One-Shot Video Object Segmentation

S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixe, D. Cremers, L.Van Gool

CVPR 2017

Authors

Sergi Caelles

Former PhD student at ETH Zurich, now at Google Research

Kevis-Kokitsi Maninis

Jordi Pont-Tuset

Laura Leal-Taixé

Daniel Cremers

Luc Van Gool

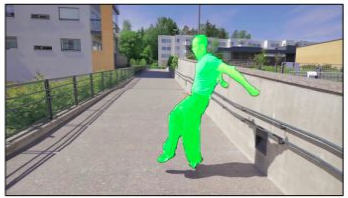
Professor at ETH Zurich & KU Leuven, Head Toyota Lab TRACE @ KUL & ETK

Problem Overview

- One-Shot Video Object Segmentation (OSVOS): Given the mask of the first frame, obtaining the segmentation for the rest frames.


frame 1


frame 21

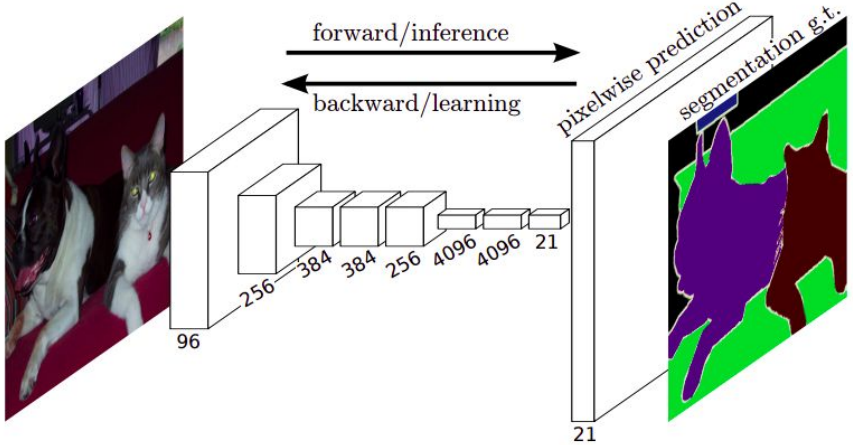

frame 41


frame 61


frame 81

Fully Convolutional Network (FCN) for Segmentation

- FCN can be constructed from a classification network by transforming its fully connected layers into convolutional layers.



“Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015

Fully Convolutional Network (FCN) for Segmentation

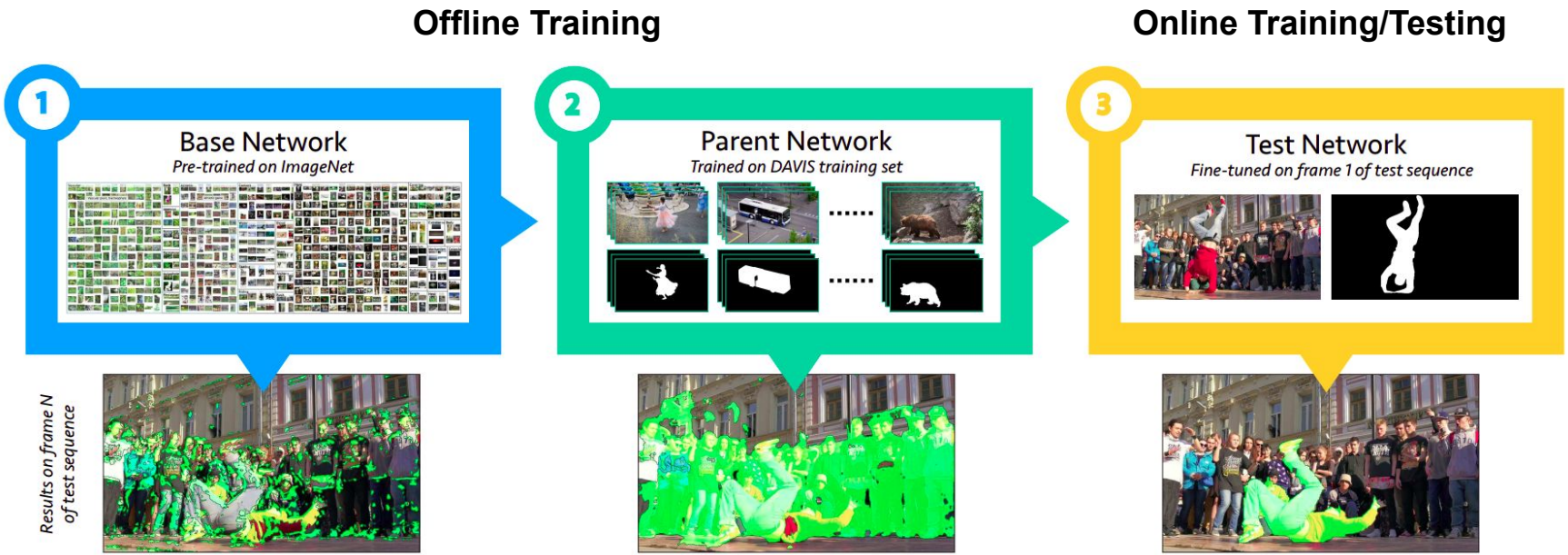
- Pixel-wise cross-entropy loss function for binary classification

