

# LEARNING VIDEO OBJECT SEGMENTATION FROM STATIC IMAGES

(\* equal contribution) \*F. Perazzi<sup>1,2</sup> \*A. Khoreva<sup>3</sup> R. Benenson<sup>3</sup> B. Schiele<sup>3</sup> A. Sorkine-Hornung<sup>1</sup>

<sup>1</sup>Disney Research, <sup>2</sup>ETH Zürich, <sup>3</sup>Max Planck Institute for Informatics



Disney Research



max planck institut  
informatik

ETH zürich



## INTRODUCTION

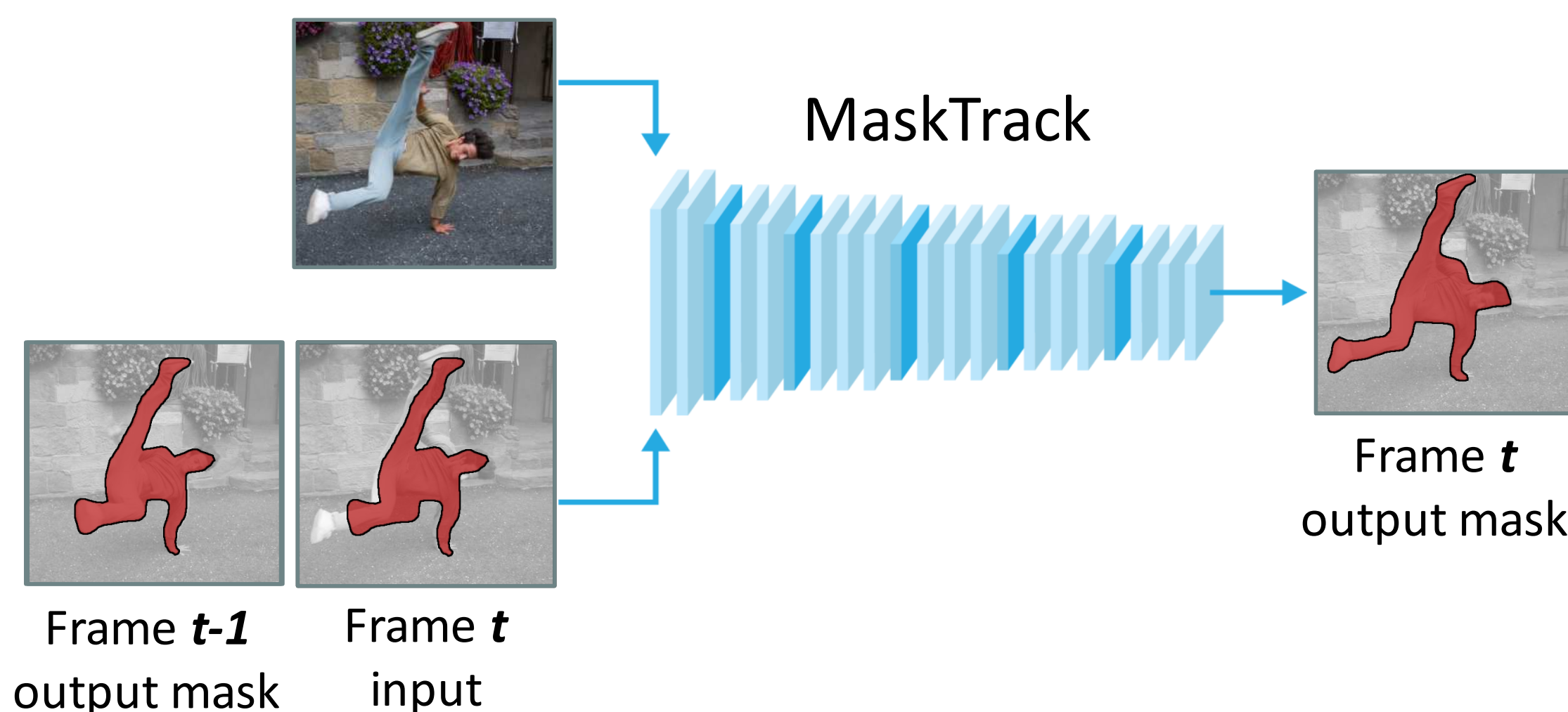
Goal: Separating **foreground objects** from the **background** in a video given the **1<sup>st</sup> frame mask annotation**.



1<sup>st</sup> frame

t

- Guided CNN trained for video object segmentation using only static images.



