A Generalized Framework for Video Instance Segmentation

Miran Heo[1]   Sukjun Hwang[1]   Jeongseok Hyun[1]   Hanjung Kim[1]

Seoung Wug Oh[2]   Joon-Young Lee[2]   Seon Joo Kim[1]

[1]Yonsei University   [2]Adobe Research

Abstract

*The handling of long videos with complex and occluded sequences has recently emerged as a new challenge in the video instance segmentation (VIS) community. However, existing methods have limitations in addressing this challenge. We argue that the biggest bottleneck in current approaches is the discrepancy between training and inference. To effectively bridge this gap, we propose a **Gen**eralized framework for **VIS**, namely **GenVIS**, that achieves state-of-the-art performance on challenging benchmarks without designing complicated architectures or requiring extra post-processing. The key contribution of GenVIS is the learning strategy, which includes a query-based training pipeline for sequential learning with a novel target label assignment. Additionally, we introduce a memory that effectively acquires information from previous states. Thanks to the new perspective, which focuses on building relationships between separate frames or clips, GenVIS can be flexibly executed in both online and semi-online manner. We evaluate our approach on popular VIS benchmarks, achieving state-of-the-art results on YouTube-VIS 2019/2021/2022 and Occluded VIS (OVIS). Notably, we greatly outperform the state-of-the-art on the long VIS benchmark (OVIS), improving 5.6 AP with ResNet-50 backbone. Code is available at* https://github.com/miranheo/GenVIS.

1. Introduction

Video Instance Segmentation (VIS) is the task of identifying, segmenting, and tracking all objects in videos simultaneously. With the emergence of datasets containing long and complex sequences, the research community is taking a step towards real-world applications. While many papers have proposed solutions, the most notable performance improvement has been achieved by recent online methods using image-based backbones [14, 34]. These results challenge the common belief that end-to-end semi-online or offline approaches (*i.e.*, [5, 13, 15, 30, 33, 38]) trained on longer video clips would better model long-range object relationships.

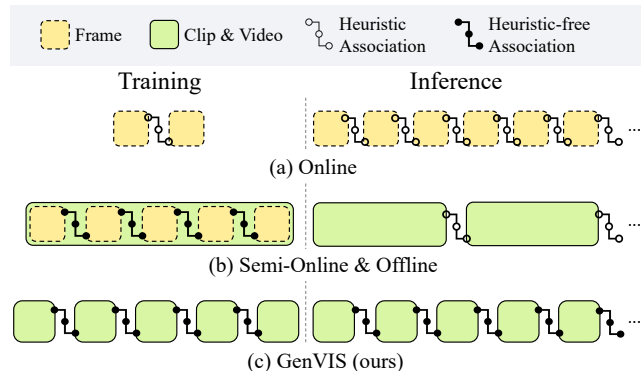We hypothesize that reason behind this somewhat surpris-

Figure 1. Comparison between current VIS paradigms and our approach. (a, b) While current methods use two separate paradigms based on the number of frames processed, we argue that the key challenge in processing real-world videos is building inter-clip associations. (b) Our proposed **GenVIS** addresses this challenge and can operate effectively in both online and semi-online manner without requiring hand-crafted post-processing.

ing result is the presence or absence of an object association scheme between frames or clips that can scale to long videos. Recent VIS methods, regardless of the approach, are driven by powerful image-level detectors [6, 8], so detection and segmentation quality are already robust and comparable to each other. To operate robustly in long videos, what the VIS really needs to focus on is the long-range tracking quality. Although semi-online and offline methods are suitable for tracking objects within clips, they need to associate objects between clips to infer long videos, which is usually achieved using simple heuristics such as IoU matching [1, 15].

While online methods are more robust than semi-online/offline VIS solutions in processing long videos, we still see a significant room for improvement in their tracking approach. These methods only consider local contexts between adjacent frames during training, while test videos can exceed hundreds of frames [27]. We believe that there is a better way to learn long-range temporal modeling that could fundamentally change the current VIS landscape.

In this paper, we argue that the biggest bottleneck in handling long videos is the discrepancy between the training and inference scenarios. Regardless of the previously defined paradigms (e.g., how many frames a method processes at

once), we need to focus on how to train the model. As illustrated in Fig. 1, all previous methodologies use only a few frames or clips (*e.g.*, one or two) for training, while real-world videos can have an unlimited length. Therefore, they have to handle typical long-range tracking scenarios (*e.g.*, newborn objects and re-identification) as exceptions through heuristics [34].

We introduce a **Gen**eralized **VIS** framework, namely **GenVIS** (Fig. 1 (c)), that is designed to *minimize* the gap between training and inference of long videos. We take an existing offline VIS model (VITA [13]) as the backbone and applay a query-propagation method [32] for object association between clips. The essence of GenVIS is the novel training strategy. By improving the training strategy of the base model, we achieve significant gains of 5.1 AP (Occluded VIS [27]) and 5.8 AP (YouTube-VIS 2022 Long Videos [36]) in long and challenging benchmarks, outperforming all the previous methods by a large margin.

Our first proposal for improvement is to load multiple clips during training. Unlike previous semi-online/offline VIS methods [1, 5, 15, 30, 33] that focus on placing multiple frames in a single clip to strengthen intra-clip tracking, we propose to prioritize inter-clip tracking by learning the temporal relationship through multiple consecutive clips. We believe that strong inter-clip reasoning is crucial for processing long videos in the real world. As videos must be split into multiple clips that fit into GPU memory when processing long videos, inter-clip association is inevitable. In this work, we use relatively short clip lengths (*e.g.*, 1 to 7), but load as many clips as possible (*e.g.*, usually more than 5).

