Tracking Anything with Decoupled Video Segmentation

Ho Kei Cheng[1†] Seoung Wug Oh[2] Brian Price[2] Alexander Schwing[1] Joon-Young Lee[2]
[1]University of Illinois Urbana-Champaign [2]Adobe Research

ICCV 2023

PRESENTER: JIAXUAN XIE

STRUCT GROUP SEMINAR

2023/12/24

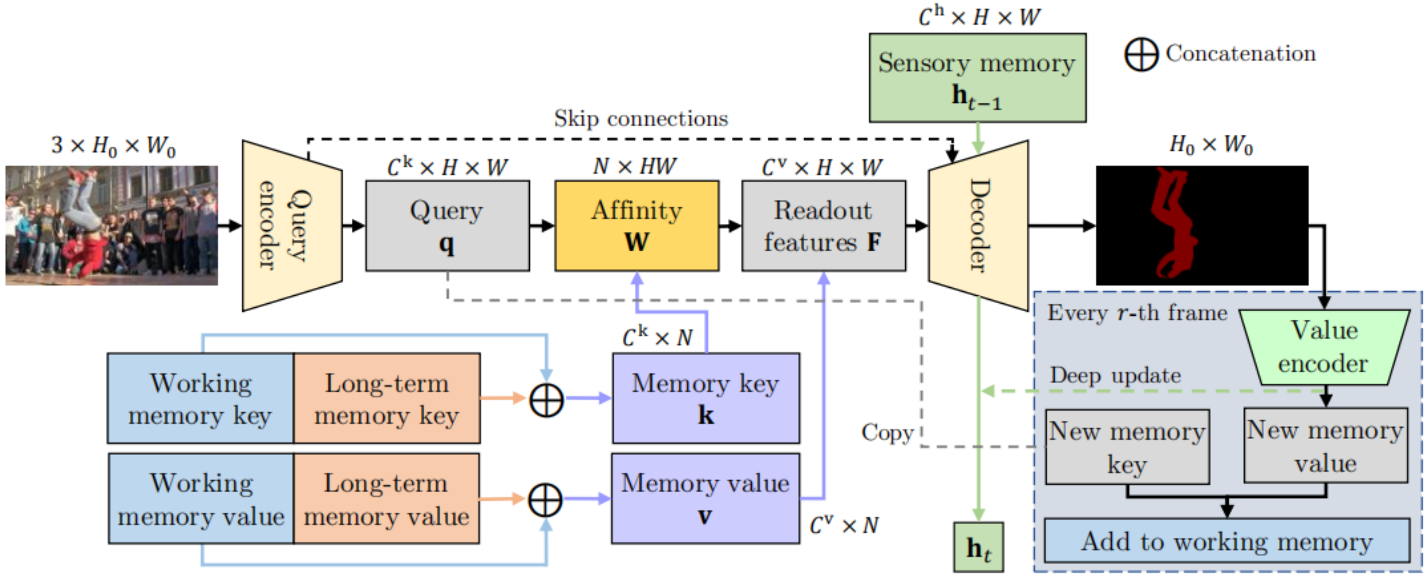
Outline

- Background
- Method
- Experiments
- Conclusion

Background

Video object segmentation (VOS)

