A brief Overview of Video Object Segmentation

Jia Zheng

SIST, ShanghaiTech

January 27, 2018

Disclaimer

Some of the material is naturally taken from online material, including

1. Blog of Eddie Smolyansky
2. DAVIS Challenge [3, 5]
3. OSVOS [1], MaskTrack [4], LucidTracker [2], OnAVOS [6]

Outline

Outline

Object Video Segmentation

Given the mask of the first frame, separate objects of interest in a video.

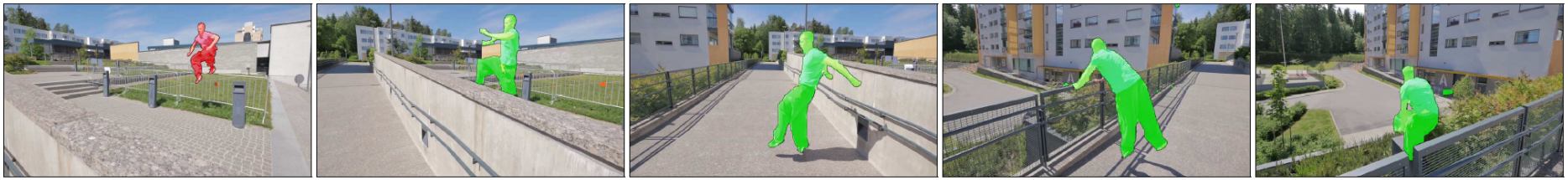

Figure: Example results of OSVOS [1]. Figure extracted from OSVOS [1]

Sub-divisions of Segmentation

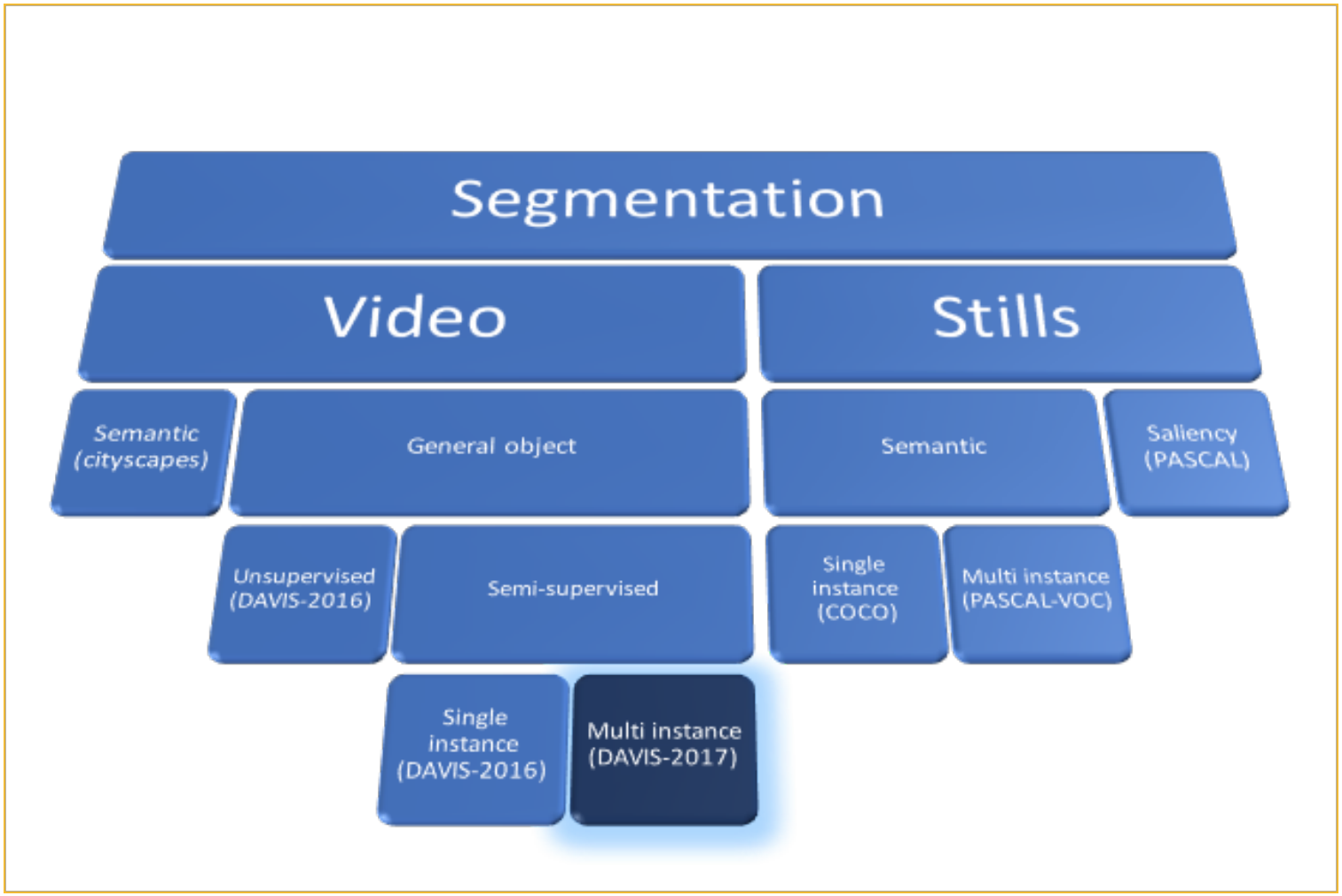

Figure: Sub-divisions of segmentation. Figure extracted from Eddie's blog.

