CAVIS: Context-Aware Video Instance Segmentation

Seunghun Lee*, Jiwan Seo*, Kiljoon Han, Minwoo Choi, and Sunghoon Im
DGIST, Daegu, Korea
{lsh5688, eccaron, kiljoon_h, subminu, sunghoonim}@dgist.ac.kr

Abstract

In this paper, we introduce the Context-Aware Video Instance Segmentation (CAVIS), a novel framework designed to enhance instance association by integrating contextual information adjacent to each object. To efficiently extract and leverage this information, we propose the Context-Aware Instance Tracker (CAIT), which merges contextual data surrounding the instances with the core instance features to improve tracking accuracy. Additionally, we design the Prototypical Cross-frame Contrastive (PCC) loss, which ensures consistency in object-level features across frames, thereby significantly enhancing matching accuracy. CAVIS demonstrates superior performance over state-of-the-art methods on all benchmark datasets in video instance segmentation (VIS) and video panoptic segmentation (VPS). Notably, our method excels on the OVIS dataset, known for its particularly challenging videos. Project page: this https URL

1. Introduction

Video Instance Segmentation (VIS) is a crucial task that involves segmenting and identifying individual objects within video sequences, applicable in a variety of fields including video understanding, autonomous driving, and video editing [49]. VIS has seen considerable advancements, with developments in both online methods [4, 19, 22, 45, 49, 51, 53], which process videos frame-by-frame to adapt in real-time, and offline methods [8, 17, 20, 42, 44], which analyze entire videos to understand inter-frame dependencies thoroughly.

Recent advances have brought robust query-based segmentation architectures [9, 10], designed to detect instance centers and cluster pixels into instance-specific groups within images. These modern VIS approaches increasingly employ instance center associations across frames to improve tracking accuracy. Techniques such as contrastive learning [45, 53] and transformer-based trackers [18, 55] leverage the similarities between instance centers for consistent identification of objects across frames. However, challenges persist in scenarios with significant occlusions or when multiple similar objects are present, leading to potential tracking inaccuracies as shown in the top row of Fig. 1. While some strategies [17, 18] attempt to use instance features for tracking across segments or entire videos, difficulties remain.

To address this issue, we propose Context-Aware Video Instance Segmentation (CAVIS), a novel framework designed to improve object identification by incorporating contextual information surrounding each instance into the tracking process. This approach draws from insights in neuroscience and cognitive science [2, 35], emphasizing the importance of contextual cues in human perception for deciphering complex scenes and resolving visual ambiguities. An example of the practical application of this principle is shown in Fig. 1, where recognizing a person is riding a bicycle, rather than just identifying the bicycle, greatly enhances the accuracy of object identification.

To achieve this, we design the Context-Aware Instance Tracker (CAIT), featuring advanced modules for extracting and matching context-aware instance features. The context-aware instance feature extractor combines the contextual information at the object's boundary with the core features of each instance. Then, we incorporate these context-aware instance features into a transformer-based tracking architecture [18, 44, 55], enhanced by our context-aware cross-attention mechanism. This adjustment allows for the precise utilization of detailed contextual nuances within each scene.

Furthermore, inspired by a recent study [23], we design the Prototypical Cross-frame Contrastive (PCC) loss, specifically designed to enhance feature consistency across frames. PCC loss correlates each pixel with an object by comparing core instance features against pixel embeddings, enhancing regional consistency within frames and establishing an intra-frame constraint reflective of the feature map's similarity. This method leverages the close relationship between object features and high-level feature maps, where their similarity guides mask predictions. By constructing instance-wise prototypes, PCC loss maintains frame-to-frame consistency, improving training efficiency and ensuring robust performance in dynamic environments.

Our extensive testing shows that CAVIS significantly

*These authors contribute equally to this work.

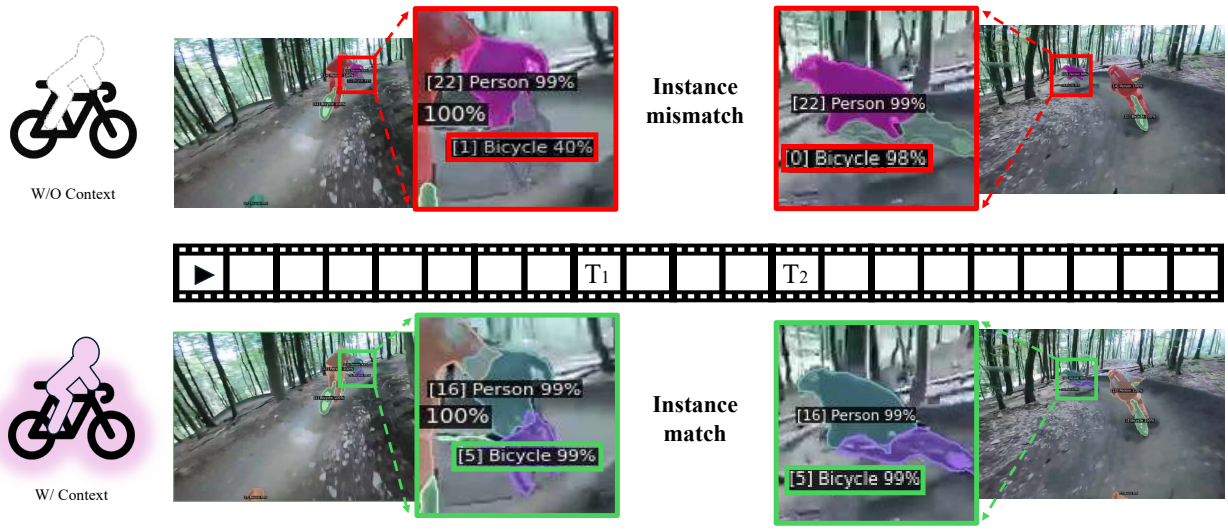


Figure 1. **Importance of contextual information**. Comparative results showing the state-of-the-art model [56] (Top) and CAVIS (Bottom). The frame on the left precedes the right by four frames, during which an occlusion takes place. The standard model, lacking contextual data, fails to consistently track the same bicycle post-occlusion, while CAVIS effectively maintains accurate instance tracking.

outperforms existing state-of-the-art methods across major video segmentation benchmarks, including YTVIS19 [49], YTVIS21 [50], OVIS [37], and VIPSeg [31], particularly excelling on OVIS dataset that includes complex videos.