$$\mathcal{L}(\mathbf{W}) = -\sum_j y_j \log P(y_j=1|X;\mathbf{W}) + (1-y_j)\log(1-P(y_j=1|X;\mathbf{W}))$$
$$= -\sum_{j\in Y_+} \log P(y_j=1|X;\mathbf{W}) - \sum_{j\in Y_-} \log P(y_j=0|X;\mathbf{W})$$

$$\mathcal{L}_{mod} = -\beta \sum_{j\in Y_+} \log P(y_j=1|X) - (1-\beta)\sum_{j\in Y_-} \log P(y_j=0|X)$$
$$\beta = |Y_-|/|Y|$$

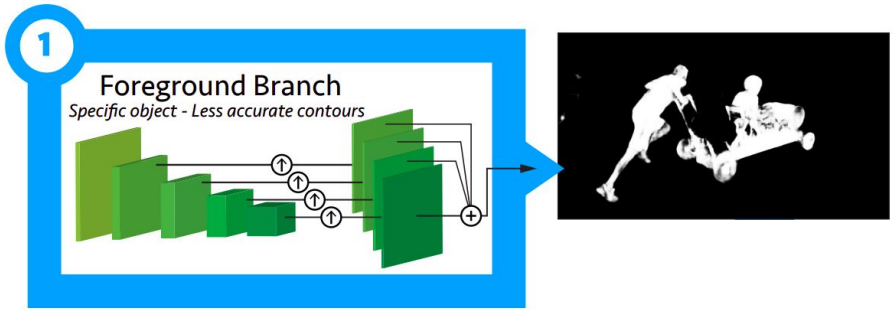
Methods

- The proposed OSVOS framework: One FCN trained three times.