Benchmark

Table: Benchmark Comparison

Benchmark	*Sequences*	*Frames*	*Objects*
YouTubeObjects	-	-	10
SegTrackv2	14	-	-
DAVIS 2016	50	3455	50
DAVIS 2017	150	10474	384
GyGo	155	-	-

DAVIS Benchmark

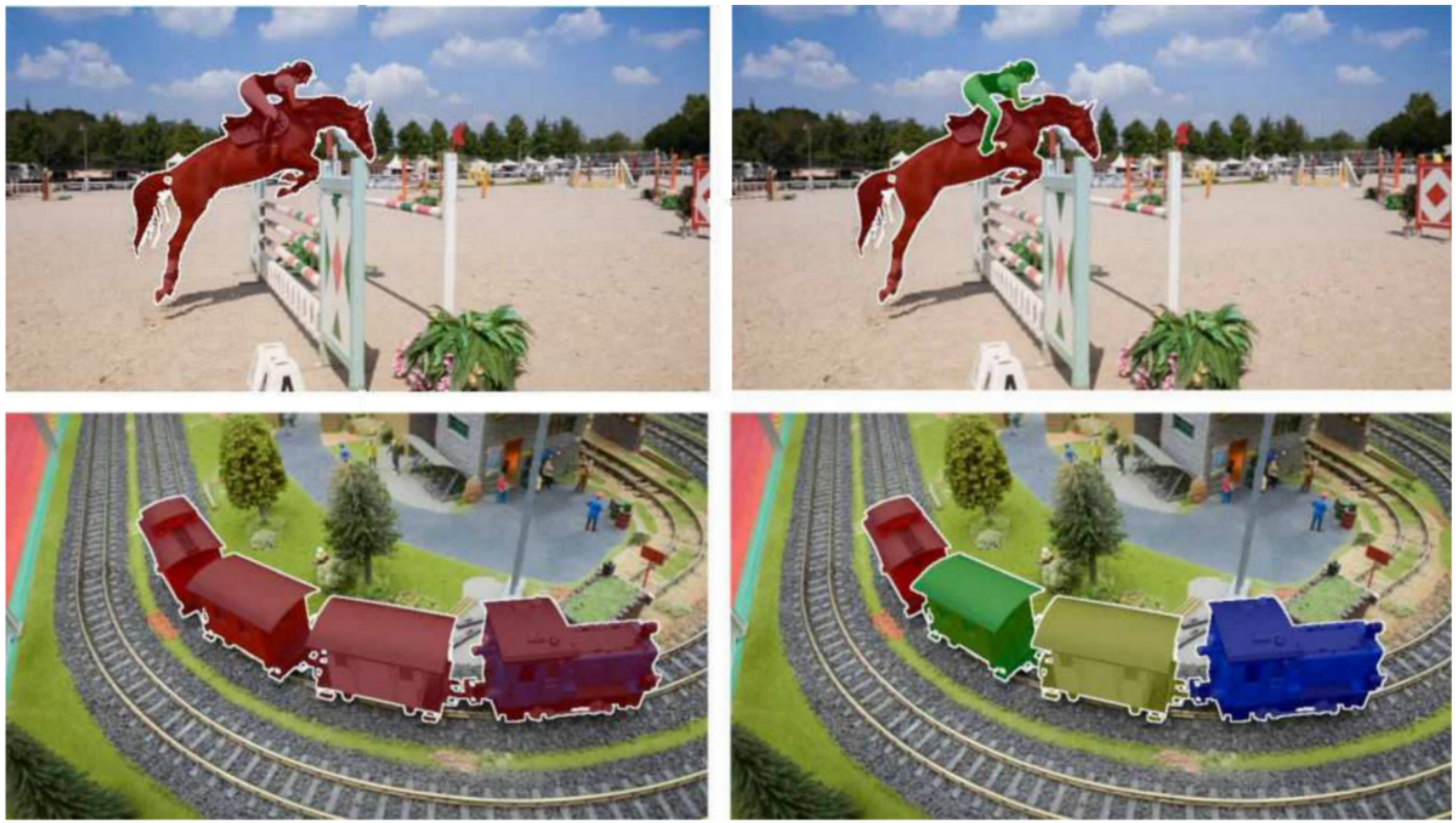

Figure: *Left*: DAVIS 2016 [3] (Single Object), *Right*: DAVIS 2017 [5] (Multiple Objects). Figure extracted from DAVIS Challenge.

Metric

- *Region Similarity* $\mathcal{J} = \frac{|M \bigcap G|}{|M \bigcup G|}$: the IOU between the estimated segmentation M and the corresponding ground-truth mask G.
- *Contour Accuracy* $\mathcal{F} = \frac{2P_c R_c}{P_c + R_c}$: interpret the masks as a set of closed contours and computes the contour-based F-measure which is a function of precision and recall.