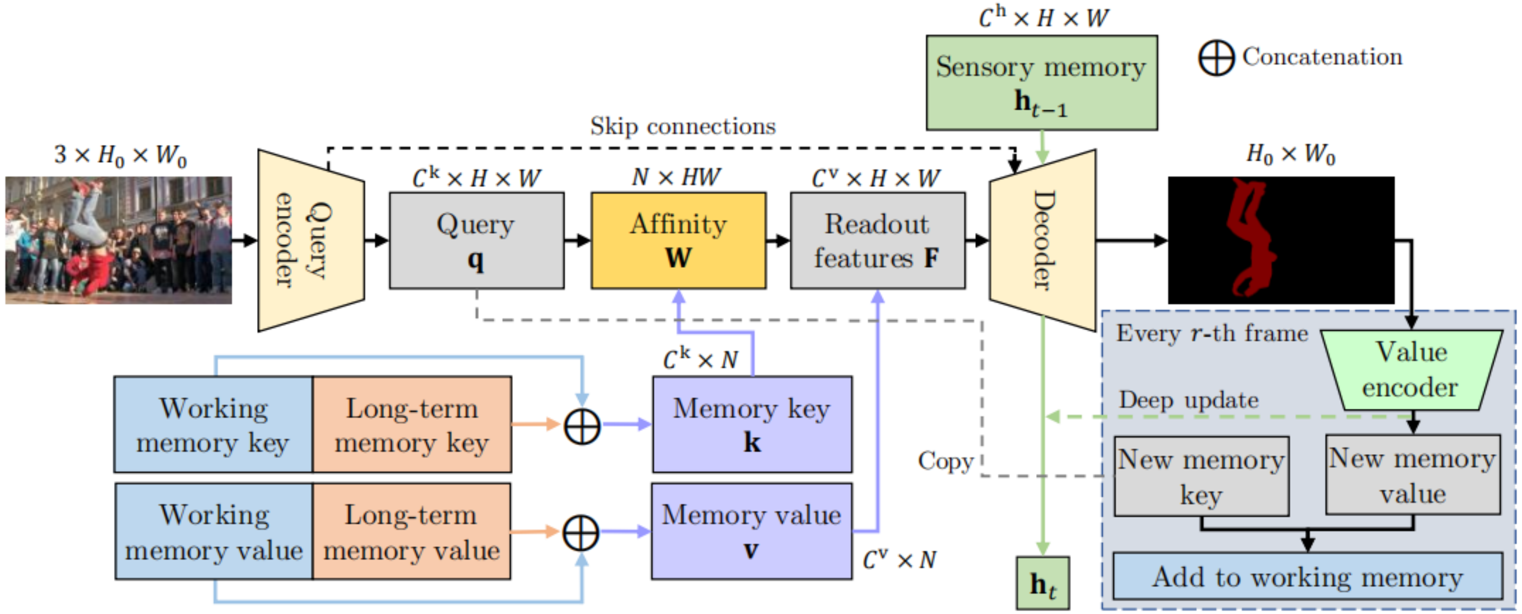
- Setting: semi-supervised, where a first-frame annotation is provided by the user, and the method segments objects in all other frames as accurately as possible while preferably running in real-time, online, and while having a small memory footprint even when processing long videos.



Background

Video object segmentation (VOS)

- $$\mathbf{F} = \mathbf{v}\mathbf{W}(\mathbf{k}, \mathbf{q}).$$



Method

- **Formulation**
  - image segmentation model $\text{Seg}(I_t) = \text{Seg}_t$
  - temporal propagation model $\text{Prop}(\mathbf{H}, I)$ (XMem as a propagation backbone)
  - represent a segmentation as a set of non-overlapping per-object binary segments $\mathbf{M}_t = \{m_i, 0 < i \leq |\mathbf{M}_t|\}$,
- **Bi-Directional Propagation**
  - In-clip consensus
  - Merging Propagation of Consensus