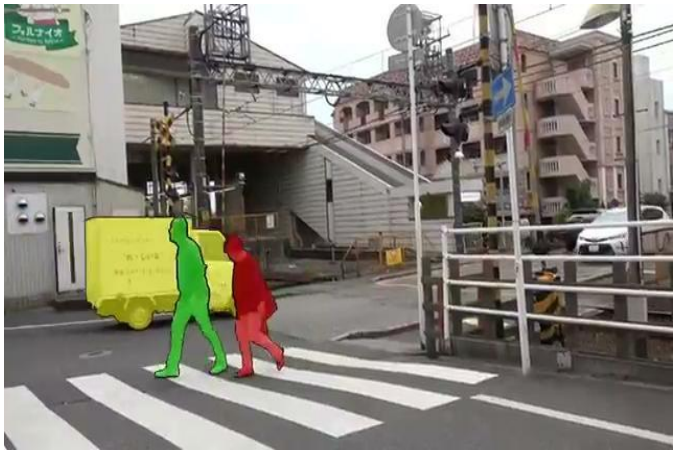
Methods

- Two-stream FCN architecture for improving contour localization

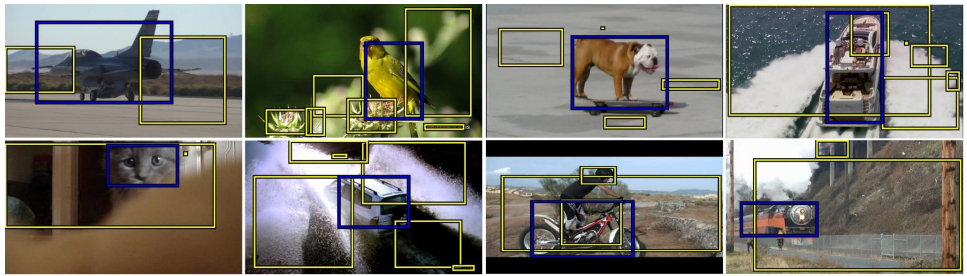


Datasets

- **DAVIS**: 50 full-HD video sequences with pixel-level annotation.
- **Youtube-Objects**: Videos of 10 object classes, each class contains 9-24 videos.



DAVIS



Youtube-Objects

Metrics

- Region Similarity (J): measures how well the pixels of two masks match

$$\mathcal{J} = \frac{|M \cap G|}{|M \cup G|}$$

- Contour Accuracy (F): measures the accuracy of the contours

$$\mathcal{F} = \frac{2 P_c R_c}{P_c + R_c}$$

- Temporal Instability (T): measures the undesired instability and jitter

"DAVIS: A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation", CVPR 2016

Metrics

- Error measure statistics: Mean, Recall, and Decay

$$\mathcal{J} = \frac{|M \cap G|}{|M \cup G|}$$

$S = \{S_i\}$ be the dataset of video sequences S_i and let $\bar{J}(S_i)$ be the error measure average on S_i.

$$Mean(S) = \frac{1}{|S|} \sum_{S_i} \bar{J}(S_i)$$

$$Recall(S) = \frac{1}{S} \sum_{S_i} 1_{\bar{J}(S_i) > \tau}$$

Let $Q_i = \{Q_i^1, .., Q_i^4\}$ be a partition of S_i in qartiles.

$$Decay(S) = \frac{1}{S} \sum_{Q_i} \bar{J}(Q_i^1) - \bar{J}(Q_i^4)$$

Results

- Ablation study on DAVIS: Comparison of OSVOS against downgraded versions without some of its components

	Measure	Ours	-BS		-PN-BS		-OS-BS		-PN-OS-BS	
$\mathcal{J}$	Mean $\mathcal{M}\uparrow$	**79.8**	77.4	*2.4*	64.6	*15.2*	52.5	*27.3*	17.6	*62.2*
	Recall $\mathcal{O}\uparrow$	**93.6**	91.0	*2.6*	70.5	*23.2*	57.7	*35.9*	2.3	*91.3*
	Decay $\mathcal{D}\downarrow$	14.9	17.4	*2.5*	27.8	*13.0*	**–1.9**	*16.7*	1.8	*13.1*
$\mathcal{F}$	Mean $\mathcal{M}\uparrow$	**80.6**	78.1	*2.5*	66.7	*13.9*	47.7	*32.9*	20.3	*60.4*
	Recall $\mathcal{O}\uparrow$	**92.6**	92.0	*0.6*	74.4	*18.3*	47.9	*44.7*	2.4	*90.2*
	Decay $\mathcal{D}\downarrow$	15.0	19.4	*4.5*	26.4	*11.4*	**0.6**	*14.3*	2.4	*12.6*
$\mathcal{T}$	Mean $\mathcal{M}\downarrow$	37.6	**33.5**	*4.0*	60.9	*23.3*	53.8	*16.2*	46.0	*8.4*

No Boundary Snapping

Results

- Ablation study on DAVIS: Comparison of OSVOS against downgraded versions without some of its components