- Handle different types of input annotations such as **bounding boxes** and **segments** while leveraging an arbitrary amount of annotated frames.
- Competitive results on three different datasets, independently from the type of input annotation



Visit the project pages:

<https://graphics.ethz.ch/~perazzif/masktrack>

<https://www.mpi-inf.mpg.de/masktrack>

Source code and trained models are available online.

## LEARNING FROM STATIC IMAGES

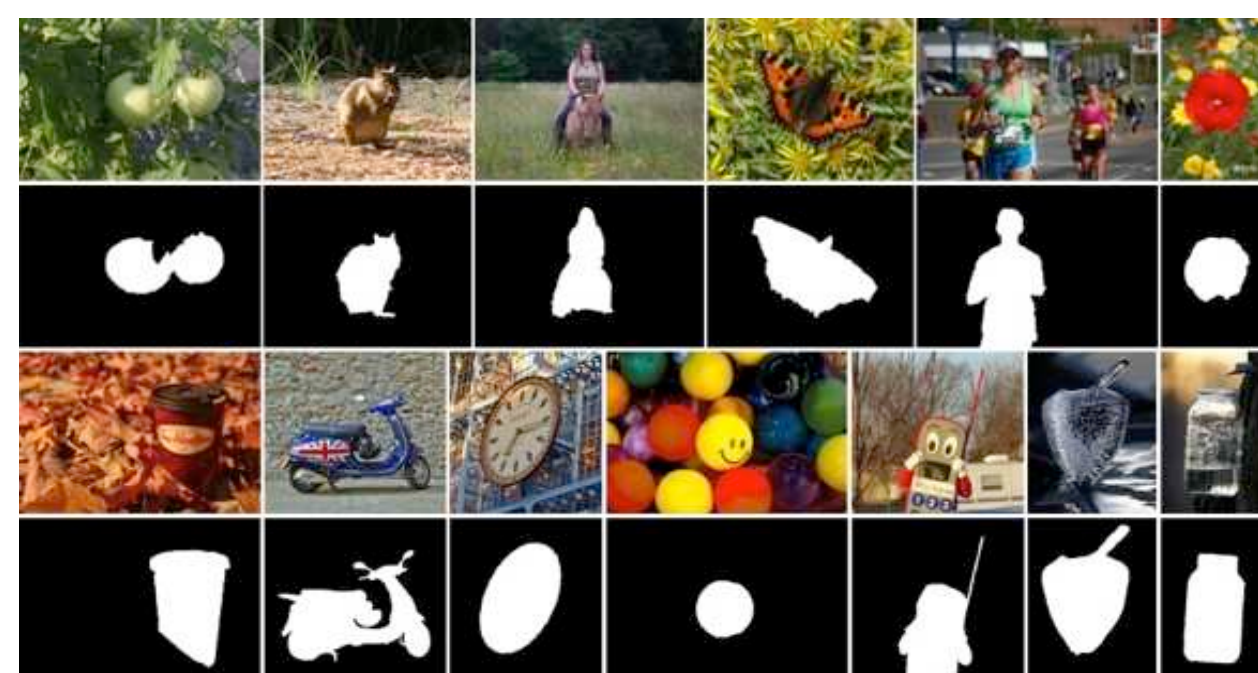
