Computer Vision 2
WS 2018/19

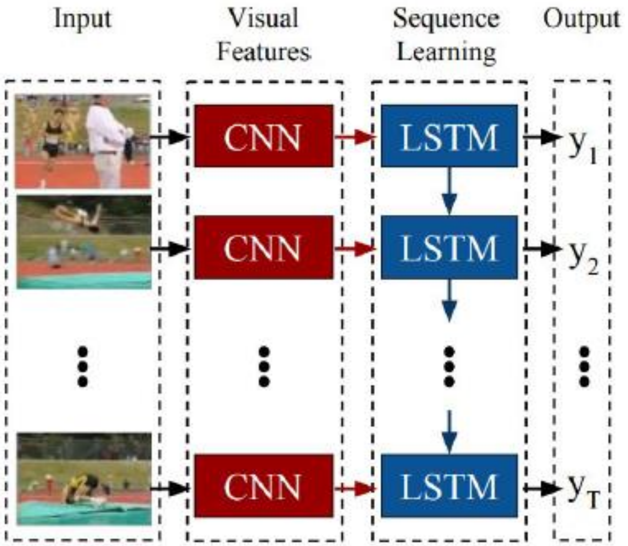
Part 18 – CNNs for Video Analysis III
23.01.2019

Guest Lecture: M.Sc. Jonathon Luiten

RWTH Aachen University, Computer Vision Group
http://www.vision.rwth-aachen.de

Course Outline

- Single-Object Tracking
- Bayesian Filtering
- Multi-Object Tracking
- Visual Odometry
- Visual SLAM & 3D Reconstruction
  - Online SLAM methods
  - Full SLAM methods
- Deep Learning for Video Analysis
  - CNNs for video analysis
  - CNNs for motion estimation
  - Video object segmentation



Topics of This Lecture

- Video Object Segmentation (VOS)
  - First-frame fine-tuning
  - Online Adaptation
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation
  - Further Approaches

- Multi-object Tracking and Segmentation (MOTS)
  - The future of segmentation based tracking

Exciting Progress in Semantic Segmentation: 2017

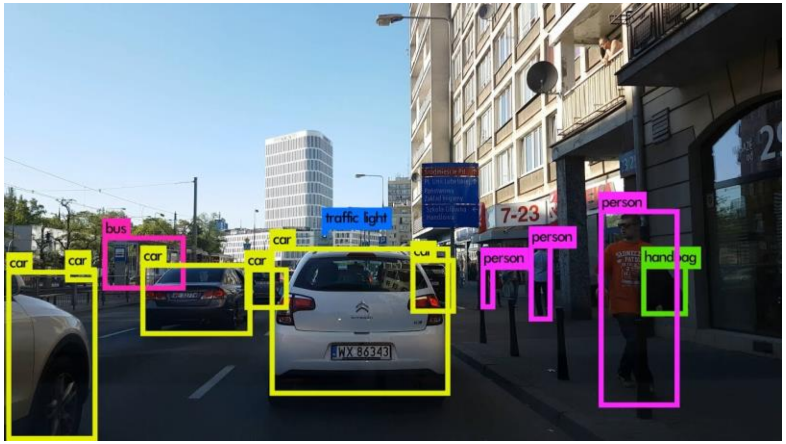
- Full-Resolution Residual Network (FRRN) [CVPR'17]
  - Single-frame processing results

[Pohlen, Hermans, Mathias, Leibe, CVPR'17]

Video Object Segmentation

- Generating **accurate** and **consistent** pixel-masks for objects in a video sequence