2. Related Works

Video Instance Segmentation. VIS methods learn to associate features frame-to-frame based on instance segmentation architectures. The pioneering MaskTrack R-CNN [49] integrates a tracking head into Mask R-CNN [16], utilizing heuristic cues for instance association Following advancements include SipMask [4] and CrossVIS [51], which enhance temporal links through cross-frame learning. IDOL[45] a contrastive learning approach with query-based architectures [59], boosting online method performance. Conversely, offline approaches like VisTR [42] and Seqformer [44] use the entire video for mask trajectory predictions, with VisTR applying DETR [5] at the clip level and Seqformer aggregating temporal information via inter-frame queries. Innovations like IFC [20] and TeViT [52] improve efficiency by adjusting attention mechanisms within transformer architectures.

Advancements in Query-based Networks. Strong query-based segmentation networks have become prevalent in current VIS methods, with many relying on Mask2Former [10] as their foundation. MinVIS [19] achieves tracking through simple post-processing based on cosine similarity between object features, without video learning. VITA [17] temporally associates frame-level queries to find instance prototypes within a video. GenVIS [18] adopts object association approach of VITA and designs a tracking network at the sub-clip level. Inspired by SimCLR [6], CTVIS [53] utilizes contrastive learning with a larger number of frames for comprehensive frame association. DVIS [55] introduces a decoupled framework for VIS, dividing it into segmentation, tracking, and refinement tasks, thereby enabling efficient and effective learning.

Object Tracking with Additional Cues. Tracking methodologies have been developed across various domains, including video object segmentation (VOS) [11, 34, 48], multiple object tracking (MOT) [3, 33, 57, 58], and VIS. Despite advancements, many challenging cases persist, prompting research into object association with supplementary data. Early approaches leverage spatial-temporal information such as geometric relation between adjacent frames [41] and aggregated object features of previous frames [47]. BeyondPixel [39] improves inter-frame object matching by proposing a new cost that captures 3D pose and shape based on monocular geometry. BATMAN [54] combines optical flow and object query features to encode motion and appearance information into bilateral space. CAROQ [12] employs a context feature defined as a memory bank of multi-level image features extracted by a pixel decoder. However, such a full-context-based approach can lead to increased complexity and memory limitations. In contrast, our method takes a memory-efficient approach by focusing on the surrounding features of each object during tracking, enabling effective object matching.

3. Preliminary

This section offers a concise introduction to the fundamentals of a query-based segmentation pipeline and outlines a VIS approach that incorporates contrastive loss [26, 28, 45, 53].

3.1. Query-based instance segmentation

Modern VIS methods adopt a query-based instance segmentation pipeline [9, 10], including three main components: a backbone encoder, a pixel decoder, and a transformer decoder. The backbone encoder and pixel decoder are responsible for extracting multi-scale feature maps from the input image. The transformer decoder employs object queries—sequences of latent vectors—as initial guesses for object centers and utilizes these features to generate object-level features. These queries undergo refinement through multiple transformer blocks via a cross-attention mechanism between the object queries and the feature maps. The refined instance features are then used for classification and segmentation tasks through respective prediction heads. Typically, the number of object queries, N, exceeds the actual number of objects, N_{GT}, present in the image. Traditionally, the process involves finding a permutation of N elements, $\sigma \in \mathfrak{S}_N$, that optimally assigns the prediction set $\{\hat{y}_i\}_{i=1}^{N}$ to maximize total similarity to the ground truth (GT) set $\{y_i\}_{i=1}^{N_{GT}}$. This is achieved by minimizing a pair-wise matching cost $\mathcal{L}_{\text{Match}}$, as defined in [8]:

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_N}{\arg\min} \sum_{k=1}^{N_{GT}} \mathcal{L}_{\text{Match}} \left(y_k, \hat{y}_{\sigma(k)}\right). \tag{1}$$

The network is trained with an objective function, $\mathcal{L}_{\text{Inst}}$, which consists of a categorical loss ($\mathcal{L}_{\text{Cls}}$), a binary cross-entropy loss for masks ($\mathcal{L}_{\text{Bce}}$), and a dice loss ($\mathcal{L}_{\text{Dice}}$) with the weights λ_{Cls}, λ_{Bce}, and λ_{Dice} balancing the contributions of each loss component as follows:

$$\mathcal{L}_{\text{Inst}} = \lambda_{\text{Cls}}\mathcal{L}_{\text{Cls}} + \lambda_{\text{Bce}}\mathcal{L}_{\text{Bce}} + \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}}. \tag{2}$$

The loss is used to train VIS framework with the frame-wise matching relation $\hat{\sigma}^t$ as follows:

$$\mathcal{L}_{\text{VIS}} = \sum_{t=1}^{T} \sum_{n=1}^{N_{GT}} \mathcal{L}_{\text{Inst}} \left(y_n^t, \hat{y}_{\hat{\sigma}^t(n)}^t\right). \tag{3}$$

3.2. Contrastive learning for VIS

In query-based architectures, the order of instance features effectively serves as the identity of each object. By aligning the sequence of instance features across frames, we can facilitate object tracking. Since instance features represent specific objects, inter-frame feature association is used for this alignment. To enhance the robustness of instance features for matching objects between frames, the following contrastive loss is integrated within the VIS framework [45]:

$$\mathcal{L}_{\text{Emb}}(v_t) = \log \left[1 + \sum_{k^+ \in \mathbf{K}_{v_t}^+} \sum_{k^- \in \mathbf{K}_{v_t}^-} \exp\left(v_t \cdot k^- - v_t \cdot k^+\right)\right], \tag{4}$$

where $\mathbf{K}_{v_t}^+$ represents the sets of positive embeddings corresponding to the same object as v_t from frames other than the t-th frame, while $\mathbf{K}_{v_t}^-$ includes negative embeddings featuring characteristics of objects different from that of v_t.

