) refines mask annotation of YouTube-VIS 2019 by self-correcting the coarsely annotated data using a novel Video Mask Transfiner (VMT) architecture.

VIPSeg. We utilize VIPSeg (Miao et al. 2022), introduced for Video Panoptic Segmentation (VPS) (Kim et al. 2020). Since VPS focuses on pixel-level classification, it typically includes high-quality mask annotations. We follow the original data split, while simply converting the "things" annotations into VIS annotations. Throughout the rest of the paper, we refer to this subset of annotations as "VIPSeg-things".

Metric

Video AP. Average Precision (AP) (Yang, Fan, and Xu 2019) is the standard metric for VIS which is an extension of the image instance segmentation metric (Lin et al. 2014). It calculates the spatio-temporal mask IoU at the trajectory level, comparing the ground truth with the predictions.

Tube-Boundary AP. While the standard AP effectively captures essential aspects of the task, including segmentation, categorization, and localization, it is not sensitive to the fidelity of masks. To measure the accuracy of mask boundaries more sensitively, Boundary AP has been proposed in the image domain (Cheng et al. 2021b). This metric focuses on the pixels near the boundary contours, calculating the IoU for mask pixels within a specified distance from these contours of the ground truth and predictions. Building on this to the video domain, (Ke et al. 2022b; Heo et al. 2022a) define Tube-Boundary AP to assess video-level mask quality. To analyze the pixel-level mask consistency, we provide Tube-Boundary AP (AP^B) using the official implementation (Ke et al. 2022b).

AP^{Pool}. The hidden key to high performance in VIS on the standard AP lies in accurate classification since it evaluates each category independently. In the current evaluation system, each trajectory comprising outputs from multiple frames is assigned a single category label along with a confidence score. However, this system presents a challenge: some false positives with low confidence scores can be beneficial for the overall score than true positives with higher confidence. This issue becomes more pronounced in the video domain due to fewer validation data samples compared to the image. Interestingly, the standard AP might favor models that generate multiple outputs while it is not consistent with practical benefits in real-world applications. To address this, we adopt AP^{Pool} from (Dave et al. 2021), which aggregates detections from all classes to compute a unified precision-recall curve. This method ensures that true positives across all classes rank higher than any class's false positives. As a result, it mitigates the favor towards multiple predictions and better highlights the precise predictions.

We note that the standard VIS benchmarks (YouTube-VIS and OVIS) do not provide access to ground-truth annotations for the validation set. Therefore, we present the results of two newly adopted metrics on HQ-YTVIS and VIPSeg.

Quantitative Results

YouTube-VIS 2019 & 2021. Due to space constraints, detailed results are provided in the supplementary material.

Backbone	Method	YouTube-VIS 2022					OVIS				
		AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50	MinVIS (Huang, Yu, and Anandkumar 2022)	23.3	47.9	19.3	20.2	28.0	25.0	45.5	24.0	13.9	29.7
	IDOL (Wu et al. 2022b)	-	-	-	-	-	30.2	51.3	30.0	15.0	37.5
	GenVIS$_{online}$ (Heo et al. 2023)	37.5	**61.6**	41.5	32.6	42.2	**35.8**	**60.8**	36.2	**16.3**	39.6
	CTVIS (Ying et al. 2023)	-	-	-	-	-	35.5	**60.8**	34.9	16.1	**41.9**
	TCOVIS (Li et al. 2023b)	**38.6**	59.4	41.6	**32.8**	**46.7**	35.3	60.7	**36.6**	15.7	39.5
	DVIS$_{online}$ (Zhang et al. 2023)	-	-	-	-	-	31.0	54.8	31.9	15.2	37.6
	RoCoVIS (ours)	**38.6**	57.0	**45.1**	**32.8**	41.4	35.1	57.9	34.6	15.8	41.0
Swin-L	MinVIS (Huang, Yu, and Anandkumar 2022)	33.1	54.8	33.7	29.5	36.6	39.4	61.5	41.3	18.1	43.3
	IDOL (Wu et al. 2022b)	-	-	-	-	-	42.6	65.7	45.2	17.9	49.6
	GenVIS$_{online}$ (Heo et al. 2023)	45.1	69.1	47.3	39.8	48.5	45.2	69.1	48.4	19.1	48.6
	CTVIS (Ying et al. 2023)	-	-	-	-	-	46.9	71.5	47.5	19.1	52.1
	TCOVIS (Li et al. 2023b)	51.0	**73.0**	53.5	41.7	**56.5**	46.7	70.9	49.5	19.1	50.8
	DVIS$_{online}$ (Zhang et al. 2023)	45.9	69.0	48.8	37.2	51.8	47.1	**71.9**	49.2	19.4	52.5
	RoCoVIS (ours)	**51.5**	72.8	**56.5**	**42.0**	56.0	**48.5**	70.5	**51.3**	**19.5**	**53.6**