Two main approaches to DAVIS-2016

- OSVOS [1]: segment the frames independently, no use of temporal information in the video.
- MaskTrack [4], LucidTracker [2]: take temporal information into account.

Outline

Pipeline

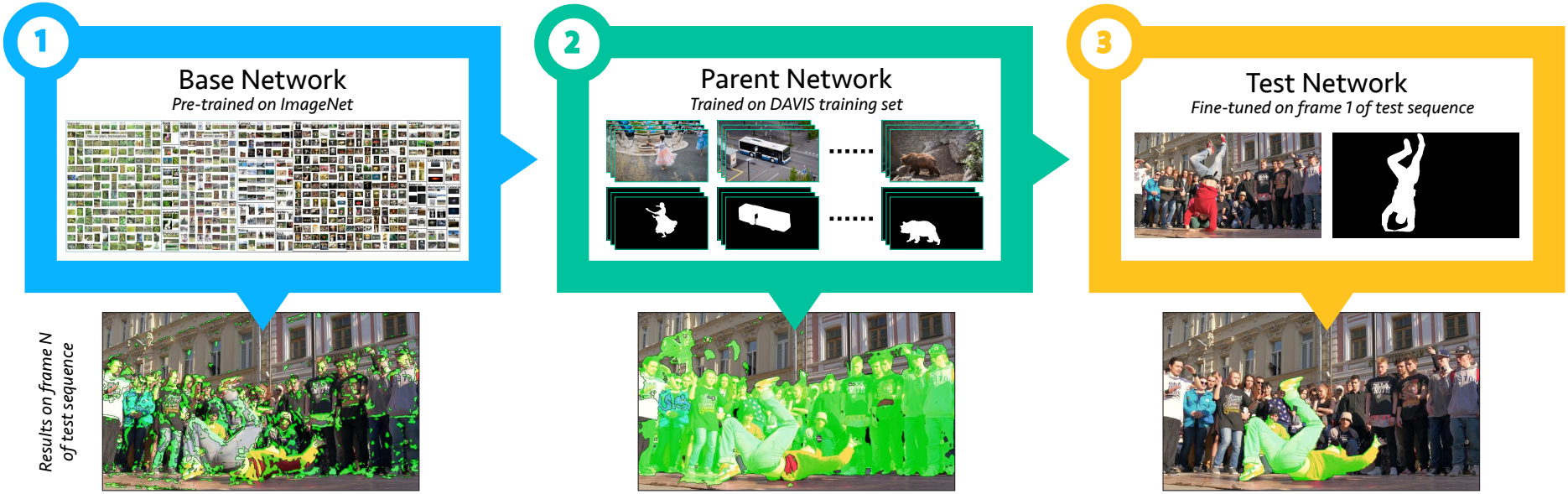

Figure: Overall pipeline. Figure extracted from OSVOS [1].

- VGG-16
- fine-tune for the first frame

Network Architecture

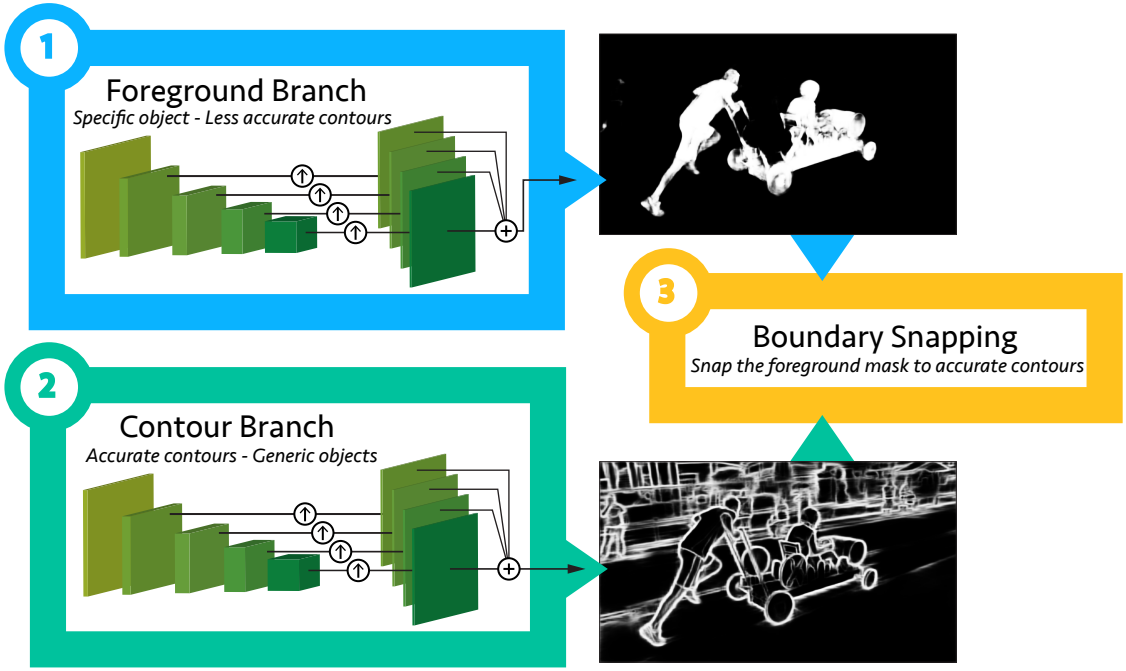

Figure: Two-stream FCN architecture (VGG-16 with balanced cross entropy loss [7]). Contour branch only trained on PASCAL-Context database. Figure extracted from OSVOS [1].