Video Object Segmentation



Object Detection

Object Tracking

Object Segmentation

Video Object Segmentation

Video Object Segmentation – Task Formulation

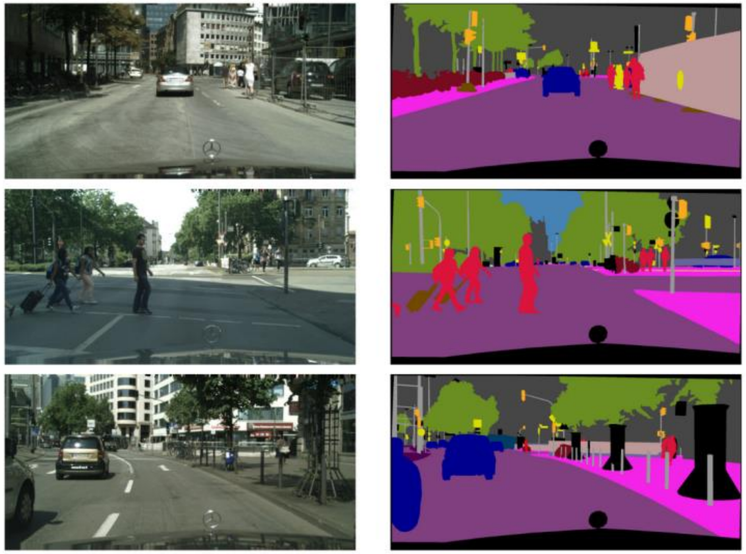
Given: First-frame ground truth

Goal: Complete video segmentation

- Task formulation
  - Given: segmentation mask of target object(s) in the first frame
  - Goal: pixel-accurate segmentation of entire video

  - Currently a major testing ground for segmentation-based tracking

Other fields related to VOS



Semantic Segmentation

Person re-identification

Optical flow estimation

VOS Datasets

DAVIS 2016
(30/20, single
objects, first frames)

DAVIS 2017
(60/90, multiple
objects, first frames)

YouTube-VOS 2018
(3471/982, multiple
objects, first frame where
object appears)

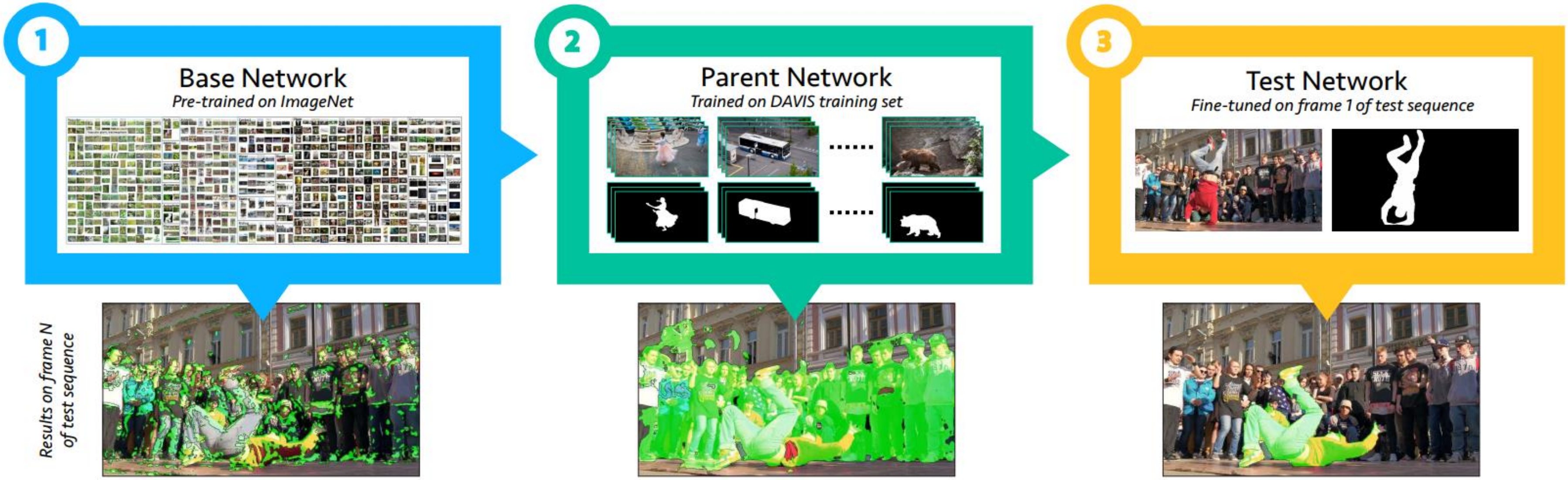
Topics of This Lecture

- Video Object Segmentation (VOS)
  - First-frame fine-tuning
  - Online Adaptation
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation
  - Further Approaches
- Multi-object Tracking and Segmentation (MOTS)
  - The future of segmentation based tracking