More importantly, we propose a new learning criterion that enables seamless association through multiple consecutive clips. Since we now deploy a sufficient number of clips, we can effectively simulate various inference scenarios at training time, covering newborn objects and objects that disappear and reappear. Specifically, we propose the Unified Video Label Assignment (UVLA) that allows unique object queries to detect newly-appeared object and to keep them consistent once matching identities are obtained. Our new learning criterion not only improves tracking performance but also removes all heuristics[1] required to handle new objects and re-identification from the inference stage. In other words, our model infers videos exactly as it learned. With these two proposals in the learning strategy, our base model outperforms all the previous methods on long video VIS benchmarks [27] without additional network modules.

To further bridge the remaining gap between training and inference, we propose adopting a memory mechanism that stores previously decoded object queries. This mechanism is particularly useful for handling very long videos (or streaming video), where there is a limit to the number of clips that can be loaded at once. To implement this, we add extra information for each object query by reading from its previous states. The memory mechanism results in meaningful improvements with only a small computational overhead.

Despite its simple framework, GenVIS achieves state-of-the-art results on VIS benchmarks, outperforming previous methods on challenging long and complex video datasets (Occluded VIS [27] and YouTube-VIS 2022 [36]). Our method also demonstrates strong generalization capability under online and semi-online settings[2]. We provide additional analysis of the training and inference settings, which can be useful for balancing accuracy and efficiency tradeoffs.

2. Related Works

2.1. Video Instance Segmentation

We categorize previous studies into two paradigms: 1) *online*, and 2) *semi-online & offline*. In this paper, we integrate the offline methods with the semi-online paradigm.

Online methods have been making considerable progress through advances in image-level object detection algorithms. MaskTrack R-CNN [36] was the first attempt on the VIS task that puts its basis on the image instance segmentation model [11]. For the extension to the video domain, MaskTrack R-CNN [36] and follow-up works [3, 22, 37] predict frame-independent outputs and make association using post-processing during the inference stage. Recently, MinVIS [14] suggested that a strong, query-based image instance segmentation model [6] inherently embeds distinctions of objects, and demonstrated competitive performance without video-based training. Based on Deformable DETR [40], IDOL [34] added a contrastive head that learns discriminative instance features between paired frames.

Inspired by Video Object Segmentation (VOS) methods [25], some previous works [9, 10, 17, 20] adopted propagation approach. VISOLO [10] and PCAN [17] integrated intermediate predictions and features to utilize the concept of memory [25] for better performance.

Semi-online & Offline paradigms leverage multiple frames to take advantage of the rich temporal context. VisTR [30] adopted [4, 29] and introduced the first end-to-end VIS model by taking a full video as an input and yielding video-level mask trajectories. IFC [15] devised inter-frame communication that alleviates the heavy computation of VisTR and effectively encodes clip-wise information. SeqFormer [33] proposed an architecture that aggregates spatio-temporal contexts, iteratively decomposing multi-level features of clips using object queries. VITA [13] introduced a new offline paradigm, showing that video-level

---

[1]Previously, [32] need to determine whether a tracked query is valid or not with a confidence threshold.

[2]The term 'general' describes how our proposed framework is designed to be versatile and able to operate in both online and semi-online manner. This flexibility is achieved by adjusting the clip length to 1, which allows us to convert the semi-online model into an online (per-frame) VIS model.
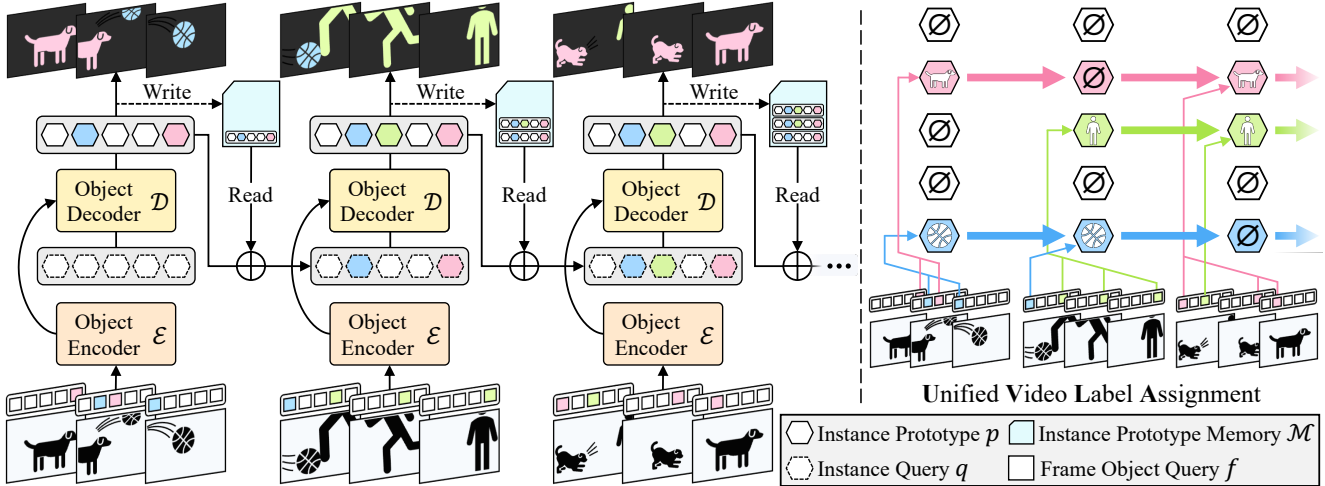
Figure 2. (Left) Overview of our framework. (Right) Unified Video Label Assignment. By presenting this simple framework with novel ground-truth assignment strategy that resembles the inference pipeline, we significantly improve the accuracy on the VIS task without any bells-and-whistles. $\varnothing$: indicates "no object" as not matched to a ground-truth.