Use Fast Bilateral Solver to snap background predictions to the image edges.

Qualitative evolution of the fine tuning

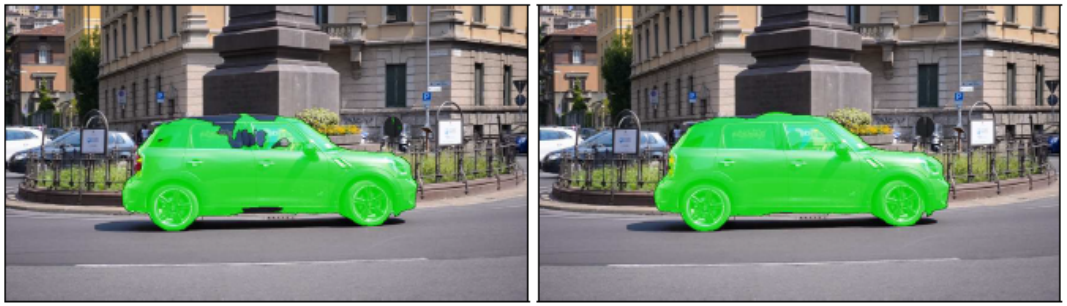

Figure: Results at 10 seconds and 1 minute per sequence. Figure extracted from OSVOS [1].

Outline

Overview

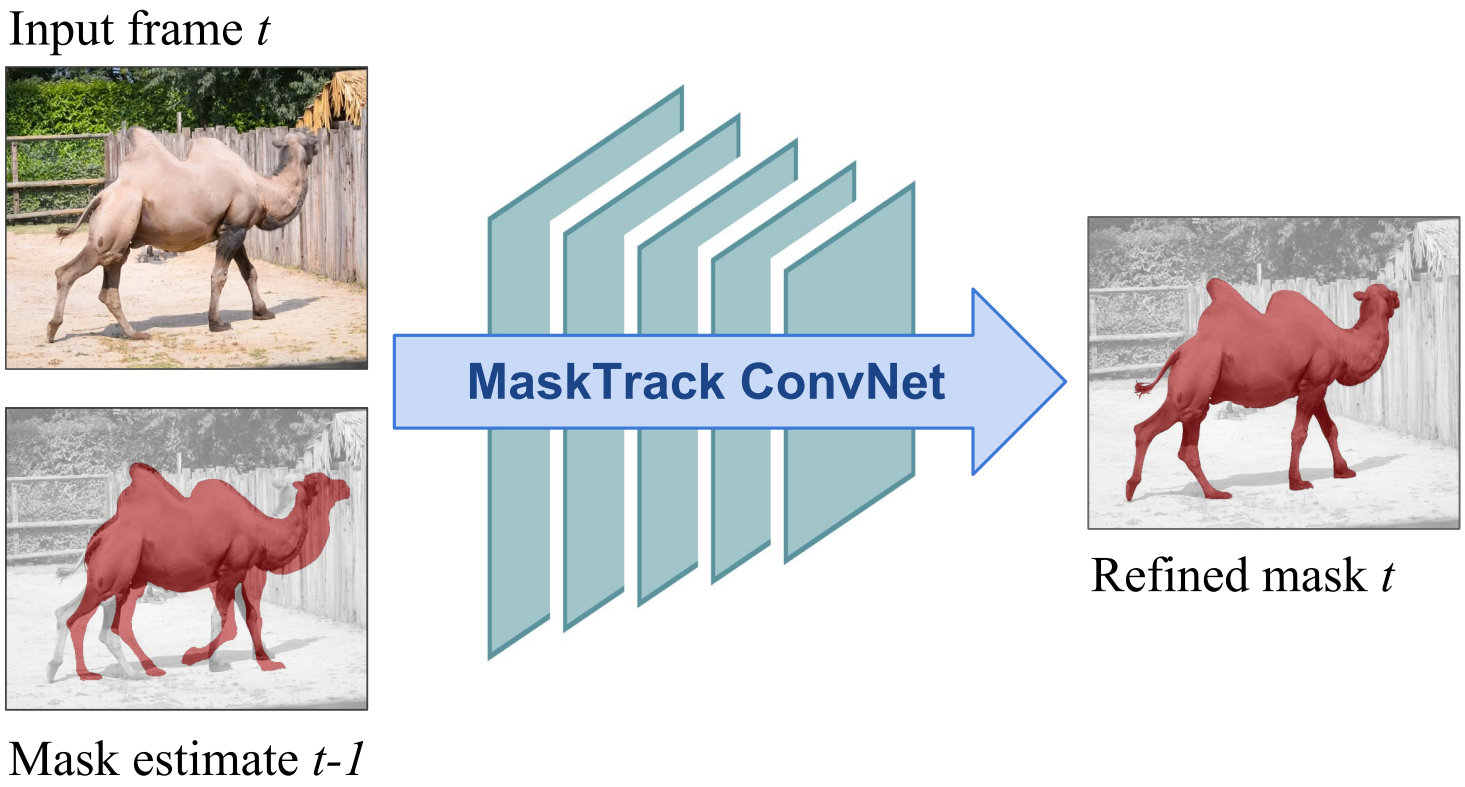

Figure: Network architecture (DeepLabv2-VGG). Expand input from RGB to RGB + mask channels. Figure extracted from MaskTrack [4].