Video Dataset



DAVIS 2016 [Perazzi et al.'16] 50\* videos, 4K images

SegTrack-v2 [Li et al.'13] 14 videos

Image Dataset



MSRA10K [Cheng et al.'14], 10K images

→ lack of large-scale diverse video data for training

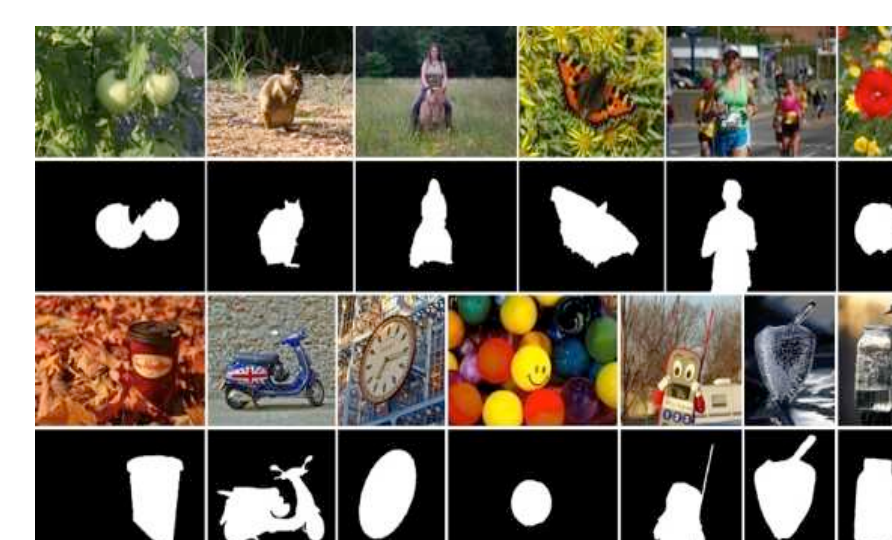
## TRAINING PHASES

1 Pre-Training



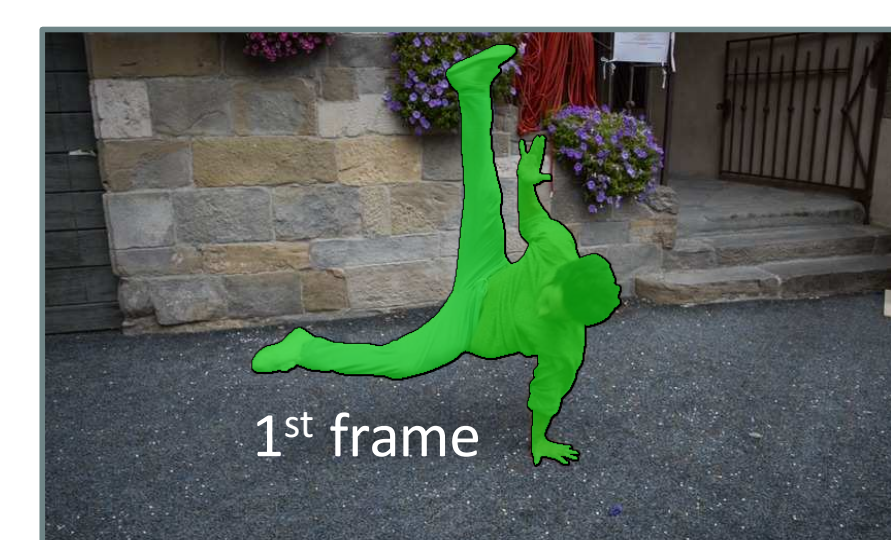
IMAGENET

2 Online Training



MSRA10K

3 Online Fine-tuning



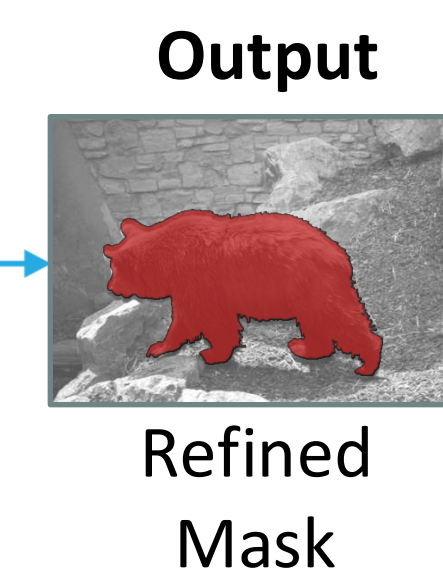
Tuning on the 1<sup>st</sup> frame  
mask of the test video