Table 1: Comparisons on **YouTube-VIS 2022 long videos and OVIS *val* sets**: we highlight the best value in **bold**.

Method	HQ-YTVIS *val.*								
	AP^B	AP^B_{50}	AP^B_{75}	AR^B_1	AP	AP_{50}	AP_{75}	AR_1	AP^{Pool}
MinVIS† (Huang, Yu, and Anandkumar 2022)	34.1	64.0	33.3	34.1	49.6	69.2	53.2	47.5	49.3
GenVIS† (Heo et al. 2023)	35.6	65.3	36.5	35.6	51.4	68.9	55.1	49.5	50.1
Direct Query Decoding	37.0	67.2	37.4	35.7	53.5	72.2	57.5	50.1	54.3
+ Instance Mask Prop.	37.5	65.6	37.5	36.6	52.3	70.8	55.3	49.9	52.7
+ Label Assign.	39.0	65.7	40.6	37.5	54.1	70.3	59.8	50.7	54.8

Table 2: Ablation study on **HQ-YTVIS *val* set** with ResNet-50 backbone. † is reproduced in our study using their official codes.

YouTube-VIS 2022. Tab. 1 showcases our performance on the challenging benchmark, the long video validation split of YouTube-VIS 2022. This benchmark, characterized by complex object trajectories and identity crossing in long videos, highlights the robustness of our method in challenging scenarios. With both ResNet-50 and Swin-L backbones, we achieve state-of-the-art results, comparable to those of TCOVIS. Importantly, we exhibit a notably higher AP_{75} compared to TCOVIS, underscoring our model's ability to generate high-fidelity mask predictions with fewer redundancies. We hypothesize that tracking-by-detection approaches are inclined to benefit from inconsistent predictions in per-frame detection, a strategy that contributes to the overall AP due to an elevated AP_{50}.

OVIS. We also present our results on the OVIS dataset in Tab. 1, which features highly occluded instances across long videos. Consistent with our findings on YouTube-VIS, the Swin-L backbone demonstrates strong scalability, yielding significant performance improvements over existing methods on challenging benchmarks. We attribute this success to the advantages of end-to-end sequential modeling. Notably, RoCoVIS is the first online joint detection-and-tracking method to showcase such strong performance, highlighting the potential of new approaches in the literature.

HQ-YTVIS & VIPSeg-things. We demonstrate that RoCoVIS produces high-quality consistent mask outputs in Tab. 3. For a deeper understanding, we also present reproduced results for two prominent online VIS methods based on Mask2Former (Cheng et al. 2022): MinVIS (Huang, Yu, and Anandkumar 2022), and GenVIS (Heo et al. 2023).
Our method outperforms others in all metrics with large margins. Notably, we also chart the boundary mask AP at various IoU thresholds on HQ-YTVIS in Fig. 3. Despite using the same image segmentation backbone as MinVIS and GenVIS, our method achieves higher AP^B at elevated IoU thresholds, underscoring the effectiveness of our approach in achieving our primary objective: pixel-level consistency.

Ablation Study

In this section, we analyze the main components of RoCoVIS and evaluate their impact in Tab. 2.

Direct Query Decoding: Direct Query Decoding effectively eliminates the bottleneck, enabling the propagated object query to directly reference the current pixel embeddings. We hypothesize that the dataset originates from YouTube-VIS 2019, characterized by relatively small instances in the scene, where removing the bottleneck is advantageous for refining predictions. Consequently, this approach consistently outperforms GenVIS across all metrics.