Training Strategy

Training

- Offline training: Input mask deformation by affine transformation and non-rigid deformations, which aim at modelling the expected motion of an object between two frames. (train on saliency segmentation dataset)
- Online training: Use first frame ground-truth annotation to synthesize additional, video specific training data.

Variants

1. Box Annotation: take bounding box annotation instead of mask
2. Optical Flow: an identical second stream network without retraining, fuse by averaging the output scores

Outline

Outline

Lucid Data Dreaming

Lucid data dreaming is a data generating strategy, i.e., synthesizing samples from the provided annotated frame (first frame) in each target video.
The outcome of this process is a large set of frame pairs in the target domain (2.5k pairs per annotation) with known optical flow and mask annotation.

Synthesis Process

Applied the process twice to get a pair of frames with known mask annotation, optical flow, and occlusion regions.

1. Illumination changes
2. Fg / bg split
3. Object Motion: affine and non-rigid deformation
4. Camera Motion: affine transformation
5. Fg / Bg merge

Lucid Data Dreaming Examples I

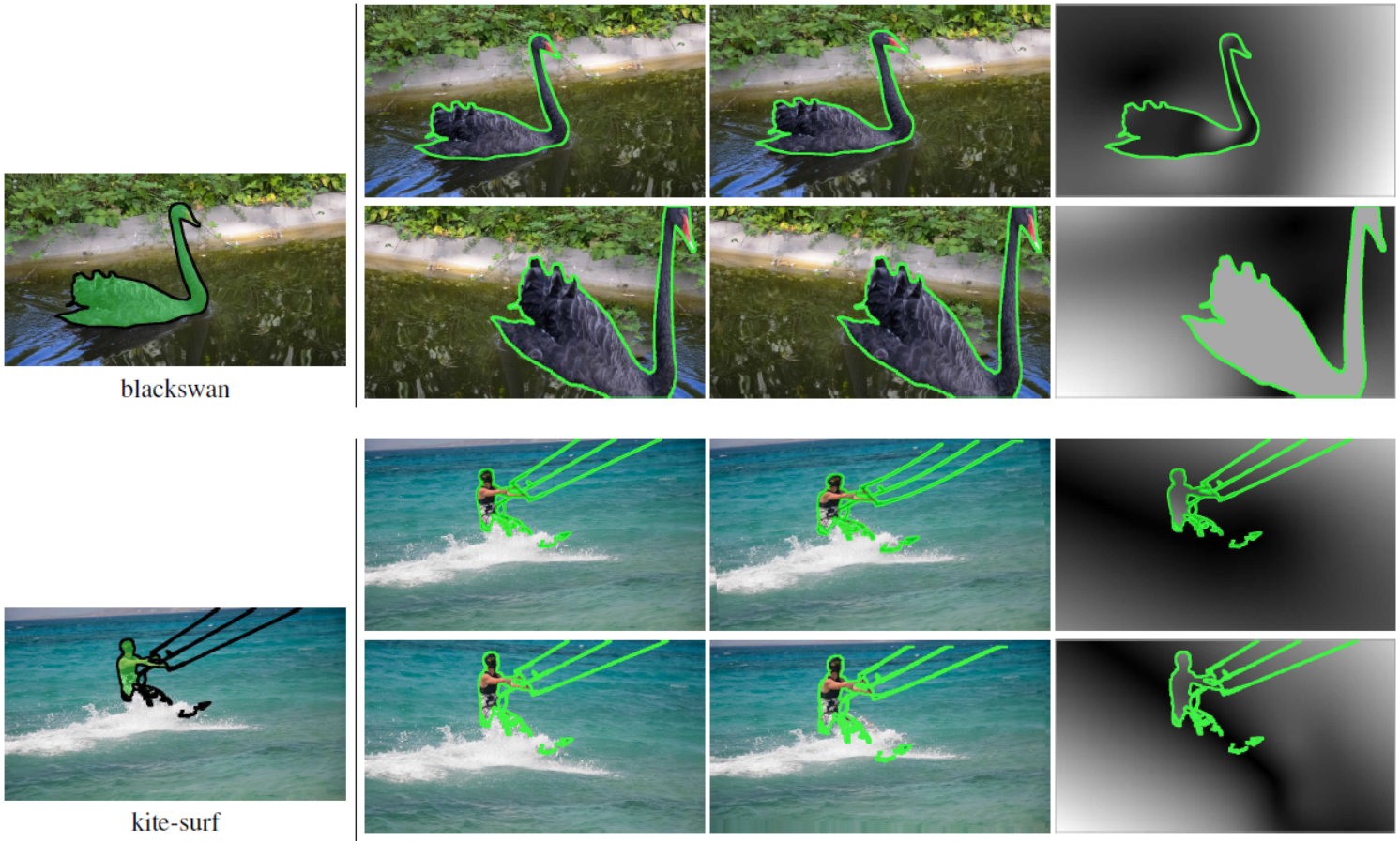

Figure: Figure extracted from LucidTracker [2].