4. Method

This section describes our Context-Aware Video Instance Segmentation (CAVIS) whose overall pipeline is illustrated in Fig. 2. Our CAVIS consists of two key components: Context-Aware Instance Tracker (CAIT) and Prototypical Cross-frame Contrastive (PCC) loss, which are detailed in Sec. 4.1 and Sec. 4.2, respectively. We describe the training losses for each network in Sec. 4.3. We provide a notation table in Tab. 1 for better readability.

4.1. Context-Aware Instance Tracker

4.1.1. Context-aware feature extraction

Following the method outlined in previous VIS studies [17–19, 53, 55], we employ Mask2Former [10] as our segmentation network $\mathcal{S}$. This framework ingests a series of input frames $\{I_t\}_{t=1}^{T}$, with T denoting the total number of frames. It extracts feature maps F, identifies instance features $\hat{Q}$, and computes both classification scores O and generates segmentation masks M as follows:

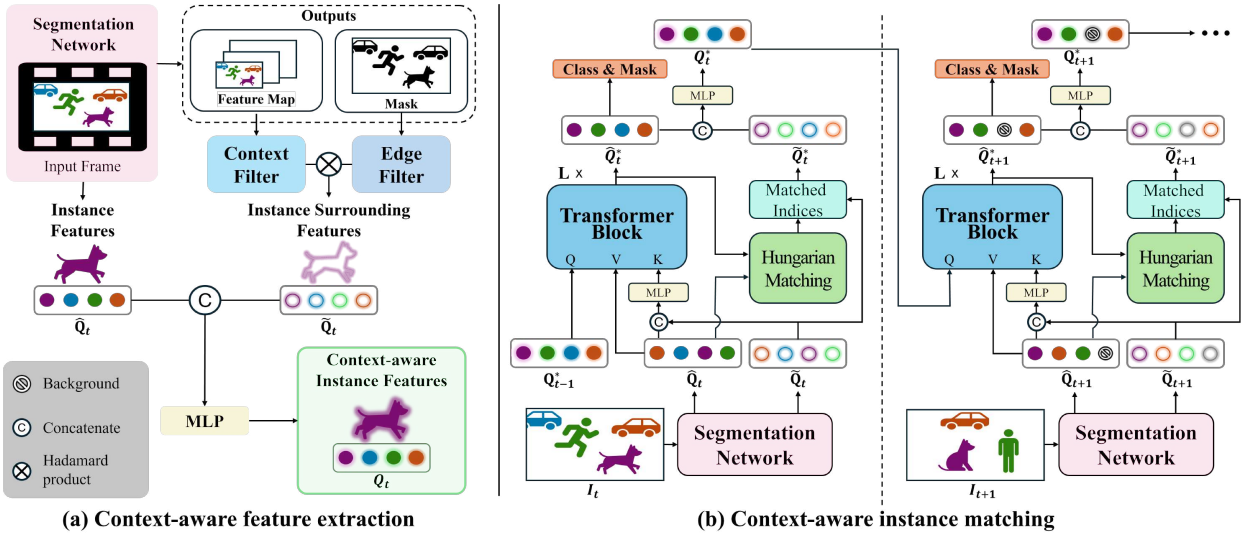
$$\left\{F_t, \hat{Q}_t, O_t, M_t\right\}_{t=1}^{T} = \mathcal{S}\left(\{I_t\}_{t=1}^{T}\right),$$
$$F_t \in \mathbb{R}^{H\times W\times C},\ \hat{Q}_t \in \mathbb{R}^{N\times C},\ O_t \in \mathbb{R}^{N\times K},\ M_t \in \mathbb{R}^{N\times H\times W}, \tag{5}$$

where H, W, and C denote the height, width, and channel dimensions of the feature maps, respectively. N indicates the maximum number of detactable objects in a single frame, and K signifies the number of object classes. We then extract the instance surrounding features $\tilde{Q}_t \in \mathbb{R}^{N\times C}$ capturing data around the object's boundaries essential for detailed context analysis as follows:

$$\tilde{Q}_t^n = \frac{\sum_{h=1}^{H}\sum_{w=1}^{W} \bar{F}_t^{\{h,w\}} * \mathbb{1}\left(\acute{M}_t^{\{n,h,w\}} > 0\right)}{\sum_{h=1}^{H}\sum_{w=1}^{W} \mathbb{1}\left(\acute{M}_t^{\{n,h,w\}} > 0\right)},$$
$$\bar{F} = \text{Avg}(F),\ \acute{M} = \text{Lap}(M),\ \forall n = \{1, ..., N\}, \tag{6}$$

where $\text{Avg}(\cdot)$ denotes an average filtering process over spatial dimensions, $\text{Lap}(\cdot)$ signifies the application of a Laplacian filter, and $\mathbb{1}(\cdot)$ is the indicator function that tests for the presence of the object within the filtered mask. We employ the average filter specifically configured with a 9×9 kernel.

Finally, we combine the core and surrounding features to further enhance instance representations. The context-aware instance feature $Q_t^n \in \mathbb{R}^C$ is generated by concatenating the core instance feature $\hat{Q}_t^n$ and the instance surrounding feature $\tilde{Q}_t^n$ along the channel dimension, and subsequently



(a) Context-aware feature extraction (b) Context-aware instance matching

Figure 2. **Overview of CAVIS**. (a) The extraction of **context-aware instance features** from the output of an instance segmentation network. (b) CAVIS pipeline through **context-aware instance matching**. This includes the organization of the surrounding instance features $\tilde{Q}_t$, facilitated by Hungarian matching between the ordered instance features $\hat{Q}_t^*$ and unordered instance features $\hat{Q}_t$.