scene understanding can be achieved by building temporal interaction between frame-level object queries.

Several methods [1,15,18] take clip-level input and run sequentially with association algorithm during post-processing. However, since the emergence of long video benchmarks [27], most existing offline methods [5,15,30,33] cannot handle these benchmarks in an end-to-end manner as GPU memories cannot hold more than hundreds of frames. Although IFC [15] proposed a post-processing algorithm to match clip-wise outputs, it depends on IoU between mask predictions of intersecting frames, thus computationally heavy. EfficientVIS [32] proposed a method that stitches clip-wise outputs with minimal heuristics by making the model itself learn to infer over sequential clips.

2.2. Multi-Object Tracking

Query-based Trackers [2,24,39] adopted query-based object detectors [4,40] to track objects across frames with bounding boxes [16]. TrackFormer [24] follows a DETR-like architecture where newly appeared objects in a video are detected with object queries, but track queries are used to re-identify target objects which were detected and tracked in previous frames. MeMOT [2] and MOTR [39] also utilized track queries, but they enhance the training approach compared to TrackFormer by incorporating multiple frames and introducing additional modules that make use of previous track queries. Nonetheless, these models entail complex heuristics with several handcrafted threshold values since both object and track queries are responsible for detecting objects in a scene, leading to duplicate detections.

Clip-based Tracker [31] proposed a novel approach to improve long-term association by reformulating the standard tracking-detection scheme as a clip-to-clip matching problem. Their observations are in line with our own, which

suggest that focusing on inter-clip association is important for effectively utilizing temporal information.

3. Method

We propose a novel generalized video instance segmentation framework GenVIS that sequentially associate clip-wise predictions. To this end, we first split an input RGB video into N_v non-overlapping clips $\{\mathcal{V}^i \in \mathbb{R}^{N_f \times H \times W \times 3}\}_{i=1}^{N_v}$, where each clip $\mathcal{V}$ consists of N_f consecutive frames. Here, N_f is an integer less than the length of the input video; if $N_f = 1$, our model runs in an online manner, otherwise semi-online. Sequentially taking a clip, GenVIS comes out with clip-level output queries that are used for predicting mask tracklets with a class probability for the input clip. Then, the clip-level output queries from the current clip again become input object queries for the next clip. Attributed to our newly defined label assignment strategy UVLA, GenVIS can 1) incorporate multiple clips (*e.g.,* five clips) during training, which leads to better modeling real-video characteristics, and 2) narrow the gap between training and inference as the inter-clip associations can also be trained. We also integrate information from past historical object queries by adopting a memory mechanism, which further improves the prediction qualities. Finally, GenVIS presents *significant accuracy* while completely *removing heuristic matching algorithms*.

3.1. Instance Prototypes

As suggested in recent studies [13,14], we hypothesize that a unique object is temporally coherent over a short period of time. Therefore, an object can be compressed into a concise representation, *instance prototype*. Simply, instance prototypes are identical to output object queries that are used for predicting masks and categories in successful

Transformer-based [29] offline VIS architectures [5,15,33]. However, such architectures have an underlying problem that only a limited number of clips can be loaded at training time, because they refer to dense spatio-temporal features to decode object queries. For example, Mask2Former-VIS [5] requires high-end GPUs with large memory capacity; in a single NVIDIA V100 GPU, 2 clips whose length is 2 frames (4 frames in total) are loaded for training.

To resolve the limitation, we adopt recently proposed VITA [13] (Fig. 2 (left)), which suggests that a video can be sufficiently represented with a collection of object-centric vectors. Given a clip $\mathcal{V}$, we discard all features but the collection of $N_f \times N_q$ frame object queries (f) from a per-frame detector [6]. After building relations within f using *Object Encoder* ($\mathcal{E}$), we aggregate temporal contexts by inserting N_q instance queries (q) into *Object Decoder* ($\mathcal{D}$) and obtain instance prototypes (p) as $p = \mathcal{D}(q, \mathcal{E}(f))$. The instance prototypes are forwarded into prediction heads that output clip-wise mask and class predictions $\{\hat{y}_k\}_{k=1}^{N_q}$.

As illustrated in Fig. 2 (left), we only train the four components that are disjoint from the backbone features: $\mathcal{E}$, $\mathcal{D}$, q, and p. This leads to efficient memory consumption; more than 30 clips of 3 frames long ($\sim$100 frames) can be used for training even with a single RTX 3090 GPU. In the VIS benchmark datasets, videos are sampled at 6 FPS; thus, 100 frames represent a video clip of 17 seconds long. We find this strategy of using numerous clips during training is essential in designing long-range relationships and modeling real-video characteristics. More training details are discussed in Sec. 4.2.