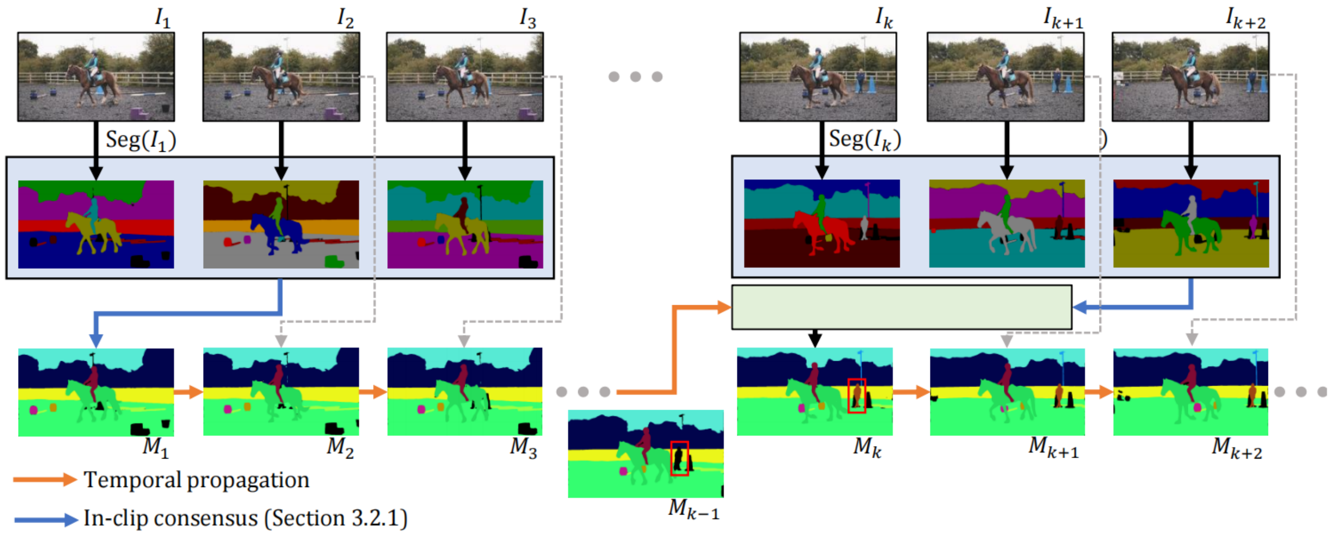
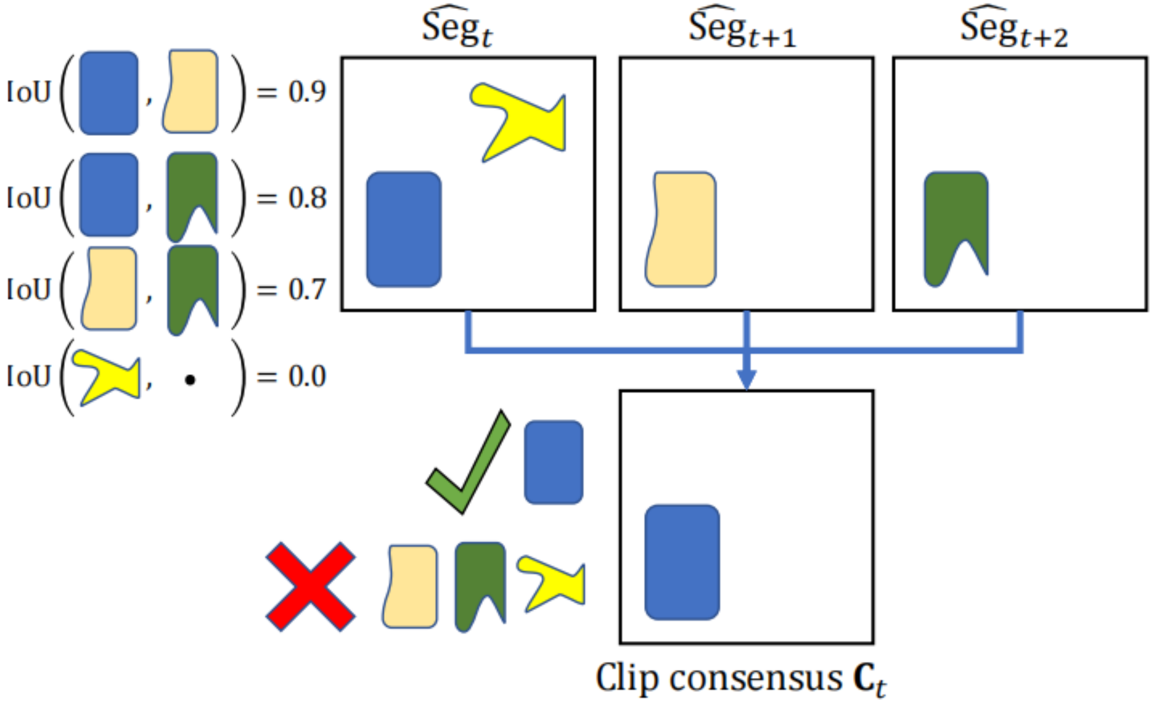


Figure 3. Overview of our framework. We first filter image-level segmentations with in-clip consensus (Section 3.2.1) and temporally propagate this result forward. To incorporate a new image segmentation at a later time step (for previously unseen objects, e.g., red box), we merge the propagated results with in-clip consensus as described in Section 3.2.2. Specifics of temporal propagation are in the appendix.