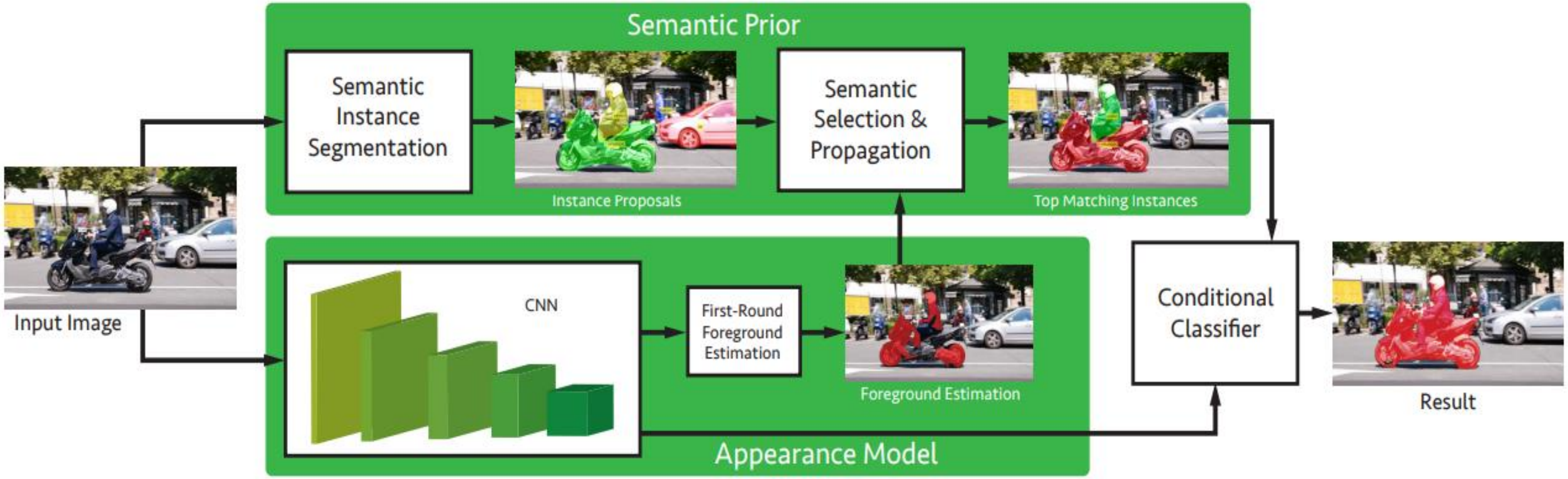
First-frame fine-tuning

- Idea
  - Semantic segmentation of one object (foreground) from background.
  - First-frame adaptation to specific object-of-interest using fine-tuning.
  - Pre-training for ‘objectness’.

OSVOS [Caelles et al. CVPR2017]



OSVOS-S [Maninis et al. PAMI18]