3.2. GenVIS

After obtaining instance prototypes from each clip, the next challenge is: how to associate predictions from adjacent clips. Previous studies [1, 14, 15] bipartitely match such separate predictions using scores gauged under customized measurements, and they show competitive performance in relatively monotonous videos. However, we empirically observe that these matching methods are vulnerable to scenes where many objects with similar appearances have complex trajectories.

Beyond the score-based associations, we take a further step to design a non-heuristic association under the motivation of eliminating the barriers between training and inference. This can be achieved by training the instance prototypes to build relations between consecutive clips. Specifically, as the instance prototypes encapsulate rich spatio-temporal and object-centric information, we formulate instance prototypes p of the previous clip to be instance queries q for the current clip as:

$$q^i = p^{i-1} = \mathcal{D}(q^{i-1}, \mathcal{E}(f^{i-1})), \qquad (1)$$

where i is the index of the current clip. Then, we place a strong condition that forces a unique instance prototype to represent a unique instance throughout the whole video. To this end, we devise a new ground-truth assignment during training to support the strong condition, as existing label assignment strategies are defined within a single clip [5, 13–15, 33] or two clips [32] only.

Unified Video Label Assignment. We define **U**nified **V**ideo **L**abel **A**ssignment (UVLA) with three goals: 1) being capable of incorporating an arbitrary number of clips during training, 2) supporting the reuse of instance prototype for the next clip (Eq. (1)), and 3) completely removing heuristics at inference. To note, there have been similar attempts [28, 39] of reusing instance prototypes as additional instance queries for different clips. However, they still require hyper-parameters and policies to remove or merge queries at inference.

For the definition of UVLA, we introduce two states to instance queries, *'occupied'* and *'unoccupied'*. As illustrated in Fig. 2 (right), instance queries become occupied (colored hexagons) if they are previously matched to a certain ground-truth, and they cannot be matched to another ground-truth object throughout a whole video. On the other hand, only unoccupied instance queries (non-colored hexagons) have the opportunity to be matched to new ground-truth objects that have not appeared in the past.

Formally, at i^{th} clip $\mathcal{V}^i$, we conduct one-to-one bipartite matching between $\{\hat{y}_k^i\}_{k=1}^{N_q}$ and $\{y_k^i\}_{k=1}^{K_{new}^i}$, where the former is the fixed-size set of N_q predictions and the latter is the set of K_{new}^i ground-truths that are newly appeared in $\mathcal{V}^i$, respectively. Following query-based detectors [4,6], we use Hungarian algorithm [19] to obtain the optimal assignment $\hat{\sigma}^i$ among a permutation of N_q elements $\sigma \in \mathfrak{S}_{N_q}$ as:

$$\hat{\sigma}^i = \arg\min_{\sigma \in \mathfrak{S}_{N_q}} \sum_{k=1}^{K_{new}^i} \Big(\mathcal{L}_{match}(y_k^i, \hat{y}_{\sigma(k)}^i) + \alpha \cdot \mathbb{1}_{\sigma(k) \in \Omega_{i-1}} \Big), \quad (2)$$

where $\mathcal{L}_{match}$ is the pair-wise matching cost defined in VITA [13], α is a large constant (*e.g.*, 10^5), and Ω_{i-1} is the indices of previously occupied instance queries. The $|\Omega^{i-1}|$ occupied instance prototypes are directly matched to their previously paired ground-truths, and we can avoid these occupied instance prototypes being matched to newly added ground-truth by simply adding $\alpha \cdot \mathbb{1}_{\sigma(k) \in \Omega_{i-1}}$ to the matching cost. Then, we update Ω with the indices of newly matched queries as follows:

$$\Omega^i = \Omega^{i-1} \cup \{\hat{\sigma}^i(k)\}_{k=1}^{K_i^{new}}, \qquad \Omega_0 = \emptyset. \qquad (3)$$

The rest $(N_q - K_i^{new} - |\Omega^{i-1}|)$ unoccupied predictions are matched to "no object".

With the proposed UVLA label assignment method, we can effectively narrow the gap between training and inference. Concretely, we can use any number of clips (*e.g.*, five)

for training, allowing GenVIS to learn real-video characteristics that appear in long sequences. Furthermore, we can infer a video without any heuristics: if an object is detected by k^{th} instance prototype at a certain clip, the kth instance query tracks the object until the input video ends.

3.3. Instance Prototype Memory

We further adopt memory [7, 25] to boost our long-range association ability by referring to information from past clips. From the definition of UVLA, the occupied instance prototypes must be unceasingly matched to their designated ground-truths until the input video ends. However, this policy would be difficult to be maintained when inferring an extremely long video, because the past information gradually fades out as sequentially processing clips. For example, if an object appeared in the remote past and then disappeared for a long term, the occupied instance query may fail to recapture the same object. To alleviate the limitation, we save concise information – instance prototypes – into the memory for each input clip and read them for improving predictions in the current clip.

Given the ith clip, all instance prototypes from past $i - 1$ clips are gathered into the memory $\mathcal{M} = \{\{p_j^k\}_{j=1}^{N_q}\}_{k=1}^{i-1}$. Using q^i as a query to decode $\mathcal{M}$, we follow standard cross-attention [29] and obtain an output z^i that holds past information. Lastly, we simply extend Eq. (1) to incorporate z^i for initializing object queries as $q^i = p^{i-1} + z^i$. To note, for a query of index j, we narrow the scope to decode only the memory of same indices $\{p_j^k\}_{k=1}^{i-1}$. Compared to globally decoding the memory, the index-wise decoding leads to better memory-reading efficiency and also improves accuracy.