Table 1. Notations used in our method.

Symbol	Description	Symbol	Description
$\hat{Q}$	: instance features	M	: mask predictions
$\tilde{Q}$	: instance surrounding features	$\acute{M}$	: boundary scores processed from M
Q	: context-aware instance features	F	: the last feature maps from pixel decoder
Q^*	: aligned context-aware instance features	$\bar{F}$	: feature maps processed by average filter

processing this combined feature through a multi-layer perceptron (MLP) as follows:

$$Q_t^n = \text{MLP}\left(\text{Concat}\left(\hat{Q}_t^n, \tilde{Q}_t^n\right)\right), \quad \forall n = \{1, ..., N\}. \quad (7)$$

The MLP is structured with three linear layers, each followed by a ReLU activation function. To promote the learning of discriminative context-aware instance features, we implement a contrastive loss specifically for these features. The context-aware contrastive loss, denoted as $\mathcal{L}_{\text{CTX}}$, leverages the established contrastive loss framework $\mathcal{L}_{\text{Emb}}$ detailed in Eq. (4) as follows:

$$\mathcal{L}_{\text{CTX}} = \sum_{t=1}^{T} \sum_{n=1}^{N_{GT}} \mathcal{L}_{\text{Emb}}\left(Q_t^{\hat{\sigma}^t(n)}\right), \quad (8)$$

where $\hat{\sigma}^t$ is a frame-wise matching relation as described in Eq. (1). This design enhances our ability to identify and track the same instances across the entire video. By not solely relying on instance centers, and instead utilizing context-aware instance embeddings, we can more accurately recognize instances throughout the video sequence.

4.1.2. Context-aware instance matching

We introduce our context-aware tracking network $\mathcal{T}$, which employs a transformer-based tracking network [18, 55] to learn associations across adjacent frames. The network comprises six transformer blocks, each featuring cross-attention, self-attention, and feed-forward layers. The conventional cross-attention mechanism, denoted as $\text{Attn}(Q, K, V)$, traditionally aligns the current unordered instance features, $\hat{Q}_t$, (serving as both the key and value), with the ordered instance features from the previous frame, $\hat{Q}_{t-1}^*$, (used as the query). To enhance accuracy, our model employs context-aware instance features, Q_{t-1}^* and Q_t, as the query and key, respectively while we still use the instance features $\hat{Q}_t$ as value. This modification leads to a context-aware cross-attention mechanism, formulated as:

$$\text{Attn}(Q_{t-1}^*, Q_t, \hat{Q}_t) = \text{softmax}\left(\frac{Q_{t-1}^* \cdot (Q_t)^T}{\sqrt{C}}\right)\hat{Q}_t, \quad (9)$$

where C is the channel dimensions of the feature maps. The aligned context-aware instance features Q_t^* is built by concatenating the aligned instance features $\hat{Q}_t^*$ and the aligned instance surrounding features $\tilde{Q}_t^*$ same as in Eq. (7). We obtain the aligned instance surrounding features $\tilde{Q}_t^*$ by us-

Figure 3. Qualitative comparisons of CAVIS (ours) against state-of-the-art methods: CTVIS [53] and DVIS++ [56] on the OVIS dataset.

ing Hungarian matching algorithm [24] on cosine similarity between $\hat{Q}_t$ and $\hat{Q}_t^*$ as follows:

$$\begin{aligned} \tilde{Q}_t^{*\sigma_H(n)} &= \tilde{Q}_t^n, \ \forall n = \{1, ..., N\}, \\ \text{where } \sigma_H &= \text{Hungarian}(\hat{Q}_t^*, \hat{Q}_t), \ \sigma_H \in \mathbb{R}^N. \end{aligned} \tag{10}$$

This process is similarly applied to the aligned context-aware features Q_{t-1}^* for the previous frame.

4.2. Prototypical Cross-frame Contrastive Loss

Current VIS methodologies emphasize the importance of object feature representation learning, especially in tracking tasks where matching object features across frames is crucial. In a query-based segmenter, object features $\hat{Q}_t^n \in \mathbb{R}^C$ are semantically correlated with each pixel embedding of the feature map $F_t^{\{h,w\}} \in \mathbb{R}^C$ from the pixel decoder, as they are used for mask prediction. This ensures consistent feature representation within object-containing regions and introduces an intra-frame constraint that reflects similarity within the feature map. By examining the inter-frame relationships of pixel embeddings, our method improves object association, essential for contrastive learning methods that strive to differentiate between similar and dissimilar objects across frames.

Given the memory-intensive nature of maintaining individual pixel consistency, we introduce Prototypical Cross-frame Contrastive (PCC) loss. This loss $\mathcal{L}_{\text{PCC}}$ maintains frame-to-frame consistency of pixel embeddings for each instance feature by constructing instance-wise prototypes from predicted masks, defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{PCC}} &= \sum_{t=1}^{T} \sum_{n=1}^{N_{GT}} \mathcal{L}_{\text{Emb}} \left(\eta_t^{\hat{\sigma}^t(n)} \right), \\ \eta_t^n &= \frac{\sum_{h=1}^{H} \sum_{w=1}^{W} F_t^{\{h,w\}} * \mathbb{1}\left(M_t^{\{h,w\}} == 1 \right)}{\sum_{h=1}^{H} \sum_{w=1}^{W} \mathbb{1}\left(M_t^{\{h,w\}} == 1 \right)}. \end{aligned} \tag{11}$$

4.3. Training Loss

To train the segmentation network $\mathcal{S}$, we implement an objective function that incorporates the standard video instance segmentation loss $\mathcal{L}_{\text{VIS}}$ in Eq. (3), context-aware constrative loss $\mathcal{L}_{\text{CTX}}$ in Eq. (8) and Prototypical Cross-frame Contrastive (PCC) loss in Eq. (11) as follows:

$$\mathcal{L}_{\mathcal{S}} = \mathcal{L}_{\text{VIS}} + \lambda_{\text{CTX}} \mathcal{L}_{\text{CTX}} + \lambda_{\text{PCC}} \mathcal{L}_{\text{PCC}}, \tag{12}$$

where λ_{CTX} and λ_{PCC} are the weights assigned to balance these losses.