Topics of This Lecture

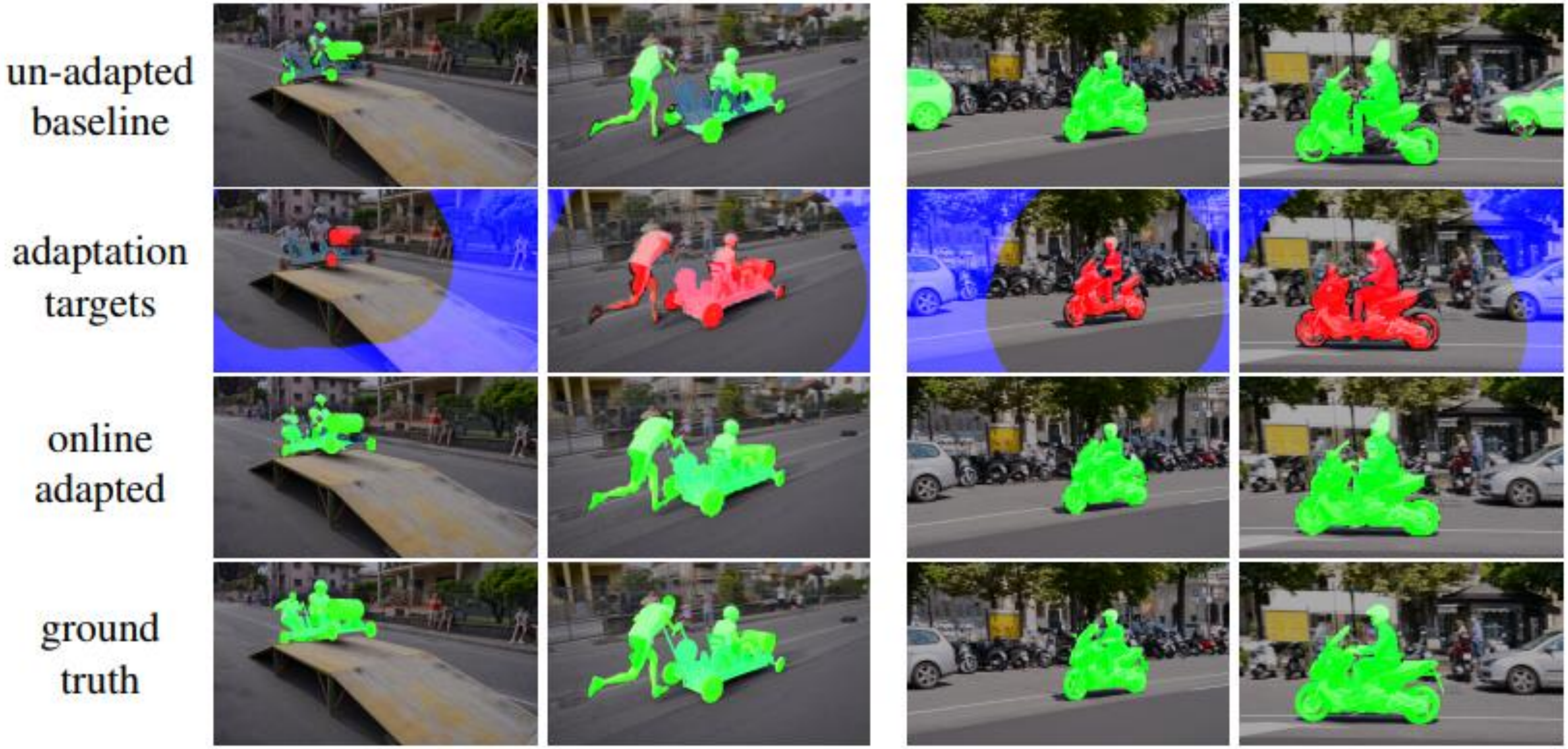
- Video Object Segmentation (VOS)
  - First-frame fine-tuning
  - Online Adaptation
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation
  - Further Approaches
- Multi-object Tracking and Segmentation (MOTS)
  - The future of segmentation based tracking

Online Adaptation

- Idea
  - adapt model to appearance changes every frame – not just in the first frame.
  - Iteratively fine-tune the model on previous prediction every frame.
  - Extremely slow.

  - *You can think of this as a Deep Learning version of Tracking by Online Classification (Lecture 5)…*

OnAVOS [Voigtlaender et al. BMVC17]