4. Experiments

4.1. Datasets

We evaluate our method on two benchmarks: YouTube-VIS (YTVIS) [36] and Occluded VIS (OVIS) [27]. The YTVIS dataset has three versions (2019, 2021, 2022). All versions tackle segmenting objects of 40 predefined categories in a video. The dataset has been updated to include more challenging videos which are long and contain complex trajectories. While the YTVIS 2021 and 2022 datasets share the same training set, YTVIS 2022 introduces 71 extra videos on top of the validation set of YTVIS 2021. We report the accuracy of the YTVIS 2021 videos and newly added YTVIS 2022 long videos, respectively.

OVIS is another challenging VIS dataset that consists of objects of 25 categories and targets a difficult scenario of heavy occlusions between objects. Also, the OVIS dataset is significantly longer than the YTVIS datasets; for example, the longest video of OVIS is ~50 seconds long while that of YTVIS 2022 is ~10 seconds long.

4.2. Implementation Details

While we adopt the architecture of VITA [13], we replace the window-based self-attention layers for object tokens with the global attention layers. As data augmentation techniques, random flipping and cropping are selected. Also, we generate pseudo videos using images of the COCO dataset [21] and jointly train the model with the VIS datasets following [33]. One batch consists of 8 videos; 5 clips ($N_v^{train} = 5$) are sampled from each video, and the length of clip (N_f^{train}) is 1 and 3 frames for training online and semi-online versions, respectively. Since we freeze the frame-level detector and backbone network, only a small size of memory is required for training. Although a large number of frames are used for training (*e.g.*, 40), the online version of GenVIS with a ResNet-50 backbone can be trained with even a single RTX 3090 GPU. Our implementation is based on the `detectron2` [35] framework and is attached in the supplementary material.

4.3. Main Results

We compare GenVIS with state-of-the-art methods using both lightweight and powerful backbones, *e.g.*, ResNet-50 [12] and Swin-L [23], on the VIS benchmarks: YTVIS 2019/2021/2022 and OVIS. GenVIS shows competitive performance on all benchmarks, and especially, it largely outperforms existing methods in the challenging datasets: YTVIS 2021/2022 and OVIS.

YouTube-VIS 2019 & 2021. In Tab. 1, we compare GenVIS with state-of-the-art methods on the YTVIS 2019 and 2021 benchmarks [36]. On the YTVIS 2019 benchmark, GenVIS achieves the highest AP with ResNet-50, while its AP is marginally lower than IDOL [34] with Swin-L. On the other hand, on the YTVIS 2021 benchmark which consists of more difficult videos, GenVIS surpasses not only IDOL [34] with Swin-L in AP by 3.5 but also with ResNet-50 by 3.2.

YouTube-VIS 2022. As shown in Tab. 2, we also present comparisons on the YouTube-VIS 2022 benchmark [36] which is more challenging than the previous benchmarks (2019 and 2021). Our GenVIS shows significant improvement over the previous state-of-the-art methods: MinVIS [14] and VITA [13]. Based on the strong ability to track objects even under complex trajectories, our method using a ResNet-50 [12] backbone achieves 37.5 AP. This is the highest accuracy among the models that use the same backbone, and also higher than MinVIS using a Swin-L [23] backbone. Attaching our method on top of a Swin-L backbone, GenVIS shows 45.1 AP which is 4.0 AP higher than VITA. Although GenVIS_{semi-online} shows slightly lower AP than its online version, it still outperforms MinVIS and VITA.