Lucid Data Dreaming Examples II

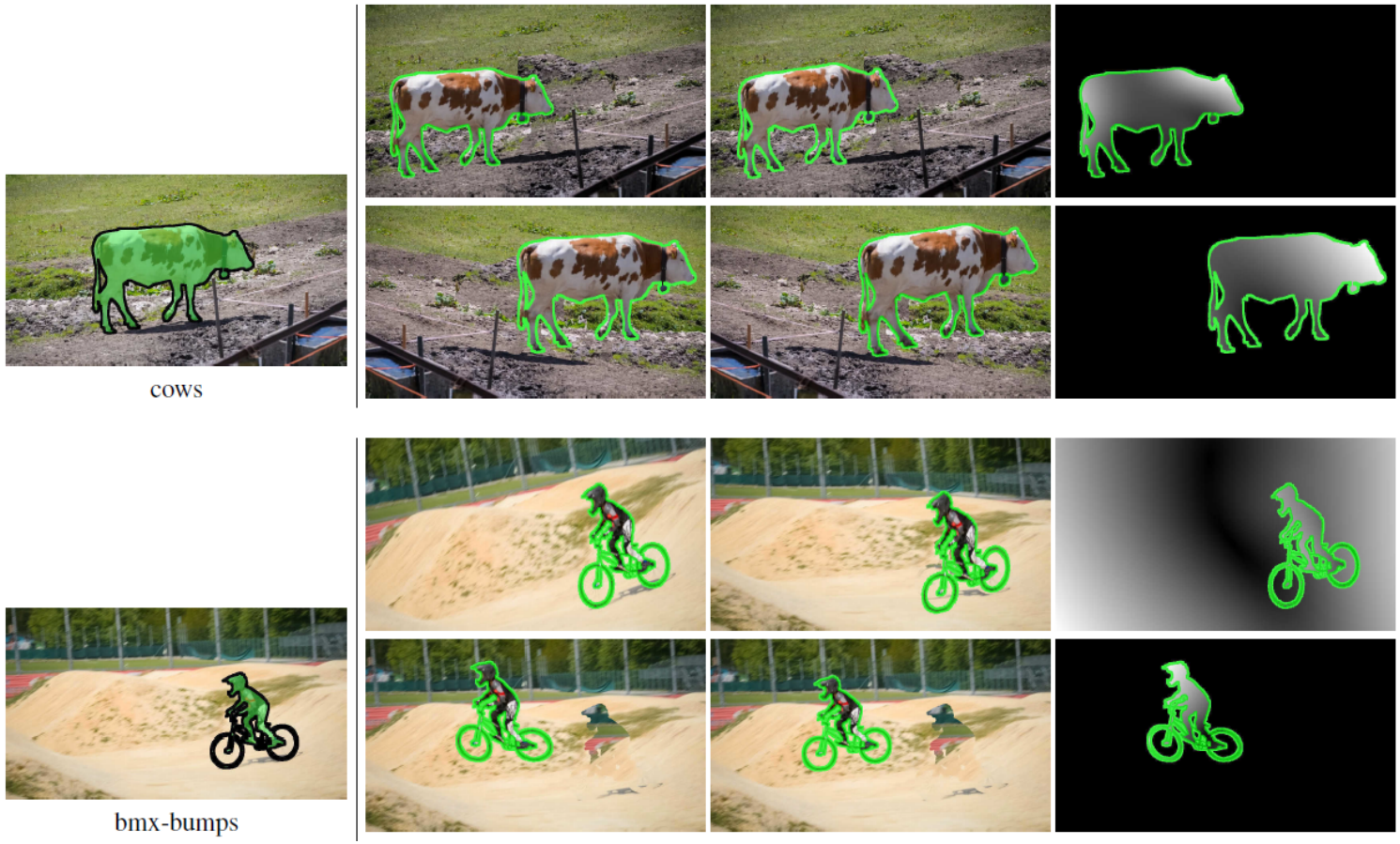

Figure: Figure extracted from LucidTracker [2].

Lucid Data Dreaming Examples with multiple objects

(a) Original image $\mathcal{I}_0$ and mask annotation M_0

(b) Generated image $\mathcal{I}_\tau$ and mask M_τ

(c) Generated flow magnitude $\|\mathcal{F}_\tau\|$

Figure: Figure extracted from LucidTracker [2].