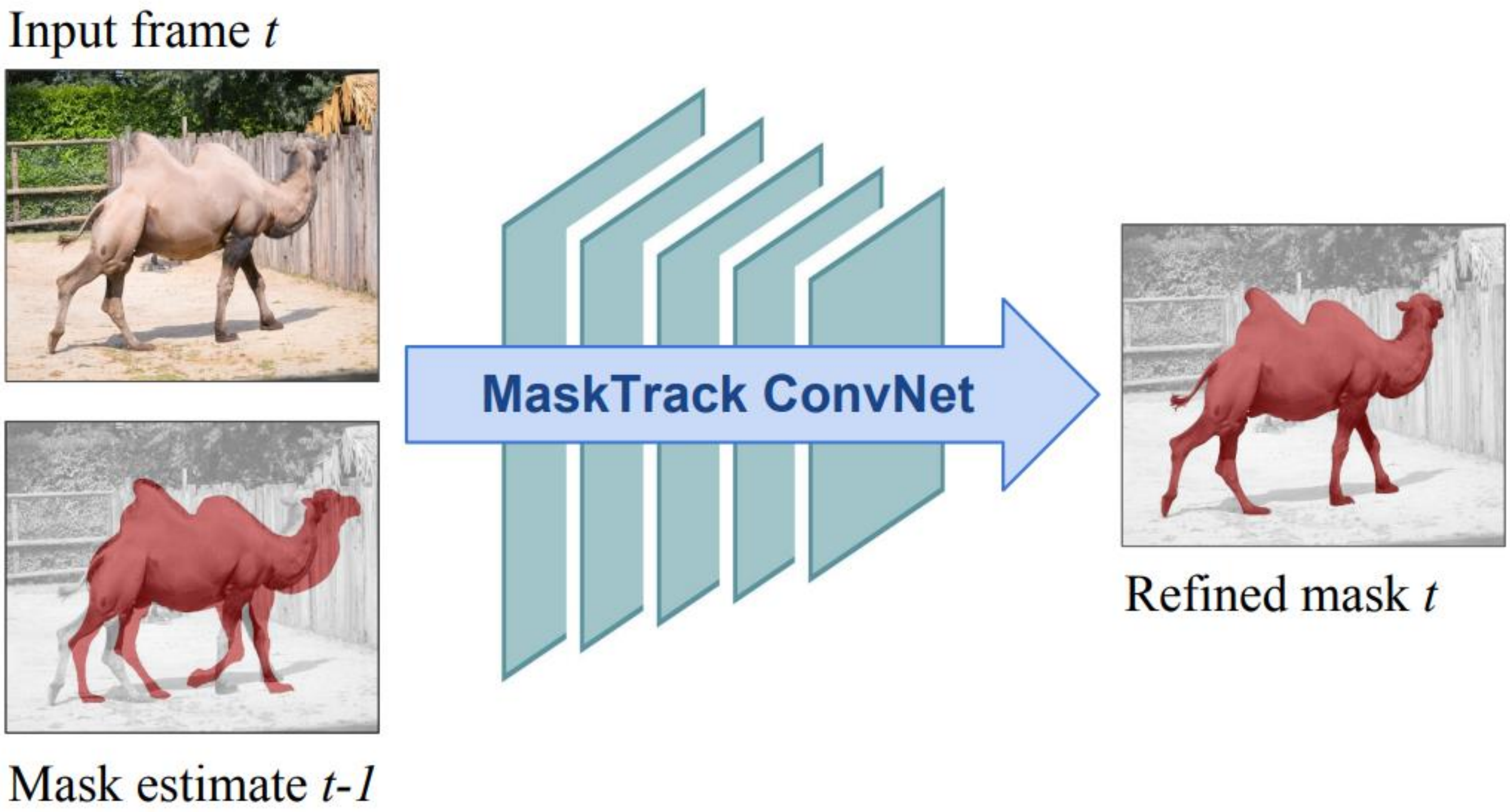
Topics of This Lecture

- Video Object Segmentation (VOS)
  - First-frame fine-tuning
  - Online Adaptation
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation
  - Further Approaches
- Multi-object Tracking and Segmentation (MOTS)
  - The future of segmentation based tracking

Mask Refinement

- Idea
  - We can often start with an approximate mask (either from previous frame or from coarse estimate).
  - Use a refinement network to accurately refine the mask estimate.
  - This can take advantage of crop-and-zoom to do segmentation at a higher resolution.