Method		Backbone	YouTube-VIS 2019					YouTube-VIS 2021				
			AP	AP$_{50}$	AP$_{75}$	AR$_1$	AR$_{10}$	AP	AP$_{50}$	AP$_{75}$	AR$_1$	AR$_{10}$
Semi-Online / Offline	EfficientVIS [32]	ResNet-50	37.9	59.7	43.0	40.3	46.6	34.0	57.5	37.3	33.8	42.5
	IFC [15]	ResNet-50	41.2	65.1	44.6	42.3	49.6	35.2	55.9	37.7	32.6	42.9
	Mask2Former-VIS [5]	ResNet-50	46.4	68.0	50.0	-	-	40.6	60.9	41.8	-	-
	TeViT [38]	MsgShifT	46.6	71.3	51.6	44.9	54.3	37.9	61.2	42.1	35.1	44.6
	SeqFormer [33]	ResNet-50	47.4	69.8	51.8	45.5	54.8	40.5	62.4	43.7	36.1	48.1
	VITA [13]	ResNet-50	49.8	**72.6**	54.5	49.4	**61.0**	45.7	**67.4**	49.5	**40.9**	**53.6**
	GenVIS$_{\text{semi-online}}$	ResNet-50	**51.3**	72.0	**57.8**	**49.5**	60.0	**46.3**	67.0	**50.2**	40.6	53.2
	SeqFormer [33]	Swin-L	59.3	82.1	66.4	51.7	64.4	51.8	74.6	58.2	42.8	58.1
	Mask2Former-VIS [5]	Swin-L	60.4	84.4	67.0	-	-	52.6	76.4	57.2	-	-
	VITA [13]	Swin-L	63.0	**86.9**	67.9	**56.3**	68.1	57.5	80.6	61.0	47.7	62.6
	GenVIS$_{\text{semi-online}}$	Swin-L	**63.8**	85.7	**68.5**	**56.3**	**68.4**	**60.1**	**80.9**	**66.5**	**49.1**	**64.7**
Online	CrossVIS [37]	ResNet-50	36.3	56.8	38.9	35.6	40.7	34.2	54.4	37.9	30.4	38.2
	VISOLO [10]	ResNet-50	38.6	56.3	43.7	35.7	42.5	36.9	54.7	40.2	30.6	40.9
	MinVIS [14]	ResNet-50	47.4	69.0	52.1	45.7	55.7	44.2	66.0	48.1	39.2	51.7
	IDOL [34]	ResNet-50	49.5	**74.0**	52.9	47.7	58.7	43.9	**68.0**	49.6	38.0	50.9
	GenVIS$_{\text{online}}$	ResNet-50	**50.0**	71.5	**54.6**	**49.5**	**59.7**	**47.1**	67.5	**51.5**	**41.6**	**54.7**
	MinVIS [14]	Swin-L	61.6	83.3	68.6	54.8	66.6	55.3	76.6	62.0	45.9	60.8
	IDOL [34]	Swin-L	**64.3**	**87.5**	**71.0**	55.6	69.1	56.1	80.8	63.5	45.0	60.1
	GenVIS$_{\text{online}}$	Swin-L	64.0	84.9	68.3	**56.1**	**69.4**	**59.6**	**80.9**	**65.8**	**48.7**	**65.0**

Table 1. Comparisons on **YouTube-VIS 2019 and 2021 validation** sets. We group the results by online or not online methods, and then, lightweight (*e.g.*, ResNet-50) or powerful (*e.g.*, Swin-L) backbone networks. For each group, we **bold** the best value in every metric.

Method		AP	AP$_{50}$	AP$_{75}$	AR$_1$	AR$_{10}$
ResNet-50	MinVIS [14]	23.3	47.9	19.3	20.2	28.0
	VITA [13]	32.6	53.9	39.3	30.3	<u>42.6</u>
	GenVIS$_{\text{online}}$	<u>37.5</u>	<u>61.6</u>	41.5	32.6	42.2
	GenVIS$_{\text{semi-online}}$	37.2	58.5	<u>42.9</u>	<u>33.2</u>	40.4
Swin-L	MinVIS† [14]	33.1	54.8	33.7	29.5	36.6
	VITA† [13]	41.1	63.0	44.0	39.3	44.3
	GenVIS$_{\text{online}}$	**45.1**	69.1	**47.3**	39.8	**48.5**
	GenVIS$_{\text{semi-online}}$	44.3	**69.9**	44.9	**39.9**	48.4

Table 2. Comparisons on the newly added **YTVIS 2022 long videos**. †: Evaluated using official repositories as the scores are not specified in original papers. <u>Underline</u> and **bold** denote the highest accuracy using ResNet-50 and Swin-L, respectively.

OVIS. We further validate the competitiveness of GenVIS on the OVIS [26] validation set as shown in Tab. 3. Compared to IDOL [34] which previously ranked the highest accuracy, our method achieves 5.6 and 2.8 AP improvements on top of ResNet-50 and Swin-L backbones, respectively.

4.4. Ablation Studies

We conduct ablation studies to verify the proposed learning method and memory module and show flexible usage of GenVIS in both online and semi-online settings. Here, we use a ResNet-50 [12] backbone and do not use pseudo

Method		AP	AP$_{50}$	AP$_{75}$	AR$_1$	AR$_{10}$
ResNet-50	CrossVIS [37]	14.9	32.7	12.1	10.3	19.8
	VISOLO [10]	15.3	31.0	13.8	11.1	21.7
	TeViT† [38]	17.4	34.9	15.0	11.2	21.8
	VITA [13]	19.6	41.2	17.4	11.7	26.0
	MinVIS [14]	25.0	45.5	24.0	13.9	29.7
	IDOL [34]	30.2	51.3	30.0	15.0	37.5
	GenVIS$_{\text{online}}$	<u>35.8</u>	<u>60.8</u>	<u>36.2</u>	16.3	<u>39.6</u>
	GenVIS$_{\text{semi-online}}$	34.5	59.4	35.0	<u>16.6</u>	38.3
Swin-L	VITA [13]	27.7	51.9	24.9	14.9	33.0
	MinVIS [14]	39.4	61.5	41.3	18.1	43.3
	IDOL [34]	42.6	65.7	45.2	17.9	**49.6**
	GenVIS$_{\text{online}}$	45.2	69.1	**48.4**	**19.1**	48.6
	GenVIS$_{\text{semi-online}}$	**45.4**	**69.2**	47.8	18.9	49.0

Table 3. Comparisons on **OVIS validation** set. <u>Underline</u> and **bold** denote the highest accuracy using ResNet-50 and Swin-L, respectively. † denotes using MsgShifT [38] backbone.

videos generated from the COCO dataset for training.