Method

In-clip consensus (Formulation)

- Input a set of n frames $(\widehat{\mathrm{Seg}}_t, \widehat{\mathrm{Seg}}_{t+1}, \ldots, \widehat{\mathrm{Seg}}_{t+n-1})$
- Output a denoised consensus $\mathbf{C}_t$
- **3 steps: Spatial Alignment**
  Representation
  Integer Programming



Method

In-clip consensus (Spatial Alignment)

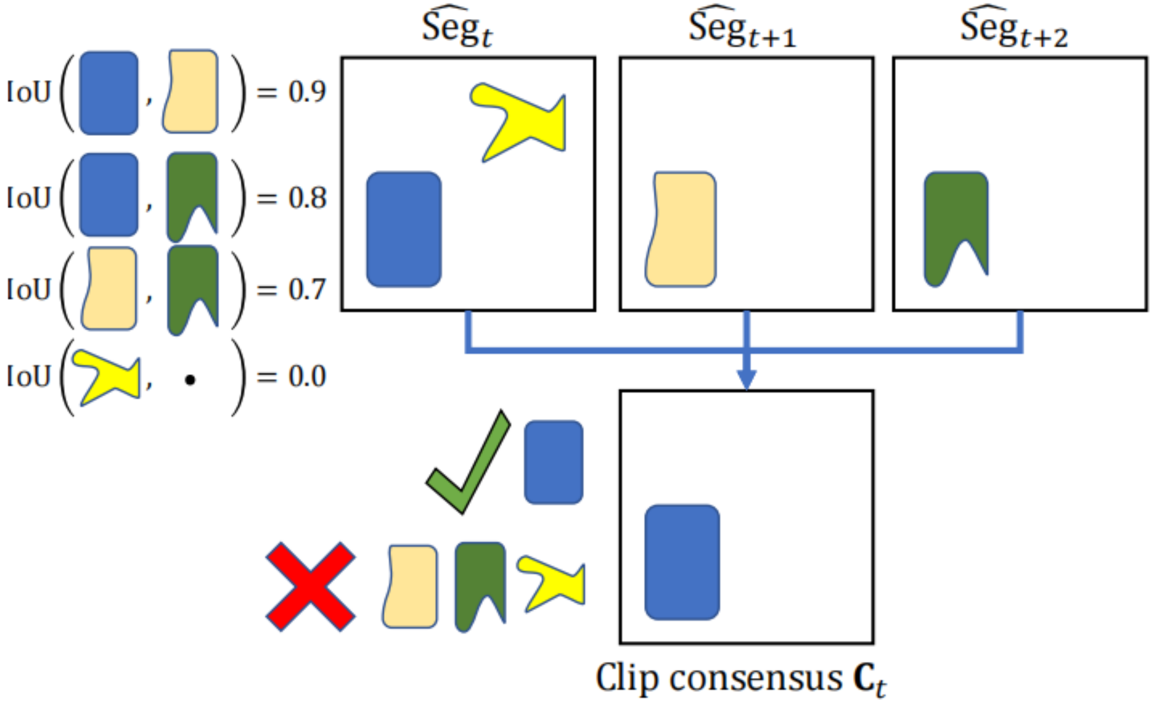
- re-use temporal propagation model

$$\widehat{\text{Seg}}_{t+i} = \text{Prop}\left(\{I_{t+i}, \text{Seg}_{t+i}\}, I_t\right), 0 < i < n.$$

In-clip consensus (Representation)

$$\mathbf{P} = \bigcup_{i=0}^{n-1} \widehat{\text{Seg}}_{t+i} = \{p_i, 0 < i \leq |\mathbf{P}|\}.$$