MaskTrack [Perazzi et al. CVPR17]



Topics of This Lecture

- Video Object Segmentation (VOS)
  - First-frame fine-tuning
  - Online Adaptation
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation
  - Further Approaches
- Multi-object Tracking and Segmentation (MOTS)
  - The future of segmentation based tracking

Optical Flow Mask Propagation

- Idea
  - Optical Flow defines correspondences between the pixels in neighboring frames.
  - Using these correspondences we can determine pixels in one frame that corresponded to a mask in the previous frame.
  - This enables us to ‘warp’ the segmentation mask from one frame to the next.
  - This propagated mask isn’t perfect, and further refinement helps.

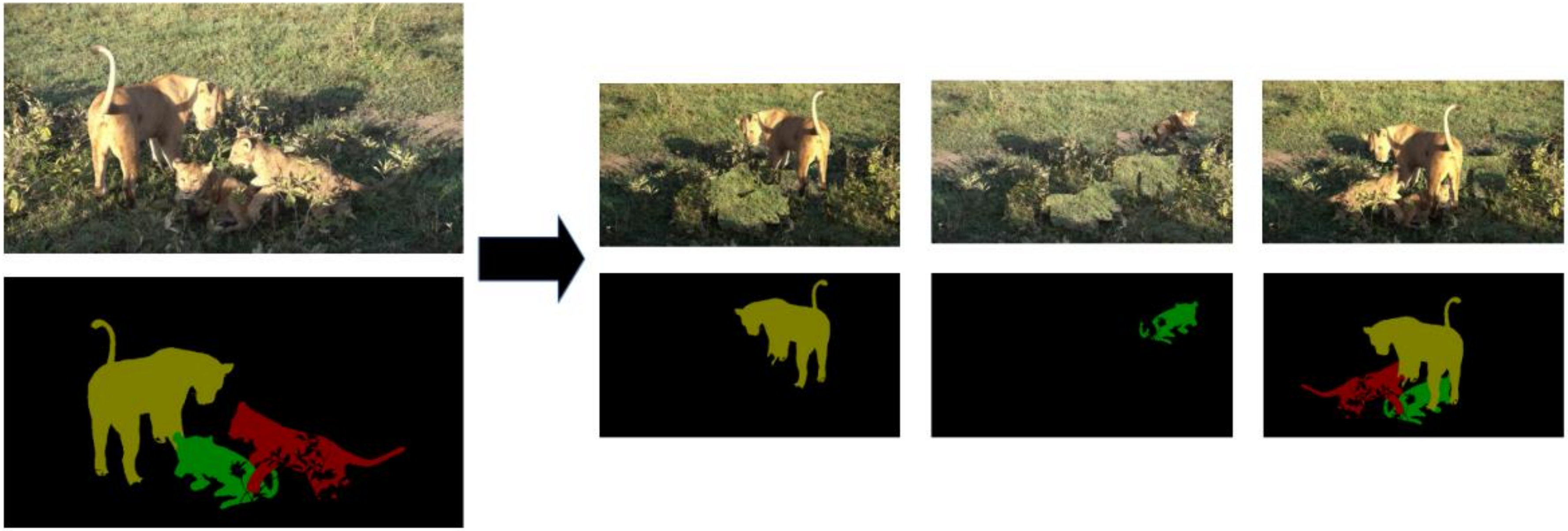
Topics of This Lecture

- Video Object Segmentation (VOS)
  - First-frame fine-tuning
  - Online Adaptation
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation
  - Further Approaches
- Multi-object Tracking and Segmentation (MOTS)
  - The future of segmentation based tracking

Data Augmentation

- Idea
  - Approaches based on fine-tuning networks on the given first frame masks work quite well – but often overfit to first frame appearance.
  - We can get around this by doing large-scale data augmentations.
  - We can crop out the objects-of-interest, fill in the background, and place objects back into the scene randomly with blending.

Lucid Data Dreaming [Khoreva et al. CVPRW17]



Topics of This Lecture

- Video Object Segmentation (VOS)
  - First-frame fine-tuning
  - Online Adaptation
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation
  - Further Approaches
- Multi-object Tracking and Segmentation (MOTS)
  - The future of segmentation based tracking

Object Appearance Re-Identification

- Idea
  – Often objects go in and out of view, or become extremely occluded.
  – In such situations, a mask-propagation based approach fails.
  – We need to re-identify objects based only on their appearance similarity.
  – We can use Siamese or Triplet Loss (see Lecture 18) based ReID networks to determine an appearance similarity score for object proposals.