Effect of the proposed learning method and memory We initiate the experiment with the VITA [13] model that regards an entire video as one long clip ($N_v = 1$) and learns the direct association of objects within a clip. Since we split a video into non-overlapping clips of three frames, we adopt

N_v^{train}	CL	UVLA	IPM	OVIS					YouTube-VIS 2022 Long Videos				
				AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
1				24.1	44.9	22.6	13.4	30.2	25.4	43.6	26.8	22.7	29.6
2	✓			27.4	47.2	27.4	13.9	32.6	29.6	49.3	32.0	26.3	35.0
5	✓			29.8	50.4	31.3	15.0	34.8	32.8	51.9	36.4	32.4	39.1
5	✓	✓		32.5	56.5	33.2	16.1	36.9	35.4	53.4	42.8	32.1	38.6
5	✓	✓	✓	33.4	57.8	34.5	15.7	37.7	36.9	56.5	44.7	32.1	41.4

Table 4. Ablation study of the methods for learning inter-clip association (CL and UVLA) and the memory module (IPM). N_v^{train} is the number of clips used for training. CL and UVLA denote correspondence learning and unified video label assignment, respectively.

	Global	Index
AP	31.8	33.4

Table 5. Ablation study of memory decoding method in IPM on the OVIS validation set.

N_f	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
1	33.0	56.8	34.3	15.7	37.5
3	**33.4**	**57.8**	**34.5**	**15.7**	**37.7**
5	31.3	56.0	32.8	15.3	35.9
7	29.7	51.4	30.7	14.6	34.0

Table 6. Ablation study of the clip length (N_f) on the OVIS validation set. The same value (N_f) is used for training (N_f^{train}) and inference (N_f^{eval}).

Figure 3. Deployment of GenVIS using the different lengths of a clip (N_f^{eval}) for inference on the Youtube-VIS 2019/21/22 and OVIS validation sets. We validated two different GenVIS models trained using N_f^{train} of 1 and 3 which represent online and semi-online settings, respectively. FPS is measured on a single A100 GPU. The models trained for the main experiments are used.

the matching algorithm of MinVIS [14] that uses the cosine similarity of instance prototypes for the association across clips (inter-clip association). This baseline shows AP of 24.1 and 25.4 in OVIS and YTVIS 2022 Long, respectively, as shown in Tab. 4. As the next baseline, we apply the correspondence learning [32] between two clips that the same query refers to the same object across clips. This improves AP to 27.4 and 29.6, respectively.

On top of these two baselines, we extend learning of correspondence between two clips to five clips. Such naive adaptation leads to improvement in both OVIS ($27.4 \rightarrow 29.8$) and YTVIS 2022 Long ($29.6 \rightarrow 32.8$). This demonstrates that training with multiple clips allows the model to effectively learn the inter-clip association. When we adopt UVLA, our label assignment strategy, carefully designed for learning correspondence between multiple clips by considering complex scenarios occurring in long videos, such as a temporary disappearance of objects, AP is significantly improved in both datasets ($29.8 \rightarrow 32.5$ and $32.8 \rightarrow 35.4$). Incorporating the instance prototype memory (IPM) additionally increases AP. In a nutshell, our proposed training method improves AP by 6.0 and 7.3 in OVIS and YTVIS 2022 Long, respectively, even though we maintain the simple and heuristic-free inter-clip association method.
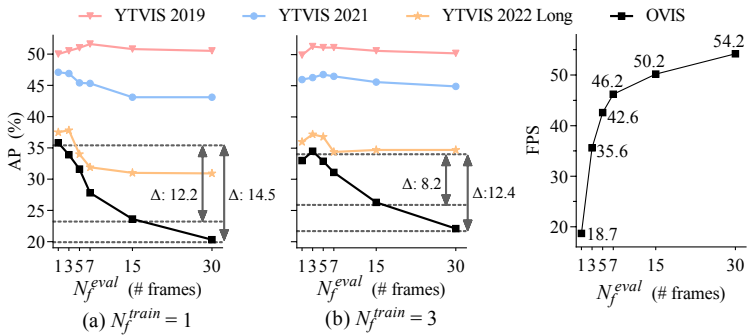
Memory decoding in IPM. As shown by Tab. 5, global decoding of instance prototypes in the memory does not show an improvement in the accuracy compared to index-wise decoding. Also, index-wise decoding is computationally more scalable with a long video, thus, we decide to use it in IPM.

Length of clip (N_f) for training. With the fixed number of clips ($N_v = 5$), we experiment with the length of clip (N_f) used for training. As shown in Tab. 6, there is no notable change in AP when N_f increases from 1 to 3. However, higher N_f values (*e.g.*, 5 and 7) bring about degradation of AP, for example, a decrease of 2.1 and 3.7, respectively, compared to $N_f = 3$. Based on this empirical observation, we use $N_f = 3$ for training the semi-online version of GenVIS.

Flexibility of GenVIS for online and semi-online settings. As shown in Fig. 3, GenVIS is flexible to use different clip lengths for inference (N_f^{eval}) although it is trained with a clip length (N_f^{train}) of 1 or 3. In the YTVIS 2019 dataset, increasing the window size from one to seven improves AP even though GenVIS is trained with a clip of one frame. Since YTVIS 2021 & 2022 and OVIS datasets have more complex trajectories of objects than YTVIS 2019, using a larger window size for inference leads to lower AP, especially

Figure 4. Qualitative comparisons of our method, GenVIS, with the state-of-the-art methods: MinVIS [14] and VITA [13]. The videos on the left and right are from YouTube-VIS 2022 [36] and OVIS [27] datasets, respectively. GenVIS shows impressive accuracy in these complicated scenes where the objects look similar crossing each other. Objects with the same identity are displayed in the same color.

in the OVIS dataset where objects are heavily occluded in each other. In these datasets, using a short clip to extract the instance prototypes is more effective than a long clip which is also observed in Tab. 6. In contrast, using a longer clip outputs temporally compact instance prototypes that lead to faster execution speed and memory efficiency. In the case of such usage, it is better to use the GenVIS trained using a long clip (*e.g.*, $N_f^{train} = 3$) since the decrease in AP is lower than the online version which is trained with $N_f^{train} = 1$.