$$\mathbf{C}_t = \{p_i | v_i^* = 1\} = \{c_i, 0 < i \leq |\mathbf{C}|\}.$$



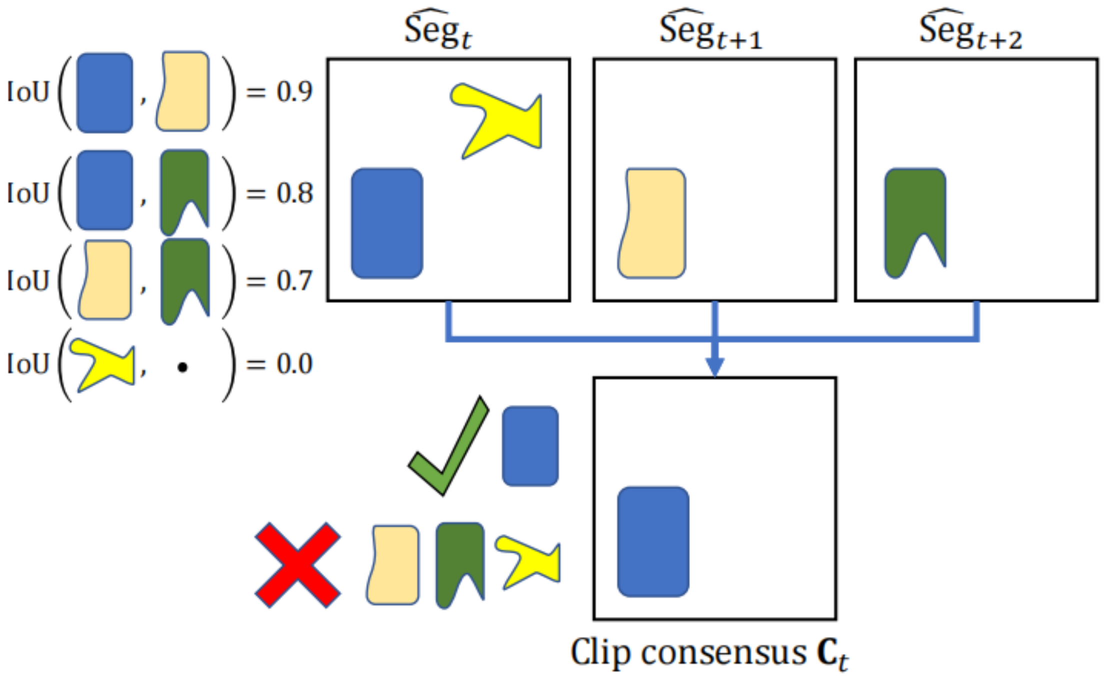
Method

In-clip consensus (Integer Programming.)

- two criteria:
  1. Lone proposals p_i are likely to be noise and should not be selected. Selected proposals should be supported by other (unselected) proposals.
  2. Selected proposals should not overlap significantly with each other.

- the process equals to:

$$v^* = \text{argmax}_v \sum_i (\text{Supp}_i + \text{Penal}_i) \text{ s.t.} \sum_{i,j} \text{Overlap}_{ij} = 0.$$



Method

In-clip consensus (Integer Programming.)

- understanding

$$v^* = \text{argmax}_v \sum_i (\text{Supp}_i + \text{Penal}_i) \ \text{s.t.} \sum_{i,j} \text{Overlap}_{ij} = 0.$$