## OFFLINE TRAINING

Guided Instance Segmentation

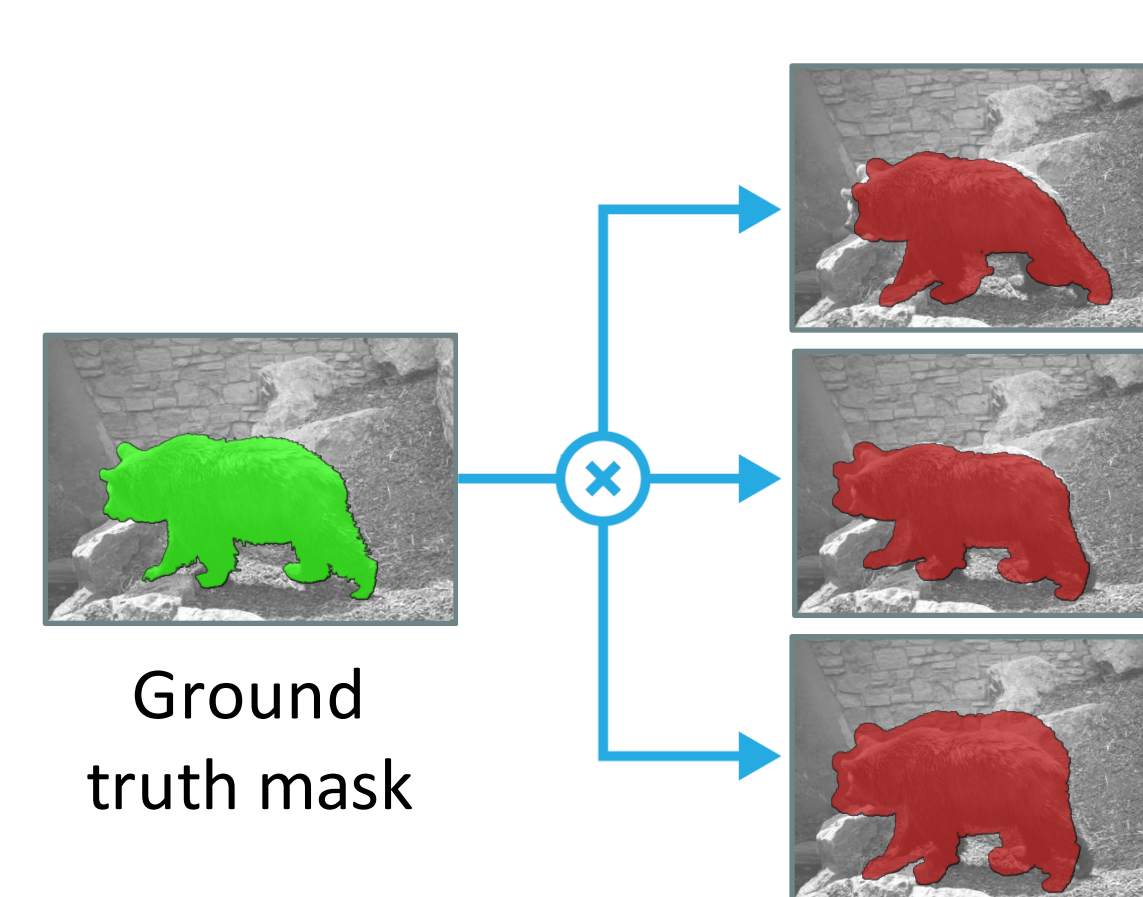


MaskTrack



Output

Previous Frame Mask Generation



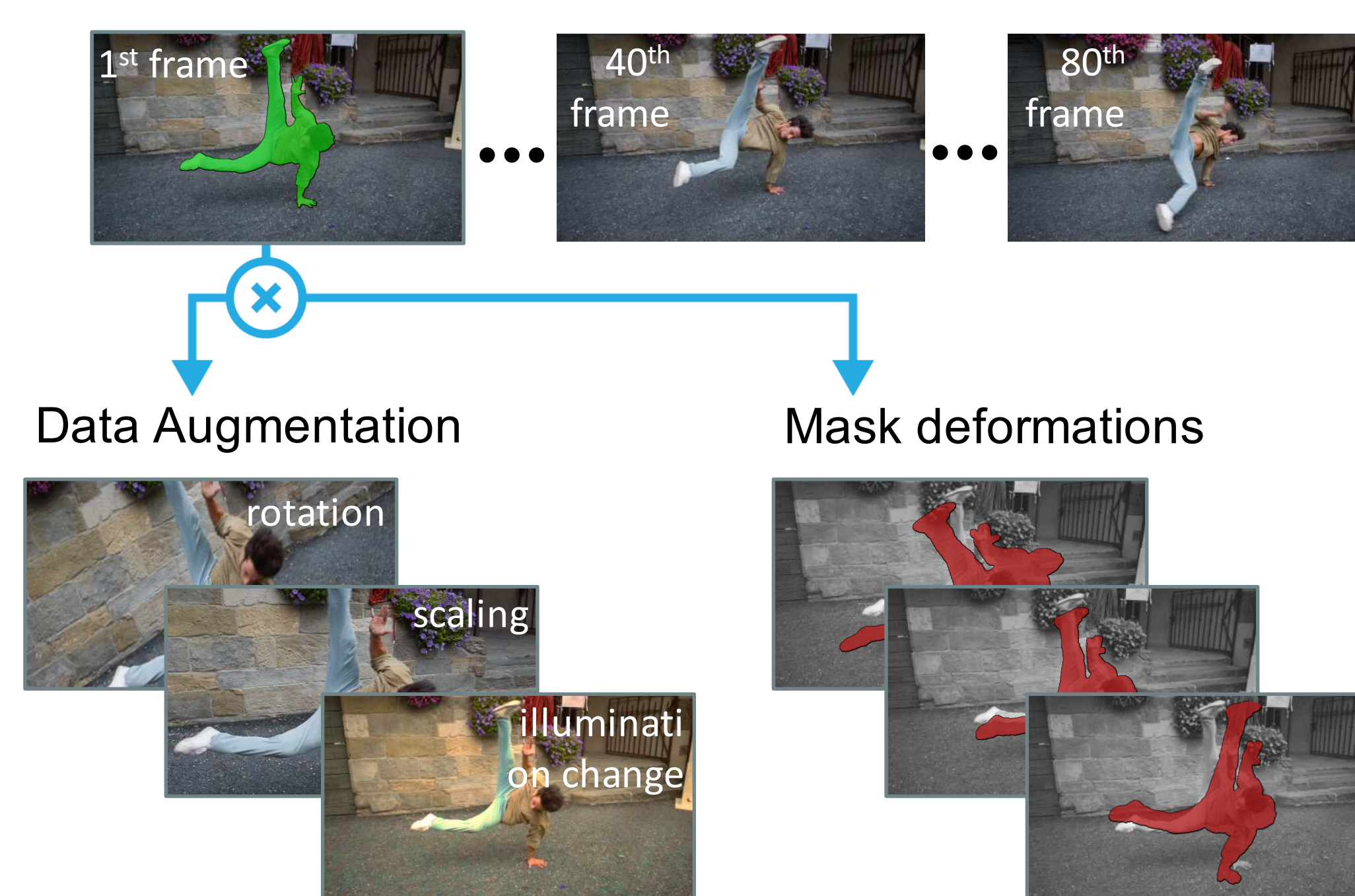
Ground truth mask

→ we train using *static images* only

→ Non-rigid and affine deformations to simulate the previous frame mask

## ONLINE FINE-TUNING

1<sup>st</sup> Frame Augmentation



## EVALUATION

Method	DAVIS, mIoU					
	Mean ↑	Recall ↑	Decay ↓	Mean ↑	Recall ↑	Decay ↓
Box oracle	45.1	39.7	-0.7	21.4	6.7	1.8
Grabcut oracle	67.3	76.9	1.5	65.8	77.2	2.9
NLC [3]	64.1	73.1	8.6	59.3	65.8	8.6
FCP [9]	63.1	77.8	3.1	54.6	60.4	3.9
BVS [6]	66.5	76.4	26.0	65.6	77.4	23.6
ObjFlow [12]	71.1	80.0	22.7	67.9	78.0	24.0
MaskTrack	74.8	87.8	14.1	75.0	84.7	14.3
MaskTrack+Flow+CRF	80.3	93.5	8.9	75.8	88.2	9.5

Method	Dataset, mIoU		
	DAVIS	YoutObjs	SegTrack-v2
Box oracle	45.1	55.3	56.1
Grabcut oracle	67.3	67.6	74.2
ObjFlow [49]	71.4	70.1	67.5
BVS [29]	66.5	59.7	58.4
NLC [15]	64.1	-	-
FCP [35]	63.1	-	-
W16 [50]	-	59.2	-
Z15 [56]	-	52.6	-
TRS [53]	-	-	69.1
MaskTrack	74.8	71.7	67.4
MaskTrack <sub>Box</sub>	73.7	69.3	62.4