4.5. Qualitative Results

Fig. 4 shows the visualization results of our GenVIS and recent online [14] and offline VIS methods [13] on the benchmarks consisting of long and complicated scenarios. In the selected scenes, the animals are similar in appearance, and their trajectories are too complex to keep tracking the same instance. In the scene of YTVIS 2022, MinVIS fails to segment the elephant which is severely occluded by the elephant in front of the target, and VITA fails to track the front elephant resulting in the change of its ID. However, GenVIS successfully segments and tracks all objects in this challenging scene. GenVIS performs well even in the scene of OVIS where the left-front cat is moving behind two cats.

5. Limitations and Future Works

Our method achieves state-of-the-art performance on long videos in both online and semi-online manners. Also, we demonstrate the trade-offs between the accuracy and the efficiency for various learning options. While the proposed framework provides a choice of the length of the clip without modifying the architecture, once the length of the clip is determined, our method treats an input video uniformly in that unit. However, there can be a variety of situations in a single video. For example, if a scene continues mono-

tonically, it may be advantageous to process it broadly and put it in memory without taking the scope small, whereas frame-level processing could be beneficial in the opposite case. Therefore, designing adjustable windows will be an interesting future direction. In addition, because we take a label assignment strategy that gives unique IDs to unique object queries, the strategy can be limited if there are very large numbers of objects across the video. Designing an algorithm integrated with memory could be another future direction.

6. Conclusion

In this paper, we propose GenVIS a generalized framework for video instance segmentation (VIS). To seamlessly bridge the gap between training and inference, we propose the following: 1) a training strategy that can involve multiple clips, 2) training the associations between separate clips with the novel label assignment - UVLA, and 3) the use of memory to alleviate vanishing information in long videos. By integrating these proposals into a single framework, we demonstrate the effectiveness of our method by achieving state-of-the-art results on multiple benchmarks.

Acknowledgements

References

[1] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*, 2020. 1, 2, 3, 4

[2] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *CVPR*, 2022. 3

[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. 2

[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, 3, 4

[5] 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, 4, 6

[6] 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, 4

[7] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 5

[8] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, 2021. 1

[9] Yang Fu, Linjie Yang, Ding Liu, Thomas S Huang, and Humphrey Shi. Compfeat: Comprehensive feature aggregation for video instance segmentation. In *AAAI*, 2021. 2

[10] Su Ho Han, Sukjun Hwang, Seoung Wug Oh, Yeonchool Park, Hyunwoo Kim, Min-Jung Kim, and Seon Joo Kim. Visolo: Grid-based space-time aggregation for efficient online video instance segmentation. In *CVPR*, 2022. 2, 6

[11] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6

[13] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. In *NeurIPS*, 2022. 1, 2, 3, 4, 5, 6, 8

[14] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. In *NeurIPS*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[15] 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, 4, 6

[16] Jeongseok Hyun, Myunggu Kang, Dongyoon Wee, and Dit-Yan Yeung. Detection recovery in online multi-object tracking with sparse graph tracker. In *WACV*, 2023. 3

[17] 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. 2

[18] Dahun Kim, Jun Xie, Huiyu Wang, Siyuan Qiao, Qihang Yu, Hong-Seok Kim, Hartwig Adam, In So Kweon, and Liang-Chieh Chen. Tubeformer-deeplab: Video mask transformer. In *CVPR*, 2022. 3

[19] Harold W Kuhn. The hungarian method for the assignment problem. In *Naval research logistics quarterly*, 1955. 4

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

[21] 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. 5

[22] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *CVPR*, 2021. 2

[23] 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. 5

[24] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, 2022. 3

[25] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 2, 5

[26] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation. *arXiv preprint arXiv:2102.01558*, 2021. 6

[27] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: Dataset and iccv 2021 challenge. In *NeurIPS Track on Datasets and Benchmarks*, 2021. 1, 2, 3, 5, 8

[28] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 4

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 4, 5

[30] 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, 2, 3

[31] Sanghyun Woo, Kwanyong Park, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Tracking by associating clips. In *ECCV*, 2022. 3

[32] Jialian Wu, Sudhir Yarram, Hui Liang, Tian Lan, Junsong Yuan, Jayan Eledath, and Gerard Medioni. Efficient video instance segmentation via tracklet query and proposal. In *CVPR*, 2022. 2, 3, 4, 6, 7

[33] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. In *ECCV*, 2022. 1, 2, 3, 4, 5, 6

[34] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, 2022. 1, 2, 5, 6

[35] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 5

[36] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 2, 5, 8

[37] 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. 2, 6

[38] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *CVPR*, 2022. 1, 6

[39] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *ECCV*, 2022. 3, 4

[40] 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, 3