	Measure	Ours	-BS		-PN-BS		-OS-BS		-PN-OS-BS	
$\mathcal{J}$	Mean $\mathcal{M}\uparrow$	**79.8**	77.4	*2.4*	64.6	*15.2*	52.5	*27.3*	17.6	*62.2*
	Recall $\mathcal{O}\uparrow$	**93.6**	91.0	*2.6*	70.5	*23.2*	57.7	*35.9*	2.3	*91.3*
	Decay $\mathcal{D}\downarrow$	14.9	17.4	*2.5*	27.8	*13.0*	**–1.9**	*16.7*	1.8	*13.1*
$\mathcal{F}$	Mean $\mathcal{M}\uparrow$	**80.6**	78.1	*2.5*	66.7	*13.9*	47.7	*32.9*	20.3	*60.4*
	Recall $\mathcal{O}\uparrow$	**92.6**	92.0	*0.6*	74.4	*18.3*	47.9	*44.7*	2.4	*90.2*
	Decay $\mathcal{D}\downarrow$	15.0	19.4	*4.5*	26.4	*11.4*	**0.6**	*14.3*	2.4	*12.6*
$\mathcal{T}$	Mean $\mathcal{M}\downarrow$	37.6	**33.5**	*4.0*	60.9	*23.3*	53.8	*16.2*	46.0	*8.4*

No Parent Network and BS

Results

- Ablation study on DAVIS: Comparison of OSVOS against downgraded versions without some of its components

	Measure	Ours	-BS		-PN-BS		-OS-BS		-PN-OS-BS	
$\mathcal{J}$	Mean $\mathcal{M}\uparrow$	**79.8**	77.4	*2.4*	64.6	*15.2*	52.5	*27.3*	17.6	*62.2*
	Recall $\mathcal{O}\uparrow$	**93.6**	91.0	*2.6*	70.5	*23.2*	57.7	*35.9*	2.3	*91.3*
	Decay $\mathcal{D}\downarrow$	14.9	17.4	*2.5*	27.8	*13.0*	**–1.9**	*16.7*	1.8	*13.1*
$\mathcal{F}$	Mean $\mathcal{M}\uparrow$	**80.6**	78.1	*2.5*	66.7	*13.9*	47.7	*32.9*	20.3	*60.4*
	Recall $\mathcal{O}\uparrow$	**92.6**	92.0	*0.6*	74.4	*18.3*	47.9	*44.7*	2.4	*90.2*
	Decay $\mathcal{D}\downarrow$	15.0	19.4	*4.5*	26.4	*11.4*	**0.6**	*14.3*	2.4	*12.6*
$\mathcal{T}$	Mean $\mathcal{M}\downarrow$	37.6	**33.5**	*4.0*	60.9	*23.3*	53.8	*16.2*	46.0	*8.4*

No One-Shot fine tuning and BS

Results

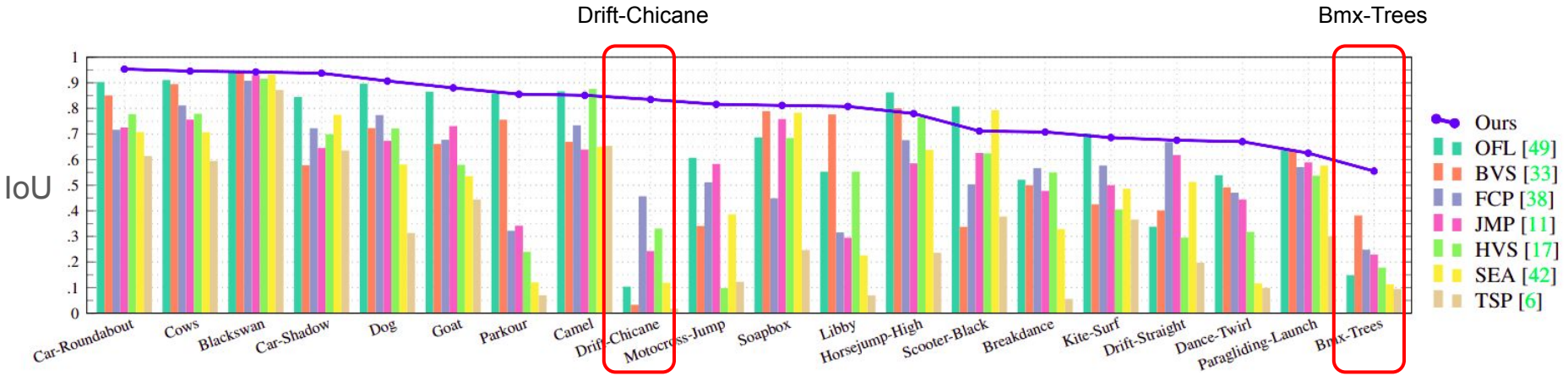
- Ablation study on DAVIS: Comparison of OSVOS against downgraded versions without some of its components