Outline

Single Object Tracking Architecture

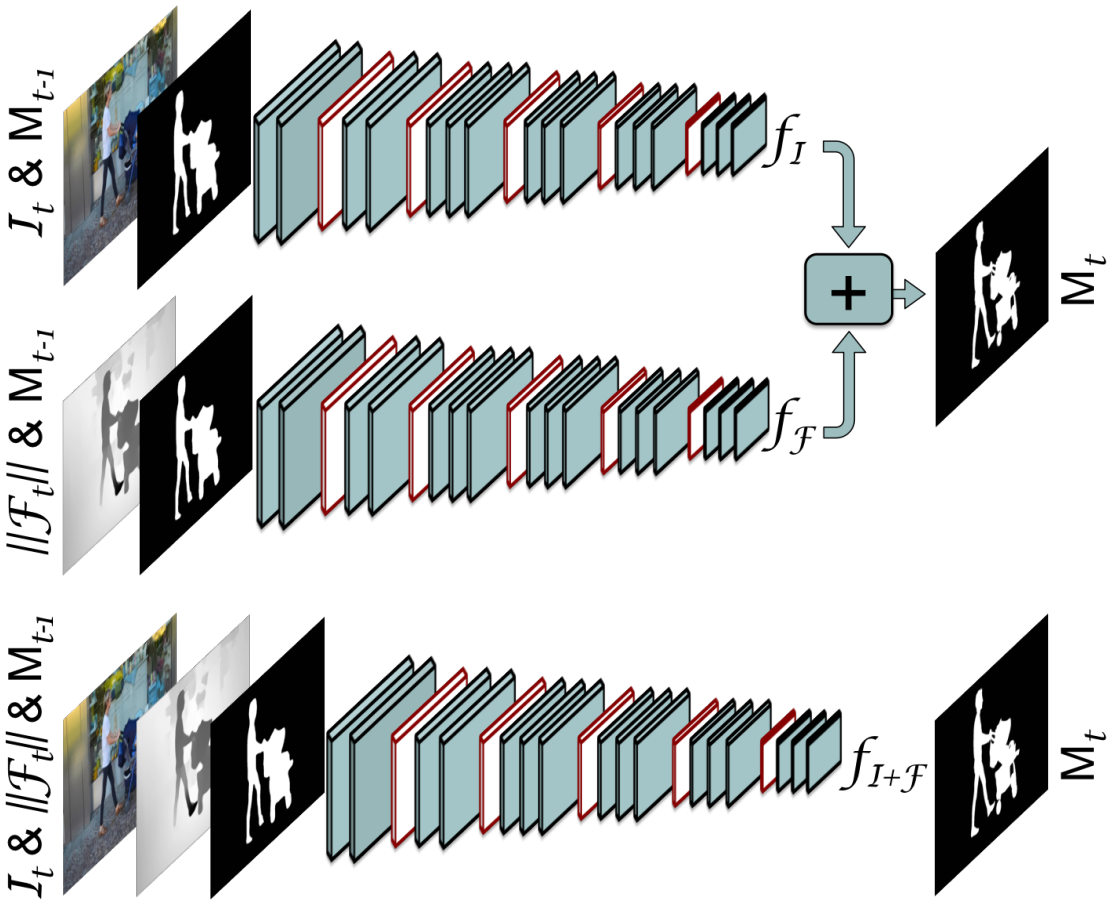

Figure: *Top*: Two stream architecture, *Bottom*: One stream architecture. Figure extracted from LucidTracker [2].