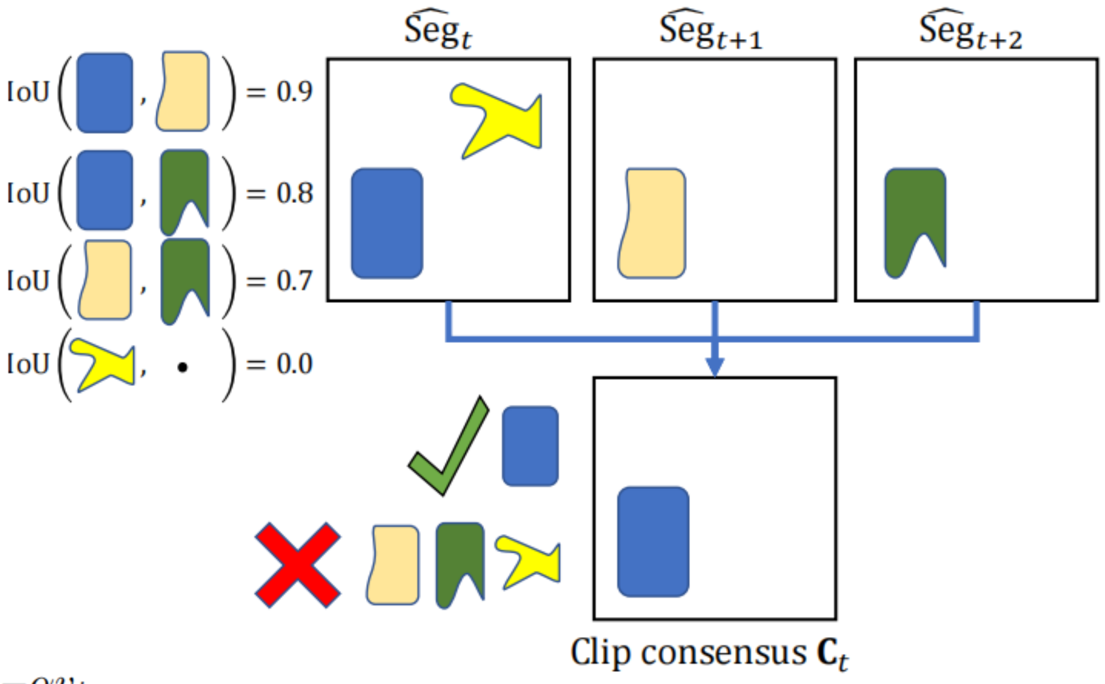
- support for proposal

$$\text{IoU}_{ij} = \text{IoU}_{ji} = \frac{|p_i \cap p_j|}{|p_i \cup p_j|}, 0 \le \text{IoU}_{ij} \le 1. \quad (5)$$

$$\text{Supp}_i = v_i \sum_j \begin{cases} \text{IoU}_{ij}, & \text{if IoU}_{ij} > 0.5 \text{ and } i \neq j \\ 0, & \text{otherwise} \end{cases}. \quad (6)$$

- do not select if overlap

$$\text{Overlap}_{ij} = \begin{cases} v_i v_j, & \text{if IoU}_{ij} > 0.5 \text{ and } i \neq j \\ 0, & \text{otherwise} \end{cases}. \qquad \text{Penal}_i = -\alpha v_i.$$



Method

Merging Propagation and Consensus

- propagation and consensus (past and future)

$$\mathrm{Prop}(\mathbf{H}, I_t) = \mathbf{R}_t = \{r_i, 0 < i \leq |\mathbf{R}|\} \quad \mathbf{C}_t = \{c_j, 0 < j \leq |\mathbf{C}|\}$$

- do not eliminate

$$\mathbf{M}_t = \{r_i \cup c_j | a_{ij} = 1\} \cup \{r_i | \forall_j a_{ij} = 0\} \cup \{c_j | \forall_i a_{ij} = 0\},$$

- maximizing association Iou

$$e_{ij} = \begin{cases} \mathrm{IoU}(r_i, c_j), & \text{if } \mathrm{IoU}(r_i, c_j) > 0.5 \\ -1, & \text{otherwise} \end{cases} . \qquad a_{ij} = 1 \text{ if } e_{ij} > 0 \text{ and } 0 \text{ otherwise}$$

- segment deletion

Experiments

- Comparison with end-to-end
- Pretrained on COCO panoptic dataset, fine-tuned on VIPSeg

Backbone				VPQ^1	VPQ^2	VPQ^4	VPQ^6	VPQ^8	VPQ^{10}	VPQ^∞	$\overline{VPQ}$	STQ
Clip-PanoFCN		end-to-end [45]	semi-online	27.3	26.0	24.2	22.9	22.1	21.5	18.1	21.1	28.3
Clip-PanoFCN		decoupled (ours)	online	29.5	28.9	28.1	27.2	26.7	26.1	25.0	26.4	35.7
Clip-PanoFCN		decoupled (ours)	semi-online	**31.3**	**30.8**	**30.1**	**29.4**	**28.8**	**28.3**	**27.1**	**28.4**	**35.8**
Video-K-Net	R50	end-to-end [34]	online	35.4	30.8	28.5	27.0	25.9	24.9	21.7	25.2	33.7
Video-K-Net	R50	decoupled (ours)	online	35.8	35.2	34.5	33.6	33.1	32.6	30.5	32.3	38.4
Video-K-Net	R50	decoupled (ours)	semi-online	37.1	36.5	35.8	35.1	34.7	34.3	32.3	33.9	38.6
Mask2Former	R50	decoupled (ours)	online	41.0	40.2	39.3	38.4	37.9	37.3	33.8	36.4	41.1
Mask2Former	R50	decoupled (ours)	semi-online	**42.1**	**41.5**	**40.8**	**40.1**	**39.7**	**39.3**	**36.1**	**38.3**	**41.5**
Video-K-Net	Swin-B	end-to-end [34]	online	49.8	45.2	42.4	40.5	39.1	37.9	32.6	37.5	45.2
Video-K-Net	Swin-B	decoupled (ours)	online	48.2	47.4	46.5	45.6	45.1	44.5	42.0	44.1	48.6
Video-K-Net	Swin-B	decoupled (ours)	semi-online	50.0	49.3	48.5	47.7	47.3	46.8	44.5	46.4	48.9
Mask2Former	Swin-B	decoupled (ours)	online	55.3	54.6	53.8	52.8	52.3	51.9	49.0	51.2	**52.4**
Mask2Former	Swin-B	decoupled (ours)	semi-online	**56.0**	**55.4**	**54.6**	**53.9**	**53.5**	**53.1**	**50.0**	**52.2**	52.2