To train the tracking network $\mathcal{T}$, we calculate the matching cost only for objects that appear for the first time to ensure consistent video-level matching pairs, as implemented in prior work [55]. The objective function $\mathcal{L}_{\mathcal{T}}$ incorporating the matching relation $\hat{\sigma}_{\text{con}}$ is defined as follows:

$$\begin{aligned} \mathcal{L}_{\mathcal{T}} &= \sum_{t=1}^{T} \sum_{n=1}^{N_{GT}} \mathcal{L}_{\text{Inst}} \left(y_n^t, \hat{y}_{\hat{\sigma}_{\text{con}}(n)}^t \right), \\ \hat{\sigma}_{\text{con}} &= \underset{\sigma \in \mathfrak{S}_N}{\arg\min} \sum_{k=1}^{N_{GT}} \mathcal{L}_{\text{Match}} \left(y_k^{f(k)}, \hat{y}_{\sigma(k)}^{f(k)} \right), \end{aligned} \tag{13}$$

Table 2. Comparisons on the validation sets of YouTube-VIS 2019, 2021, and OVIS datasets. The best and second-best scores are highlighted in **red** and **blue**, respectively. † denotes the model trained with the temporal refiner [55]. Rows in cyan indicate comparisons with a top-performing model.

	Methods	OVIS					YTVIS19					YTVIS21				
		AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50	MaskTrack R-CNN [49]	10.8	25.3	8.5	7.9	14.9	30.3	51.1	32.6	31.0	35.5	28.6	48.9	29.6	26.5	33.8
	SipMask [4]	10.2	24.7	7.8	7.9	15.8	33.7	54.1	35.8	35.4	40.1	31.7	52.5	34.0	30.8	37.8
	IFC [20]	13.1	27.8	11.6	9.4	23.9	41.2	65.1	44.6	42.3	49.6	35.2	55.9	37.7	32.6	42.9
	CrossVIS [51]	14.9	32.7	12.1	10.3	19.8	36.3	56.8	38.9	35.6	40.7	34.2	54.4	37.9	30.4	38.2
	EfficientVIS [46]	-	-	-	-	-	37.9	59.7	43.0	40.3	46.6	34.0	57.5	37.3	33.8	42.5
	SeqFormer [44]	15.1	31.9	13.8	10.4	27.1	47.4	69.8	51.8	45.5	54.8	40.5	62.4	43.7	36.1	48.1
	VISOLO [14]	15.3	31.0	13.8	11.1	21.7	38.6	56.3	43.7	35.7	42.5	36.9	54.7	40.2	30.6	40.9
	Mask2Former-VIS [8]	17.3	37.3	15.1	10.5	23.5	46.4	68.0	50.0	-	-	40.6	60.9	41.8	-	-
	VITA [17]	19.6	41.2	17.4	11.7	26.0	49.8	72.6	54.5	49.4	61.0	45.7	67.4	49.5	40.9	53.6
	MinVIS [19]	25.0	45.5	24.0	13.9	29.7	47.4	69.0	52.1	45.7	55.7	44.2	66.0	48.1	39.2	51.7
	CAROQ [12]	25.8	47.9	25.4	14.2	33.9	46.7	70.4	50.9	45.7	55.9	43.3	64.9	47.1	39.3	52.7
	IDOL [45]	28.2	51.0	28.0	14.5	38.6	49.5	74.0	52.9	47.7	58.7	43.9	68.0	49.6	38.0	50.9
	DVIS [55]	30.2	55.0	30.5	14.5	37.3	51.2	73.8	57.1	47.2	59.3	46.4	68.4	49.6	39.7	53.5
	TCOVIS [25]	35.3	60.7	**36.6**	15.7	39.5	52.3	73.5	57.6	49.8	60.2	49.5	71.2	53.8	41.3	55.9
	CTVIS [53]	35.5	**60.8**	34.9	16.1	**41.9**	**55.1**	**78.2**	59.1	**51.9**	**63.2**	50.1	73.7	54.7	41.8	**59.5**
	GenVIS [18]	35.8	**60.8**	36.2	**16.3**	39.6	50.0	71.5	54.6	49.5	59.7	47.1	67.5	51.5	41.6	54.7
	VISAGE [23]	**36.2**	60.3	35.3	16.1	40.3	**55.1**	78.1	**60.6**	51.0	62.3	**51.6**	**73.8**	**56.1**	**43.6**	**59.3**
	Ours	**37.6**	**63.4**	**38.2**	**16.5**	**43.5**	**55.7**	**78.3**	**61.7**	**51.5**	**63.3**	**50.5**	**74.1**	**54.9**	**42.6**	58.5
Swin-L	SeqFormer [44]	-	-	-	-	-	59.3	82.1	66.4	51.7	64.4	51.8	74.6	58.2	42.8	58.1
	Mask2Former-VIS [8]	25.8	46.5	24.4	13.7	32.2	60.4	84.4	67.0	-	-	52.6	76.4	57.2	-	-
	VITA [17]	27.7	51.9	24.9	14.9	33.0	63.0	86.9	67.9	56.3	68.1	57.5	80.6	61.0	47.7	62.6
	CAROQ [12]	38.2	60.7	39.5	17.7	44.1	61.4	82.8	68.6	55.2	68.1	54.5	75.4	60.5	45.5	61.4
	MinVIS [19]	39.4	61.5	41.3	18.1	43.3	61.6	83.3	68.6	54.8	66.6	55.3	76.6	62.0	45.9	60.8
	IDOL [45]	40.0	63.1	40.5	17.6	46.4	64.3	87.5	71.0	55.6	69.1	56.1	80.8	63.5	45.0	60.1
	GenVIS [18]	45.2	69.1	48.4	**19.1**	48.6	64.0	84.9	68.3	56.1	69.4	59.6	80.9	65.8	**48.7**	65.0
	DVIS [55]	45.9	71.1	48.3	18.5	51.5	63.9	87.2	70.4	56.2	69.0	58.7	80.4	66.6	47.5	64.6
	TCOVIS [25]	46.7	70.9	**49.5**	**19.1**	50.8	64.1	86.6	69.5	55.8	69.0	**61.3**	82.9	68.0	**48.6**	65.1
	CTVIS [53]	**46.9**	**71.5**	47.5	**19.1**	**52.1**	**65.6**	**87.7**	**72.2**	**56.5**	**70.4**	**61.2**	**84.0**	**68.8**	48.0	**65.8**
	Ours	**48.6**	**74.0**	**52.5**	**19.5**	**53.3**	**66.0**	**89.5**	**73.3**	**56.8**	**71.4**	61.1	**84.1**	**69.2**	48.2	**66.3**
ViT-L	MinVIS [19]	42.9	65.7	45.4	19.8	46.5	65.6	85.4	72.7	57.5	70.6	59.2	79.9	66.7	47.8	64.1
	DVIS++ [56]	**49.6**	**72.5**	**55.0**	**20.8**	**54.6**	**67.7**	**88.8**	**75.3**	**57.9**	**73.7**	**62.3**	**82.7**	**70.2**	**49.5**	**68.0**
	Ours	**53.2**	**75.9**	**59.1**	**20.9**	**58.2**	**68.9**	**89.3**	**76.2**	**58.3**	**73.6**	**64.6**	**85.6**	**72.5**	**49.5**	**69.3**
	DVIS++† [56]	**53.4**	**78.9**	**58.5**	**21.1**	**58.7**	**68.3**	**90.3**	**76.1**	**57.8**	**73.4**	**63.9**	**86.7**	**71.5**	**48.8**	**69.5**
	Ours†	**57.1**	**82.6**	**63.5**	**21.2**	**61.8**	**69.4**	**90.9**	**77.2**	**58.3**	**74.7**	**65.3**	**87.3**	**73.2**	**49.7**	**70.3**