Multiples Objects Tracking Architecture

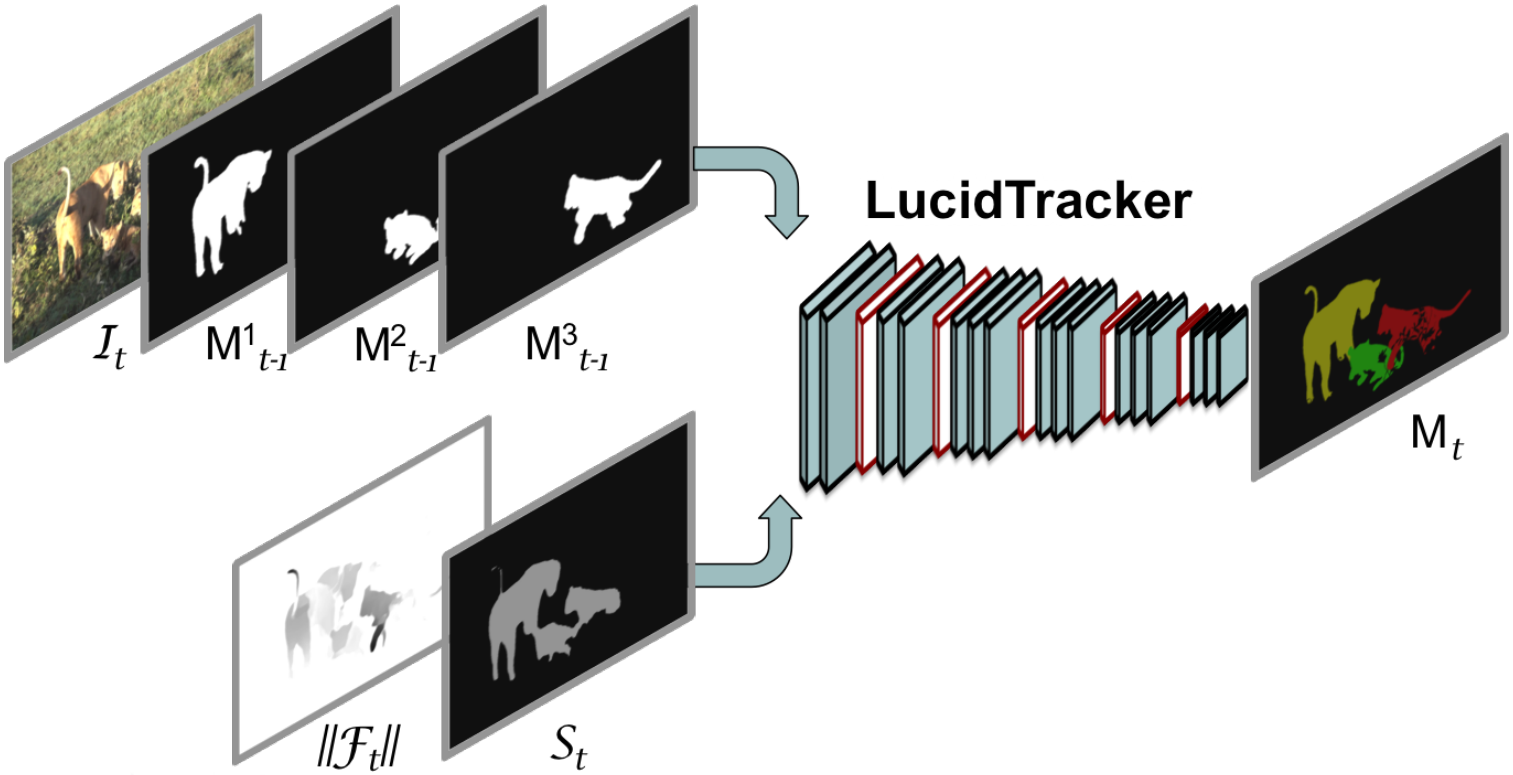

Figure: Image I, N object masks M, semantic segmentation S from PSPNet [8], optical flow F. Figure extracted from LucidTracker [2].

Outline

Comparison

Table: Comparison of segment tracking results

Method	# training images	Flow	mIoU on DAVIS
OSVOS [1]	∼2.3k	No	79.8
MaskTrack [4]	∼11k	Yes	80.3
LucidTracker [2]	24∼126	Yes	91.2

Outline

Pipeline

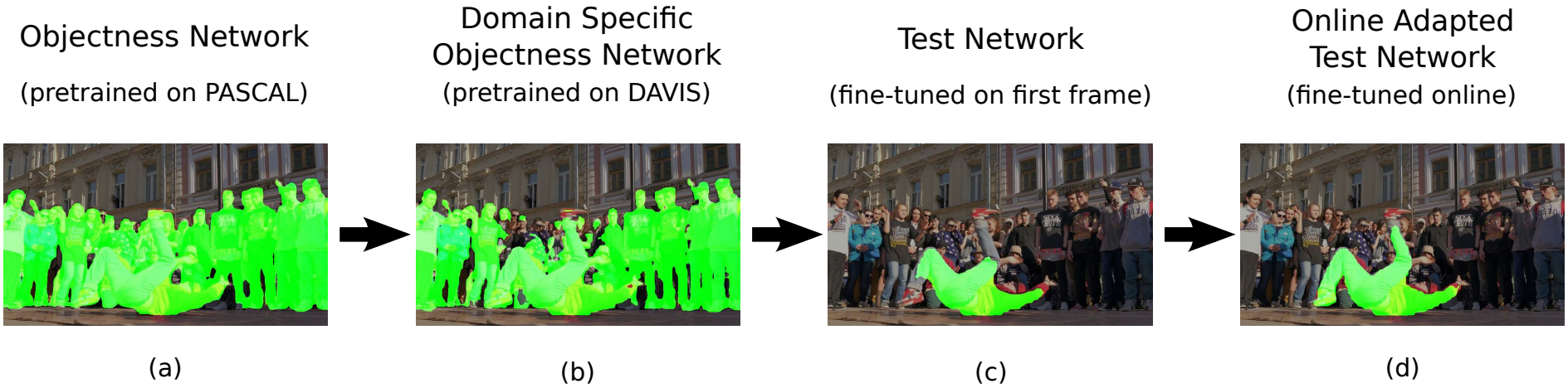

Figure: Pipeline. Figure extracted from OnAVOS [6].

- ResNet: Wider or Deeper.
- Online fine-tune with current frame and first frame

Summary

Video Object Segmentation

Pixel-wise object tracking problem

Approach

- Data Generation
- Offline + online training
- Two-stream: Appearance + Motion information

Reference

Sergi Caelles et al. "One-shot video object segmentation". In: *CVPR*. 2017.

Anna Khoreva et al. "Lucid data dreaming for object tracking". In: *arXiv preprint arXiv:1703.09554* (2017).

F. Perazzi et al. "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation". In: *CVPR*. 2016.

Federico Perazzi et al. "Learning video object segmentation from static images". In: *CVPR*. 2017.

Jordi Pont-Tuset et al. "The 2017 DAVIS Challenge on Video Object Segmentation". In: *arXiv:1704.00675* (2017).

Paul Voigtlaender and Bastian Leibe. "Online Adaptation of Convolutional Neural Networks for Video Object Segmentation". In: *BMVC*. 2017.

Saining Xie and Zhuowen Tu. "Holistically-nested edge detection". In: *CVPR*. 2015.

Hengshuang Zhao et al. "Pyramid Scene Parsing Network". In: *CVPR*. 2017.

Thanks

Thanks for Attention!