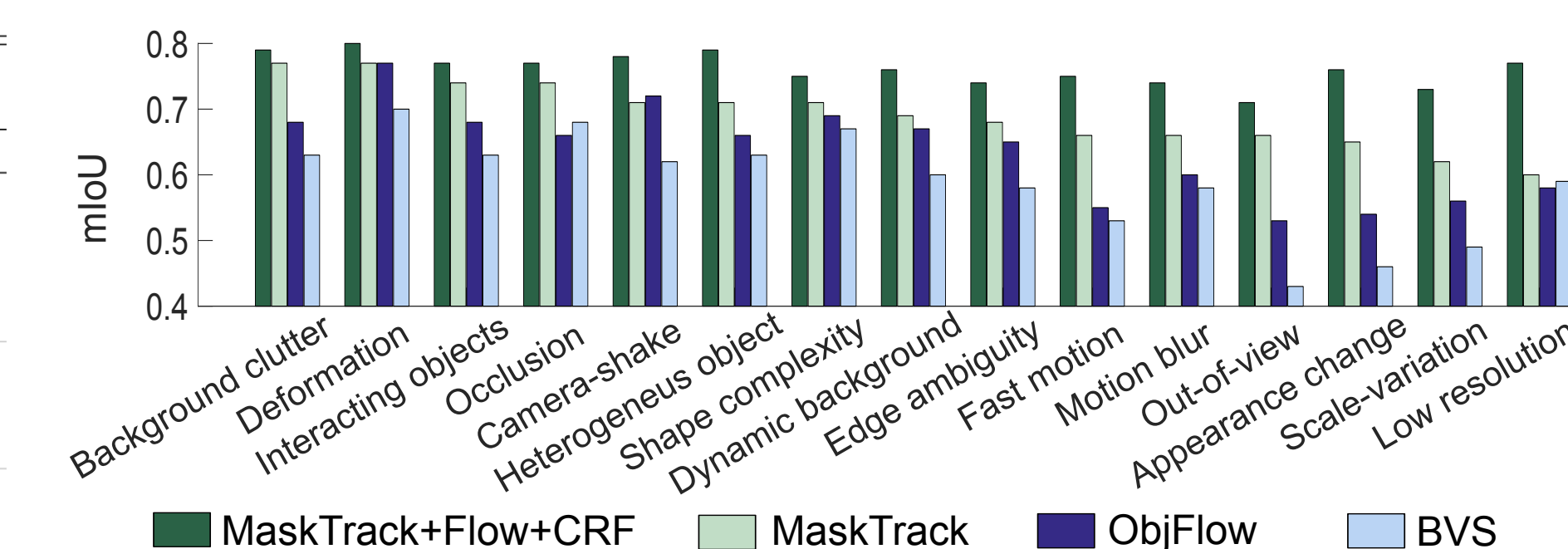
Competitive results despite using the same model and parameters across all videos.