Instance Mask Propagation: We observe a modest increase in AP^B (by +0.5) but a decrease in general stability in overall detection performance (AP) when implementing instance mask propagation. We determined that the primary issue lies in the existing UVLA criterion (Heo et al. 2023), which is not tailored to address inconsistent mask predictions for temporarily disappeared objects. This limitation

Method	HQ-YTVIS *test*						VIPSeg-things *val*					
	AP^B	AP^B_{75}	AR^B_1	AP	AP_{75}	AR_1	AP^B	AP^B_{75}	AR^B_1	AP	AP_{75}	AR_1
SeqFormer‡ (Wu et al. 2022a)	28.6	21.4	29.3	48.5	52.2	48.5	-	-	-	-	-	-
VMT (Ke et al. 2022b)	30.7	24.2	31.5	50.5	54.5	50.2	-	-	-	-	-	-
MinVIS† (Huang, Yu, and Anandkumar 2022)	36.8	34.9	36.0	54.6	58.4	52.0	16.5	16.0	15.0	22.5	23.2	19.8
GenVIS† (Heo et al. 2023)	37.5	35.0	36.9	55.2	57.0	52.8	18.1	18.1	16.3	24.2	24.9	21.6
RoCoVIS (ours)	**39.6**	**42.8**	**39.4**	**57.2**	**61.6**	**55.1**	**19.2**	**20.6**	**17.5**	**24.9**	**25.8**	**21.8**

Table 3: Comparisons on **HQ-YTVIS *test*** and **VIPSeg-things *val*** with ResNet-50 backbone. ‡ is the reproduced result by (Ke et al. 2022b), and † is reproduced in our study.

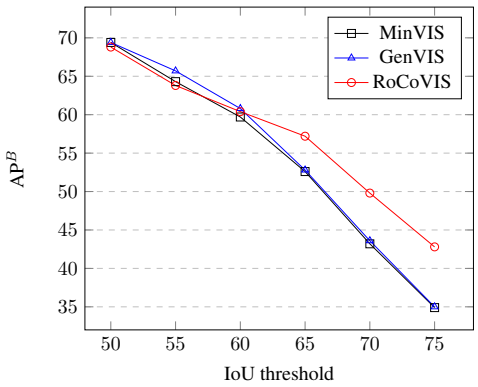

Figure 3: AP^B vs. IoU threshold on HQ-YTVIS.

leads to error propagation, adversely affecting the robustness of the detection performance.

UVLA++: Finally, by implementing UVLA++, we effectively suppress such errors. Aligning with the instance mask propagation, this leads to noticeable performance improvements across all metrics.

AP^{Pool}: Each component of our method consistently exhibits improvements in AP^{Pool}, in contrast to the standard AP whereas GenVIS demonstrates a decrease. Since AP^{Pool} eliminates the limitations of class-independent boundaries, our method, eliminating the bottleneck, gains an advantage in confidence score calibration and fewer redundancies.

Qualitative Results

Fig. 4 provides qualitative comparisons of our method against GenVIS, showcasing our contributions: 1) robust pixel-level mask consistency and 2) instance consistency. **In the first video**, GenVIS displays inconsistent mask predictions across frames. The inconsistency arises because GenVIS relies solely on frame-level predictions, which can vary under various ambiguous conditions within the same video. In contrast, our method demonstrates consistent mask predictions for the target object. **The second video** highlights the issue of redundant mask predictions in complex scenes by GenVIS. While GenVIS generates multiple mask predictions for the same individual, our method avoids overlapped masks, showcasing visually satisfying results.

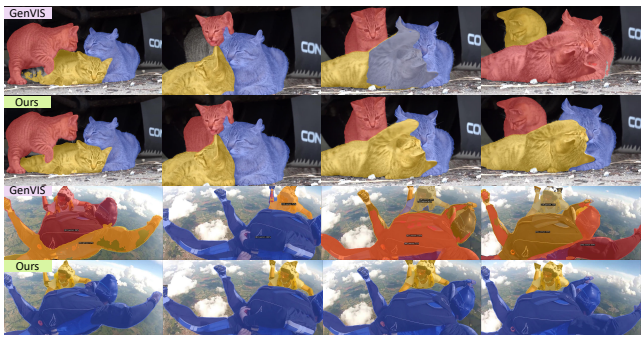

Figure 4: **Qualitative comparisons** with GenVIS.