where $f(k)$ denotes the frame in which the k-th instance first appears.

5. Experiments

We evaluate CAVIS on two major tasks: video instance segmentation (VIS) and video panoptic segmentation (VPS) on four benchmark datasets recognized for their challenges and prevalence in the research community: YouTubeVIS-2019 [49], YouTubeVIS-21 [50], OVIS [37], and VIPSeg [31]. For VIS, performance metrics include average precision (AP) and average recall (AR) as established in previous studies [49]. In the realm of VPS [21], we further examine our model's capabilities using the Segmentation and Tracking Quality (STQ) metric, and the Video Panoptic Quality (VPQ) metric.

5.1. Implementation Details

We employ Mask2Former [10] as our segmentation network, utilizing three distinct backbone encoders: ResNet-50 [15], Swin-L [30], and ViT-L [13]. The ResNet-50 and Swin-L backbones are initialized with parameters pre-trained on the COCO dataset [29], while the ViT-L backbone uses initialization parameters from DINOv2 [36]. Additionally, for the ViT-L backbone, we employ a memory-efficient version of VIT-Adapter [7], aligning with recent advancements in network efficiency [56]. Further details are described in the Supplementary Material.

5.2. Comparison to State-of-the-Art Methods

Video Instance Segmentation (VIS). We benchmark CAVIS against leading methods on three established VIS datasets, as detailed in Tab. 2. CAVIS sets a new state-of-the-art, outperforming the previous top model, DVIS++, by margins of 1.1, 1.4, and 3.7 average precision (AP) points on YouTube-VIS2019, YouTube-VIS2021, and OVIS, respectively. Particularly noteworthy is CAVIS's performance on the OVIS dataset, where it significantly outstrips all competitors. This dataset is renowned for its diversity and the complexity of its video sequences. Fig. 3 illustrates how our

Table 3. Comparison on VIPSeg validation sets. 'Th' and 'St' denote 'things' and 'stuff' classes. † denotes the model trained with the temporal refiner [55].

Method	ResNet-50				ViT-L			
	VPQ	VPQ^{Th}	VPQ^{St}	STQ	VPQ	VPQ^{Th}	VPQ^{St}	STQ
VPSNet-SiamTrack[43]	17.2	17.3	17.3	21.1	-	-	-	-
VIP-Deeplab[38]	16.0	12.3	18.2	22.0	-	-	-	-
Clip-PanoFCN[32]	22.9	25.0	20.8	31.5	-	-	-	-
Video K-Net [27]	26.1	-	-	31.5	-	-	-	-
TarVIS[1]	33.5	39.2	28.5	43.1	-	-	-	-
Tube-Link[28]	39.2	-	-	39.5	-	-	-	-
Video-kMax [40]	38.2	-	-	39.9	-	-	-	-
DVIS [55]	39.4	38.6	40.1	36.3	-	-	-	-
DVIS++ [56]	41.9	41.0	42.7	38.5	56.0	58.0	54.3	49.8
Ours	42.4	43.1	41.8	39.7	56.9	60.1	54.2	51.0
DVIS † [55]	43.2	43.6	42.8	42.8	-	-	-	-
DVIS++ † [56]	**44.2**	**44.5**	**43.9**	**43.6**	**58.0**	**61.2**	**55.2**	**56.0**
Ours †	**45.3**	**47.5**	**43.4**	**45.3**	**58.5**	**63.1**	**54.5**	**56.1**

model proficiently tracks objects even in scenarios marked by severe occlusion. This capability underscores the strength of our context-aware video learning approach, which effectively leverages information from surrounding objects for accurate instance matching, even under severe occlusion.