## ANALYSIS

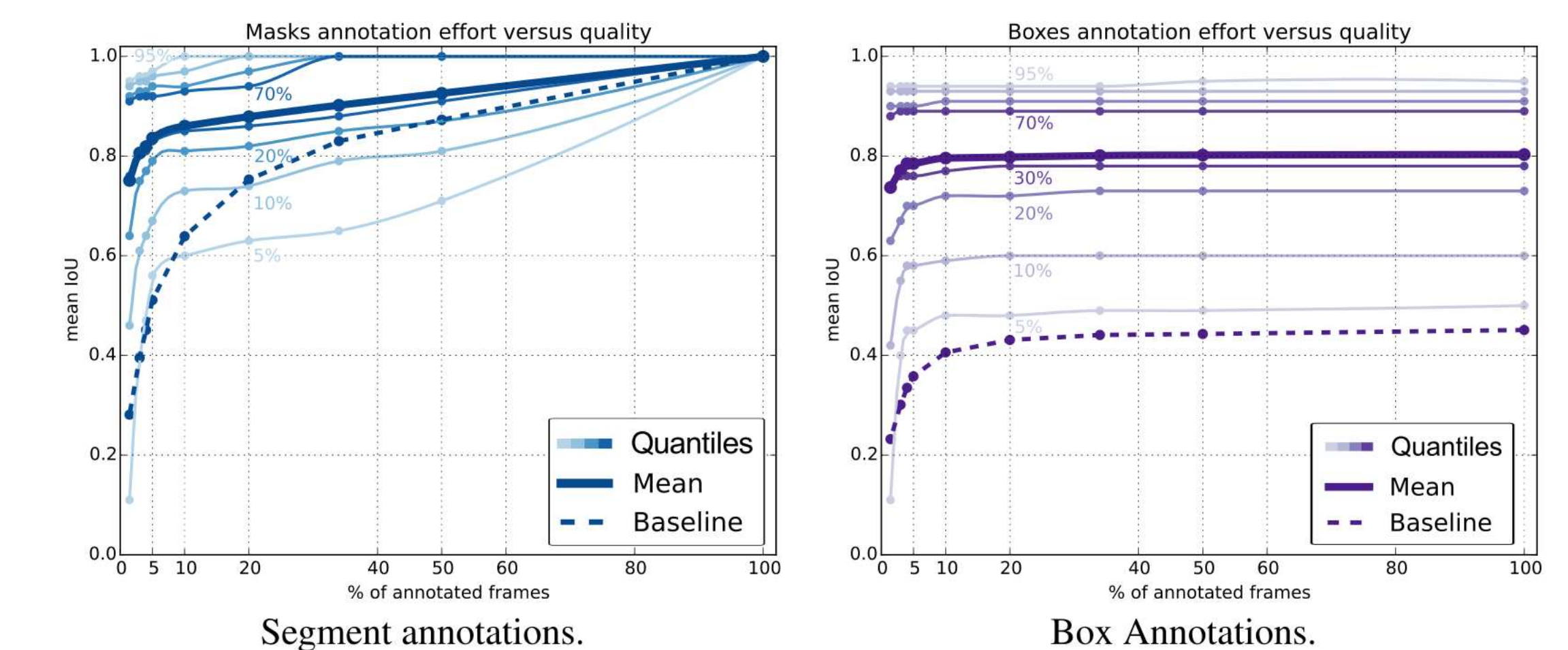
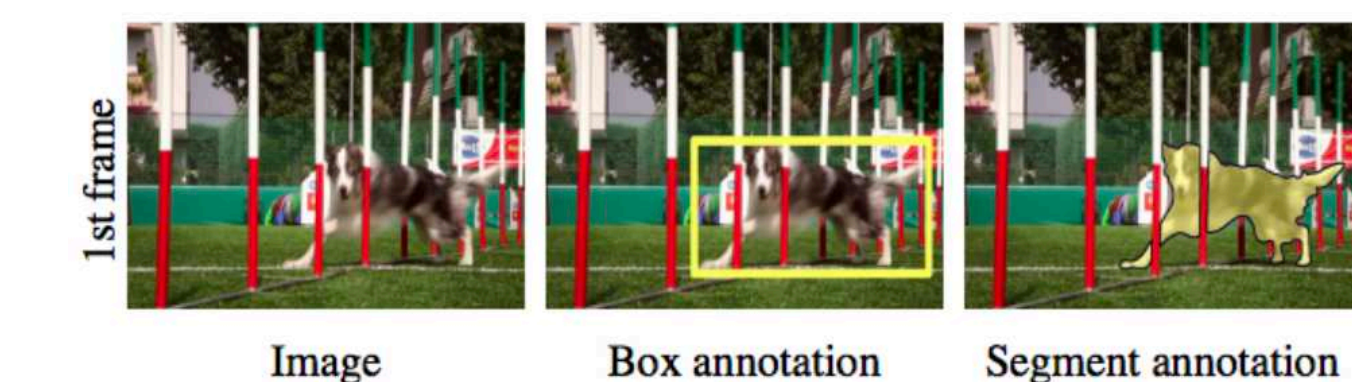
Ablation Study

Aspect	System variant	mIoU	ΔmIoU
Add-ons	MaskTrack+Flow+CRF	80.3	+1.9
	MaskTrack+Flow	78.4	+3.6
	MaskTrack	74.8	-
Training	No online fine-tuning	69.9	-4.9
	No offline training	57.6	-17.2
	Reduced offline training	73.2	-1.6
	Training on video	72.0	-2.8
Mask deformation	No dilation	72.4	-2.4
	No deformation	17.1	-57.7
	No non-rigid deformation	73.3	-1.5
Input channel	Boxes	69.6	-5.2
	No input	72.5	-2.3

Attribute Based Analysis



## SUPERVISION EFFORT



## REFERENCES

- ObjFlow**: Video Segmentation via Object Flow, Yi-Hsuan Tsai et al, CVPR 201
- NLC**: Video segmentation by non-local consensus voting, A. Faktor and M. Irani BMVC 2014
- BVS**: Bilateral Space Video Segmentation, N. Maerki et al, CVPR 2016
- FCP**: Fully connected object proposals for video segmentation, F. Perazzi et al, ICCV 2015
- TRS**: Track and segment: An iterative unsupervised approach for video object proposals, Xiao et al, CVPR 2016
- W16**: Semi-supervised domain adaptation for weakly labeled semantic video object segmentation, Chan et al.
- Z15**: Semantic object segmentation via detection in weakly labeled video, Zhang et al, CVPR 2015

Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs, Chen et al, ICLR 2015.