	Measure	Ours	-BS		-PN-BS		-OS-BS		-PN-OS-BS	
$\mathcal{J}$	Mean $\mathcal{M}\uparrow$	**79.8**	77.4	*2.4*	64.6	*15.2*	52.5	*27.3*	17.6	*62.2*
	Recall $\mathcal{O}\uparrow$	**93.6**	91.0	*2.6*	70.5	*23.2*	57.7	*35.9*	2.3	*91.3*
	Decay $\mathcal{D}\downarrow$	14.9	17.4	*2.5*	27.8	*13.0*	**–1.9**	*16.7*	1.8	*13.1*
$\mathcal{F}$	Mean $\mathcal{M}\uparrow$	**80.6**	78.1	*2.5*	66.7	*13.9*	47.7	*32.9*	20.3	*60.4*
	Recall $\mathcal{O}\uparrow$	**92.6**	92.0	*0.6*	74.4	*18.3*	47.9	*44.7*	2.4	*90.2*
	Decay $\mathcal{D}\downarrow$	15.0	19.4	*4.5*	26.4	*11.4*	**0.6**	*14.3*	2.4	*12.6*
$\mathcal{T}$	Mean $\mathcal{M}\downarrow$	37.6	**33.5**	*4.0*	60.9	*23.3*	53.8	*16.2*	46.0	*8.4*

It failed.