Limitation

As we adopt end-to-end learning, our method entails a higher training cost compared to decoupled methods that employ minimal embedding space adjustments from pre-trained image-level detection models. Additionally, we recognize a potential area for further development in our current design, particularly regarding the integration of a large number of historical frames.

Conclusion

We present RoCoVIS, a video instance segmentation model designed to correct inconsistencies at both pixel and instance levels. By employing a straightforward query propagation framework enhanced by an instance mask propagation module, RoCoVIS surpasses leading models on high-quality VIS benchmarks, particularly highlighted in the newly employed metrics. Our model also achieves state-of-the-art performance with a large-scale backbone across challenging benchmarks, demonstrating its scalability and effectiveness in end-to-end sequential modeling. We hope our research will lead to discussions on the often-neglected yet vital aspects of VIS challenges, focusing on refining and accurately evaluating VIS performance for online video applications.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00457882, National AI Research Lab Project, and No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University))

References

Cao, J.; Anwer, R. M.; Cholakkal, H.; Khan, F. S.; Pang, Y.; and Shao, L. 2020. SipMask: Spatial Information Preservation for Fast Image and Video Instance Segmentation. In *ECCV*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*.

Cheng, B.; Choudhuri, A.; Misra, I.; Kirillov, A.; Girdhar, R.; and Schwing, A. G. 2021a. Mask2Former for Video Instance Segmentation. *arXiv preprint arXiv:2112.10764*.

Cheng, B.; Girshick, R.; Dollár, P.; Berg, A. C.; and Kirillov, A. 2021b. Boundary IoU: Improving object-centric image segmentation evaluation. In *CVPR*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *CVPR*.

Cheng, H. K.; and Schwing, A. G. 2022. XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model. In *ECCV*.

Cheng, H. K.; Tai, Y.-W.; and Tang, C.-K. 2021. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NeurIPS*.

Dave, A.; Dollár, P.; Ramanan, D.; Kirillov, A.; and Girshick, R. 2021. Evaluating large-vocabulary object detectors: The devil is in the details. *arXiv preprint arXiv:2102.01066*.

Hannan, T.; Koner, R.; Bernhard, M.; Shit, S.; Menze, B.; Tresp, V.; Schubert, M.; and Seidl, T. 2023. GRAtt-VIS: Gated Residual Attention for Auto Rectifying Video Instance Segmentation. *arXiv preprint arXiv:2305.17096*.

He, F.; Zhang, H.; Gao, N.; Jia, J.; Shan, Y.; Zhao, X.; and Huang, K. 2022. Inspro: Propagating instance query and proposal for online video instance segmentation. In *NeurIPS*.

Heo, M.; Hwang, S.; Hyun, J.; Kim, H.; Oh, S. W.; Lee, J.-Y.; and Kim, S. J. 2023. A generalized framework for video instance segmentation. In *CVPR*.

Heo, M.; Hwang, S.; Oh, S. W.; Lee, J.-Y.; and Kim, S. J. 2022a. Integrating Pose and Mask Predictions for Multi-person in Videos. In *CVPRW*.

Heo, M.; Hwang, S.; Oh, S. W.; Lee, J.-Y.; and Kim, S. J. 2022b. VITA: Video Instance Segmentation via Object Token Association. In *NeurIPS*.

Huang, D.-A.; Yu, Z.; and Anandkumar, A. 2022. MinVIS: A Minimal Video Instance Segmentation Framework without Video-based Training. In *NeurIPS*.

Hwang, S.; Heo, M.; Oh, S. W.; and Kim, S. J. 2021. Video instance segmentation using inter-frame communication transformers. In *NeurIPS*.

Ke, L.; Danelljan, M.; Li, X.; Tai, Y.-W.; Tang, C.-K.; and Yu, F. 2022a. Mask transfiner for high-quality instance segmentation. In *CVPR*.

Ke, L.; Ding, H.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; and Yu, F. 2022b. Video mask transfiner for high-quality video instance segmentation. In *ECCV*.

Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F.; et al. 2023. Segment anything in high quality. In *NeurIPS*.