Video Panoptic Segmentation (VPS). In the realm of VPS, CAVIS also achieves the best performance on the VIPSeg dataset as shown in Tab. 3. For the ResNet-50 backbone, it achieves 45.3 in both Video Panoptic Quality (VPQ) and Segmentation and Tracking Quality (STQ). For the ViT-L backbone, the figures reach 58.5 VPQ and 56.1 STQ, demonstrating substantial advancements. Specifically, our model shows significant gains in VPQ^{Th}—which assesses performance on 'thing' classes—with increases of 3.0 and 1.9 for ResNet-50 and ViT-L backbones, respectively, over the previous best models. These improvements highlight the versatility of our context-aware object matching strategy across various video segmentation tasks.

5.3. Ablation Study

We conduct ablation studies on the OVIS dataset [37] with the ResNet-50 [15] backbone, detailed in Tab. 4. We use the same experimental setting as those in the main experiments. To evaluate the segmentation network across these setups in Tab. 4-(a-c), we employ the minimal post-processing method proposed by MinVIS [19].

Ablation study on technical contributions of CAVIS. We conduct a series of experiments in Tab. 4-(a) to demonstrate the effectiveness of our key components: the context-aware instance tracker (CAIT) and prototypical cross-frame contrastive (PCC) loss.

Tab. 4-(a) includes six experiments (i- vi), where experiment (i) present our baseline performance of retraining MinVIS [19] and DVIS [55] to match our settings. For segmentation netowrk ($\mathcal{S}$), experiments (iii- iv) show that implementing contrastive learning with context-aware instance features results in a notable +3.3 AP improvement over the baseline MinVIS. Further performance boosts are noted when PCC is introduced in experiments (v- vi), which records the highest performance of 30.0 AP by utilizing both $\mathcal{L}_{\text{CTX}}$ and $\mathcal{L}_{\text{PCC}}$. This highlights the synergistic effect of integrating context-aware tracking with cross-frame contrastive loss, significantly enhancing the system's accuracy and effectiveness.

We also evaluate the tracking network ($\mathcal{T}$) by initially using each fixed pre-trained segmentation network listed in Tab. 4-(i- vi). Comparing setups (v) and (vi), our newly designed context-aware cross-attention improves performance by +2.3 AP over the standard cross-attention, achieving 37.6 AP. These findings validate the efficacy of the context-aware feature, confirming its significant advantages for instance matching in complex video scenarios.

Context filter. Given that images feature objects at various scales, identifying the optimal receptive field size that functions effectively across different scenarios is essential. Our experiments, detailed in Tab. 4-(b), explore the effects of varying context filter sizes from 3 to 11. The optimal performance is achieved with a filter size of 9; larger sizes led to decreased performance, suggesting that excessively large receptive fields may detract from effective object matching by homogenizing the context features across all objects. Extended analysis on this aspect is provided in our supplementary materials. Further investigation into different types of context filters shown in Tab. 4-(c) reveals that the average filter, which evenly reflects surrounding information, offers a well-defined benefit. In contrast, the learnable filter, which

Table 4. Ablation studies on each component of CAVIS.

(a) CAIT, PCC loss

	Segmenter ($\mathcal{S}$)			Tracker ($\mathcal{T}$)	
	$\mathcal{L}_{\text{CTX}}$	$\mathcal{L}_{\text{PCC}}$	AP	Context-aware matching	AP
(i)			26.4		33.2
(ii)		✓	28.1		34.2
(iii)	✓		29.7		34.8
(iv)	✓		29.7	✓	37.2
(v)	✓	✓	**30.0**		35.3
(vi)	✓	✓	**30.0**	✓	**37.6**

(b) Context filter size, the number of frames

Metric: AP		The numbe of frames		
		2	3	4
Filter size	3×3	27.3	27.3	27.2
	5×5	28.3	28.7	28.1
	7×7	28.7	29.3	28.1
	9×9	29.5	**30.0**	28.7
	11×11	28.7	29.1	28.0

(c) Context filter type in $\mathcal{S}$

Metric	Context filter type in $\mathcal{S}$			
	Average	Max	Median	Learnable
AP	**30.0**	29.4	29.6	28.7

(d) Context alignment in $\mathcal{T}$

Context alignment in $\mathcal{T}$	
✗	✓
33.2	**37.6**

(e) Value for CAIT

Value for CAIT	
Q	$\hat{Q}$
36.8	**37.6**

lacks specific directives on characterizing surrounding information, performs similarly to scenarios without enhanced context features, as demonstrated in Tab. 4-(a)-(ii). This indicates the importance of a clearly defined context filter in improving the segmentation and tracking accuracy.

The number of adjacent frames used during training. Our fundamental assumption is that the surrounding information between adjacent frames remains relatively stable, thereby aiding object matching. Tab. 4-(b) details the performance comparison based on the number of adjacent frames used during training. Utilizing three frames results in the highest performance, achieving 30.0 AP. Increasing the number of frames decreases performance, likely due to larger gaps between sampled frames which lead to more significant changes in the surrounding information, thus complicating object matching. Consequently, we have found that using three frames optimizes training effectiveness.

Context feature alignment. The tracking network aligns instance features and produces outputs in varying orders, necessitating the precise alignment of context features for accurate matching in subsequent frames. Misalignment of these features can lead to incorrect matching of context information for each object, significantly impacting performance. As demonstrated in Tab. 4-(d), misalignment results in a notable performance drop of 4.4 AP. This underscores the critical need for accurate alignment of context information to ensure robust object tracking performance.

Value for context-aware instance matching. Context-aware features (Q), which include information on instance features, could be used as the value in context-aware matching. However, during segmenter training, the instance features ($\hat{Q}$) specifically drive the segmentation prediction. Therefore, it is more effective to use instance features as the value for matching, leading to better performance, as shown in Tab. 4-(e).