Results

- Comparison with state-of-the-art methods on DAVIS Validation

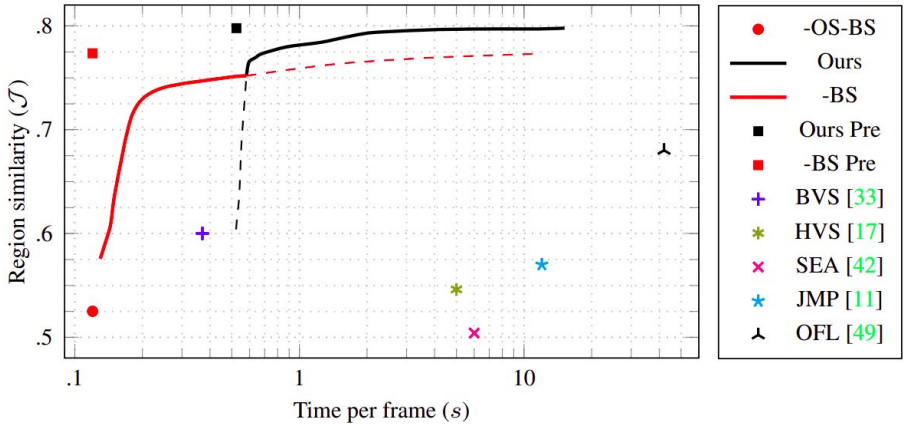


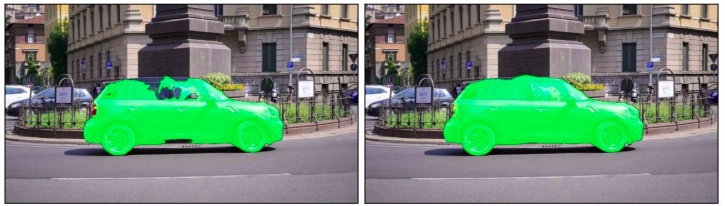
Results

- Qualitative results on the two challenging cases

Results

- Trade-off between speed and accuracy



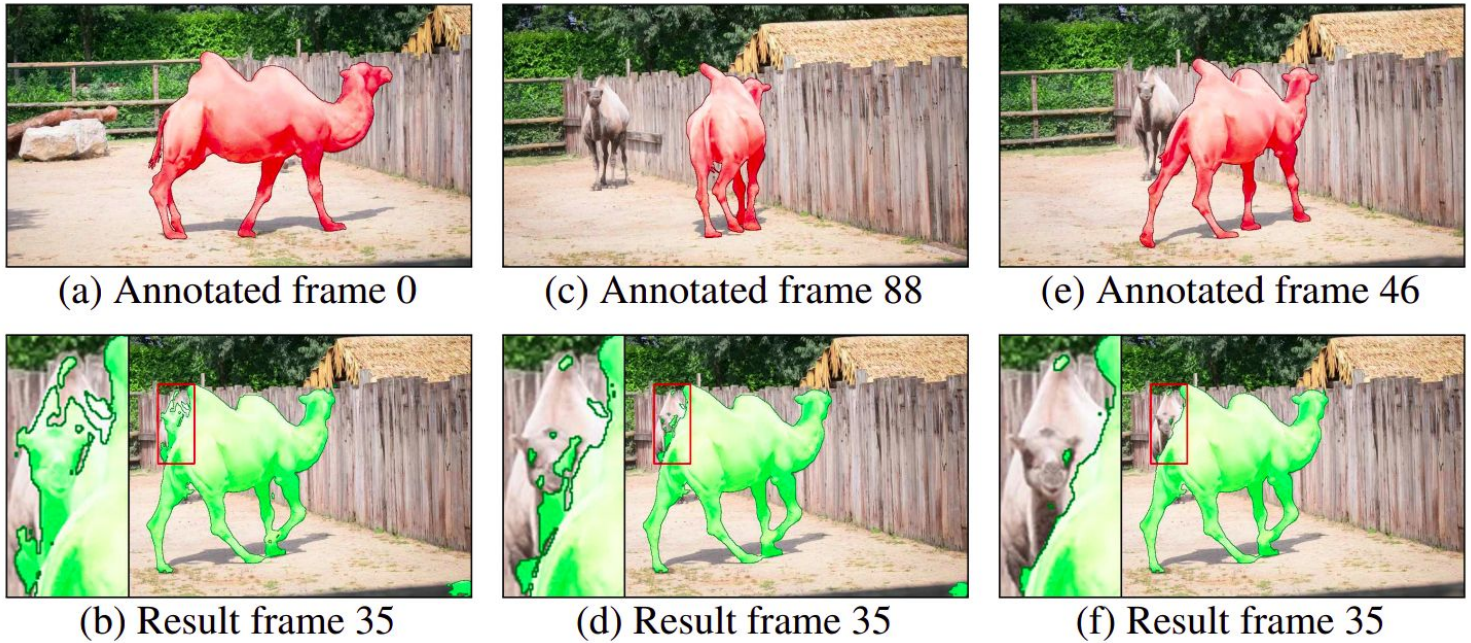


Fine tuning 10 seconds

Fine tuning 60 seconds

Results

- Progressive refinement: annotating more than one frame for fine tuning



(a) Annotated frame 0 (c) Annotated frame 88 (e) Annotated frame 46

(b) Result frame 35 (d) Result frame 35 (f) Result frame 35

Results

- Progressive refinement: annotating more than one frame for fine tuning

Annotations	0	1	2	3	4	5	All
Quality ($\mathcal{J}$)	58.5	79.8	84.6	85.9	86.9	87.5	88.7

Contributions

- This work adapts the CNN to a particular object instance given a single annotated image (hence *one-shot*).
- OSVOS processes each frame of a video independently, obtaining temporal consistency as a by-product rather than as the result of an explicitly imposed, expensive constraint.
- OSVOS can work at various points of the trade-off between speed and accuracy.

Critiques

- Online fine-tuning is not time-efficient, and thus may be impractical for real applications.

Questions

- Why this method can beat state-of-the-art methods though it doesn’t assume temporal constraints like other methods normally do?
- Is it possible to further extend this method for zero-shot video segmentation? Meaning the query and support images can be same object in different domains (i.e., same objects in different background).