Kim, D.; Woo, S.; Lee, J.-Y.; and Kweon, I. S. 2020. Video panoptic segmentation. In *CVPR*.

Kirillov, A.; Wu, Y.; He, K.; and Girshick, R. 2020. Pointrend: Image segmentation as rendering. In *CVPR*.

Li, F.; Zhang, H.; Xu, H.; Liu, S.; Zhang, L.; Ni, L. M.; and Shum, H.-Y. 2023a. Mask DINO: Towards a Unified Transformer-Based Framework for Object Detection and Segmentation. In *CVPR*.

Li, J.; Yu, B.; Rao, Y.; Zhou, J.; and Lu, J. 2023b. TCOVIS: Temporally Consistent Online Video Instance Segmentation. In *ICCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Liu, D.; Cui, Y.; Tan, W.; and Chen, Y. 2021. SG-Net: Spatial Granularity Network for One-Stage Video Instance Segmentation. In *CVPR*.

Miao, J.; Wang, X.; Wu, Y.; Li, W.; Zhang, X.; Wei, Y.; and Yang, Y. 2022. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*.

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019. Video Object Segmentation Using Space-Time Memory Networks. In *ICCV*.

Pang, J.; Qiu, L.; Li, X.; Chen, H.; Li, Q.; Darrell, T.; and Yu, F. 2021. Quasi-dense similarity learning for multiple object tracking. In *CVPR*.

Qi, J.; Gao, Y.; Hu, Y.; Wang, X.; Liu, X.; Bai, X.; Belongie, S.; Yuille, A.; Torr, P. H.; and Bai, S. 2021. Occluded video instance segmentation. *arXiv preprint arXiv:2102.01558*.

Qi, L.; Kuen, J.; Shen, T.; Gu, J.; Li, W.; Guo, W.; Jia, J.; Lin, Z.; and Yang, M.-H. 2023. High Quality Entity Segmentation. In *ICCV*.

Seong, H.; Hyun, J.; and Kim, E. 2020. Kernelized memory network for video object segmentation. In *ECCV*.

Seong, H.; Oh, S. W.; Lee, J.-Y.; Lee, S.; Lee, S.; and Kim, E. 2021. Hierarchical memory matching network for video object segmentation. In *ICCV*.

Wang, Y.; Xu, Z.; Wang, X.; Shen, C.; Cheng, B.; Shen, H.; and Xia, H. 2020. End-to-End Video Instance Segmentation with Transformers. In *CVPR*.

Wen, Q.; Yang, J.; Yang, X.; and Liang, K. 2023. Patchdct: Patch refinement for high quality instance segmentation. In *ICLR*.

Wu, J.; Jiang, Y.; Zhang, W.; Bai, X.; and Bai, S. 2022a. Seqformer: a frustratingly simple model for video instance segmentation. In *ECCV*.

Wu, J.; Liu, Q.; Jiang, Y.; Bai, S.; Yuille, A.; and Bai, X. 2022b. In Defense of Online Models for Video Instance Segmentation. In *ECCV*.

Yang, L.; Fan, Y.; and Xu, N. 2019. Video instance segmentation. In *ICCV*.

Yang, S.; Fang, Y.; Wang, X.; Li, Y.; Fang, C.; Shan, Y.; Feng, B.; and Liu, W. 2021. Crossover Learning for Fast Online Video Instance Segmentation. In *ICCV*.

Yang, Z.; Wei, Y.; and Yang, Y. 2021. Associating objects with transformers for video object segmentation. In *NeurIPS*.

Yang, Z.; and Yang, Y. 2022. Decoupling features in hierarchical propagation for video object segmentation. In *NeurIPS*.

Ying, K.; Zhong, Q.; Mao, W.; Wang, Z.; Chen, H.; Wu, L. Y.; Liu, Y.; Fan, C.; Zhuge, Y.; and Shen, C. 2023. Ctvis: Consistent training for online video instance segmentation. In *ICCV*.

Zhang, T.; Tian, X.; Wu, Y.; Ji, S.; Wang, X.; Zhang, Y.; and Wan, P. 2023. DVIS: Decoupled Video Instance Segmentation Framework. In *ICCV*.