6. Conclusion

In this paper, we introduce Context-Aware Video Instance Segmentation (CAVIS), a pioneering framework designed to enhance the accuracy and reliability of object tracking in complex video scenarios by integrating contextual information surrounding each instance. The introduction of the Context-Aware Instance Tracker (CAIT) and the innovative Prototypical Cross-frame Contrastive (PCC) loss are central to CAVIS's effectiveness. CAIT leverages the surrounding context to enrich the core features of each instance, providing a more holistic view that significantly improves instance detection and segmentation under challenging conditions. Simultaneously, PCC loss ensures consistency of these enriched features across frames, reinforcing the temporal linkage between instances and enhancing the overall tracking robustness. Our experiments across multiple challenging benchmarks demonstrate that CAVIS significantly outperforms existing state-of-the-art methods, particularly excelling in scenarios that demand robust tracking capabilities. The integration of CAIT and PCC not only addresses the primary challenges of occlusions and motion but also effectively manages the presence of visually similar objects that can often mislead traditional VIS approaches.

Acknowledgments

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.RS-2014-II140123, Development of High Performance Visual Big-Data Discovery Platform for Large-Scale Realtime Data Analysis), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00210908) and the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2025-02219277, AI Star Fellowship Support Project(DGIST)).

References

[1] Ali Athar, Alexander Hermans, Jonathon Luiten, Deva Ramanan, and Bastian Leibe. Tarvis: A unified approach for target-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18738–18748, 2023. 7

[2] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004. 1

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 941–951, 2019. 2

[4] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 1–18. Springer, 2020. 1, 2, 6

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[7] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 6

[8] 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, 3, 6

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34: 17864–17875, 2021. 1, 3

[10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1, 2, 3, 6

[11] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2

[12] Anwesa Choudhuri, Girish Chowdhary, and Alexander G. Schwing. Context-aware relative object queries to unify video instance and panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6377–6386, 2023. 2, 6

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6

[14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2896–2905, 2022. 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[17] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *Advances in Neural Information Processing Systems*, 35:23109–23120, 2022. 1, 2, 3, 6

[18] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14623–14632, 2023. 1, 2, 4, 6

[19] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Advances in Neural Information Processing Systems*, 35:31265–31277, 2022. 1, 2, 3, 6, 7

[20] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *Advances in Neural Information Processing Systems*, 34:13352–13363, 2021. 1, 2, 6

[21] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[22] Hojin Kim, Seunghun Lee, Hyeon Kang, and Sunghoon Im. Offline-to-online knowledge distillation for video instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 159–168, 2024. 1

[23] Hanjung Kim, Jaehyun Kang, Miran Heo, Sukjun Hwang, Seoung Wug Oh, and Seon Joo Kim. Visage: Video instance segmentation with appearance-guided enhancement. In *European Conference on Computer Vision*, pages 93–109. Springer, 2025. 1, 6

[24] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[25] Junlong Li, Bingyao Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. Tcovis: Temporally consistent online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1097–1107, 2023. 6

[26] Minghan Li, Shuai Li, Wangmeng Xiang, and Lei Zhang. Mdqe: Mining discriminative query embeddings to segment occluded instances on challenging videos. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10524–10533, 2023. 2

[27] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18847–18857, 2022. 7

[28] Xiangtai Li, Haobo Yuan, Wenwei Zhang, Guangliang Cheng, Jiangmiao Pang, and Chen Change Loy. Tube-link: A flexible cross tube framework for universal video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13923–13933, 2023. 2, 7

[29] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[30] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[31] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4133–4143, 2021. 2, 6

[32] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21033–21043, 2022. 7

[33] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 2

[34] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9226–9235, 2019. 2

[35] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007. 1

[36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[37] 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: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039, 2022. 2, 6, 7

[38] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3997–4008, 2021. 7

[39] Sarthak Sharma, Junaid Ahmed Ansari, J Krishna Murthy, and K Madhava Krishna. Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3508–3515. IEEE, 2018. 2

[40] Inkyu Shin, Dahun Kim, Qihang Yu, Jun Xie, Hong-Seok Kim, Bradley Green, In So Kweon, Kuk-Jin Yoon, and Liang-Chieh Chen. Video-kmax: A simple unified approach for online and near-online video panoptic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 229–239, 2024. 7

[41] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3539–3548, 2017. 2

[42] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8741–8750, 2021. 1, 2

[43] Sanghyun Woo, Dahun Kim, Joon-Young Lee, and In So Kweon. Learning to associate every segment for video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2705–2714, 2021. 7

[44] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *European Conference on Computer Vision*, pages 553–569. Springer, 2022. 1, 2, 6

[45] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, pages 588–605. Springer, 2022. 1, 2, 3, 6

[46] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 959–968, 2022. 6

[47] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3988–3998, 2019. 2

[48] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2

[49] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 1, 2, 6

[50] Linjie Yang, Yuchen Fan, Yang Fu, and Ning Xu. The 3rd large-scale video object segmentation challenge - video instance segmentation track, 2021. 2, 6

[51] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8043–8052, 2021. 1, 2, 6

[52] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2885–2895, 2022. 2

[53] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. Ctvis: Consistent training for online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 899–908, 2023. 1, 2, 3, 5, 6

[54] Ye Yu, Jialin Yuan, Gaurav Mittal, Li Fuxin, and Mei Chen. Batman: Bilateral attention transformer in motion-appearance neighboring space for video object segmentation. In *European Conference on Computer Vision*, pages 612–629. Springer, 2022. 2

[55] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. *arXiv preprint arXiv:2306.03413*, 2023. 1, 2, 3, 4, 5, 6, 7

[56] Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *arXiv preprint arXiv:2312.13305*, 2023. 2, 5, 6, 7

[57] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. 2

[58] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pages 474–490. Springer, 2020. 2

[59] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2