Table 1. Comparisons of end-to-end approaches (e.g., state-of-the-art Video-K-Net [34]) with our decoupled approach on the large-scale video panoptic segmentation dataset VIPSeg [45]. Our method scales with better image models and performs especially well with large k where long-term associations are considered. All baselines are reproduced using official codebases.

Experiments

- In open-world video segmentation dataset BURST

Method		Validation			Test		
		OWTA_{all}	OWTA_{com}	OWTA_{unc}	OWTA_{all}	OWTA_{com}	OWTA_{unc}
Mask2Former	w/ Box tracker [2]	60.9	66.9	24.0	55.9	61.0	24.6
Mask2Former	w/ STCN tracker [2]	64.6	71.0	25.0	57.5	62.9	23.9
OWTB [39]		55.8	59.8	38.8	56.0	59.9	38.3
Mask2Former	w/ ours online	69.5	74.6	42.3	70.1	75.0	44.1
Mask2Former	w/ ours semi-online	**69.9**	**75.2**	41.5	**70.5**	**75.4**	44.1
EntitySeg	w/ ours online	68.8	72.7	49.6	69.5	72.9	53.0
EntitySeg	w/ ours semi-online	69.5	73.3	**50.5**	69.8	73.1	**53.3**

Table 2. Comparison to baselines in the open-world video segmentation dataset BURST [2]. 'com' stands for 'common classes' and 'unc' stands for 'uncommon classes'. Our method performs better in both – in the common classes with Mask2Former [7] image backbone, and in the uncommon classes with EntitySeg [49]. The agility to switch image backbones is one of the main advantages of our decoupled formulation. Baseline performances are transcribed from [2].

Experiments

- Referring video segmentation takes a text description of an object as input and segments the target object.

Method	Ref-DAVIS [25]	Ref-YTVOS [55]
URVOS [55]	51.6	47.2
ReferFormer [64]	60.5	62.4
VLT [17]	61.6	63.8
Ours	**66.3**	**66.0**

Table 3. $\mathcal{J}\&\mathcal{F}$ comparisons on two referring video segmentation datasets. Ref-YTVOS stands for Ref-YouTubeVOS [55].

Experiments

- Unsupervised video segmentation.
- DAVIS-16(single-object) and DAVIS-17(multi-object).

Method	D16-val	D17-val	D17-td
RTNet [54]	85.2	-	-
PMN [31]	85.9	-	-
UnOVOST [43]	-	67.9	58.0
Propose-Reduce [36]	-	70.4	-
Ours	**88.9**	**73.4**	**62.1**

Experiments

- Different hyperparameters

Varying clip size	VPQ^1	VPQ^{10}	$\overline{\text{VPQ}}$	STQ	FPS
$n = 1$	41.0	37.3	36.4	41.1	**10.3**
$n = 2$	40.4	37.2	36.3	39.0	9.8
$n = 3$	**42.1**	**39.3**	38.3	41.5	7.8
$n = 4$	**42.1**	39.1	**38.5**	42.3	6.6
$n = 5$	41.7	38.9	38.3	**42.8**	5.6
Varying merge freq.	VPQ^1	VPQ^{10}	$\overline{\text{VPQ}}$	STQ	FPS
Every 3 frames	**42.2**	39.2	**38.4**	**42.6**	5.2
Every 5 frames	42.1	**39.3**	38.3	41.5	7.8
Every 7 frames	41.5	39.0	35.7	40.5	**8.4**
Spatial Align?	VPQ^1	VPQ^{10}	$\overline{\text{VPQ}}$	STQ	FPS
Yes	**42.1**	**39.3**	**38.3**	**41.5**	7.8
No	36.7	33.9	32.8	33.7	**9.2**

Table 5. Performances of our method on VIPSeg [45] with different hyperparameters and design choices. By default, we use a clip size of $n = 3$ and a merge frequency of every 5 frames with spatial alignment for a balance between performance and speed.

Experiments

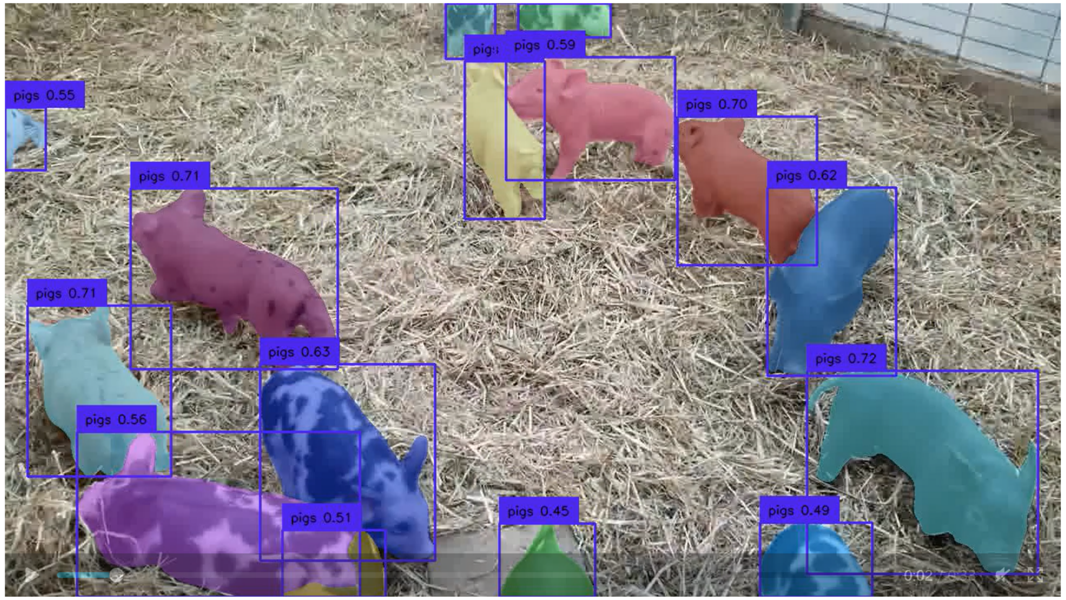
- bi-directional propagation

Temporal scheme	VPQ^1	VPQ^4	VPQ^{10}	$\overline{VPQ}$	STQ
Mask IoU	39.9	32.7	27.7	27.6	34.5
Mask IoU+flow	40.2	33.7	28.8	28.6	37.0
Query assoc.	40.4	33.1	28.1	28.0	35.8
'ShortTrack'	40.6	33.3	28.3	28.2	37.2
'TrustImageSeg'	40.3	37.5	33.7	33.2	37.9
Ours, bi-direction	**41.0**	**39.3**	**37.3**	**36.4**	**41.1**

Table 6. Performances of different temporal schema on VIPSeg [45]. Our bi-directional propagation scheme is necessary for the final high performance.

Experiments

Demo with Grounded Segment Anything (text prompt: "pigs"):



Source: https://youtu.be/FbK3SL97zf8

Demo with Segment Anything (automatic points-in-grid prompting); original video follows DEVA result overlaying the video:

Source: DAVIS 2017 validation set "soapbox"

Conclusion

- Using decoupled video segmentation that leverages external data, generalize better and able to incorporate existing universal segmentation models (like SAM)
- bi-directional propagation that denoises image segmentations and merges image segmentations with temporally propagated segmentations gracefully

Thanks for your listening!