ReID-VOS [Li et al. CVPRW17]

Topics of This Lecture

- Video Object Segmentation (VOS)
  - First-frame fine-tuning
  - Online Adaptation
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation
  - Further Approaches
- Multi-object Tracking and Segmentation (MOTS)
  - The future of segmentation based tracking

Proposal Generation

- Idea
  - Instance Segmentation Networks (E.g. Mask-RCNN) give excellent single image object instance segmentation proposal results.
  - One can approach video object segmentation as taking these proposals in each frame and then linking them over time using a merging algorithm.

PReMVOS [Luiten et al. ACCV18]

- An approach that combines all of the previous VOS principles and gives state-of-the-art results.
- Combines the following principles:
  - First-frame fine-tuning
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation

PReMVOS – Overview

Proposal generation

Refinement

Merging

- Proposal generation
  - Category-agnostic Mask R-CNN proposals
  - ResNet101 backbone, joint training on COCO and Mapillary
- Refinement
  - Fully-convolutional segmentation network trained to refine the segmentation given a proposal bounding box
  - DeepLabV3+ backbone

[Luiten, Voigtlaender, Leibe, ACCV'18]

PReMVOS – Overview

Proposal generation | Refinement | Merging

- Merging
  - Greedy decision process, chooses proposal(s) with best score
  - Optional proposal expansion through Optical Flow propagation
  - Proposal score as combination of
    - Objectness score
    - Mask propagation IoU score (Optical Flow warping)
    - ReID score
    - Object-Object interaction scores

[Luiten, Voigtlaender, Leibe, ACCV'18]

PReMVOS – Results on Benchmarks

- DAVIS Challenge 2018 Winner

			Ours (Ens)	Ours	Lixx	Dawns	ILC_R	Apata	UIT
17/18 T-C	$\mathcal{J}\&\mathcal{F}$	Mean	**74.7**	71.8	73.8	69.7	69.5	67.8	66.3
	$\mathcal{J}$	Mean	71.0	67.9	**71.9**	66.9	67.5	65.1	64.1
		Recall	**79.5**	75.9	79.4	74.1	77.0	72.5	75.0
		Decay	19.0	23.2	19.8	23.1	15.0	27.7	**11.7**
	$\mathcal{F}$	Mean	**78.4**	75.6	75.8	72.5	71.5	70.6	68.6
		Recall	**86.7**	82.9	83.0	80.3	82.2	79.8	80.7
		Decay	20.8	24.7	20.3	25.9	18.5	30.2	**13.5**

- Youtube-VOS Challenge 2018 Winner

	Overall	$\mathcal{J}$ seen	$\mathcal{J}$ unseen	$\mathcal{F}$ seen	$\mathcal{F}$ unseen
Ours	**72.2**	**73.7**	64.8	**77.8**	72.5
Seq-2-Seq []	70.0	66.9	**66.8**	74.1	72.3
2nd	72.0	72.5	66.3	75.2	**74.1**
3rd	69.9	73.6	62.1	75.5	68.4
4th	68.4	70.6	62.3	72.8	67.7

[Luiten, Voigtlaender, Leibe, ACCV'18]

PReMVOS – Visual Results

[Luiten, Voigtlaender, Leibe, ACCV'18]

Lessons Learned

- Challenge 1: How to generate proposals?
  – Deep-learning based region proposal generators are fit for the task
  – Experimented with SharpMask and Mask R-CNN

- Challenge 2: How to track region proposals?
  – Region overlap works as a consistency measure
  – Optical flow based propagation really helps
  – ReID score also helpful

- Open issues
  – PReMVOS has no notion of 3D objects moving through 3D space.
  – Track initialization / termination logic needed for real tracking.
  – How to obtain the initial segmentation?

Topics of This Lecture

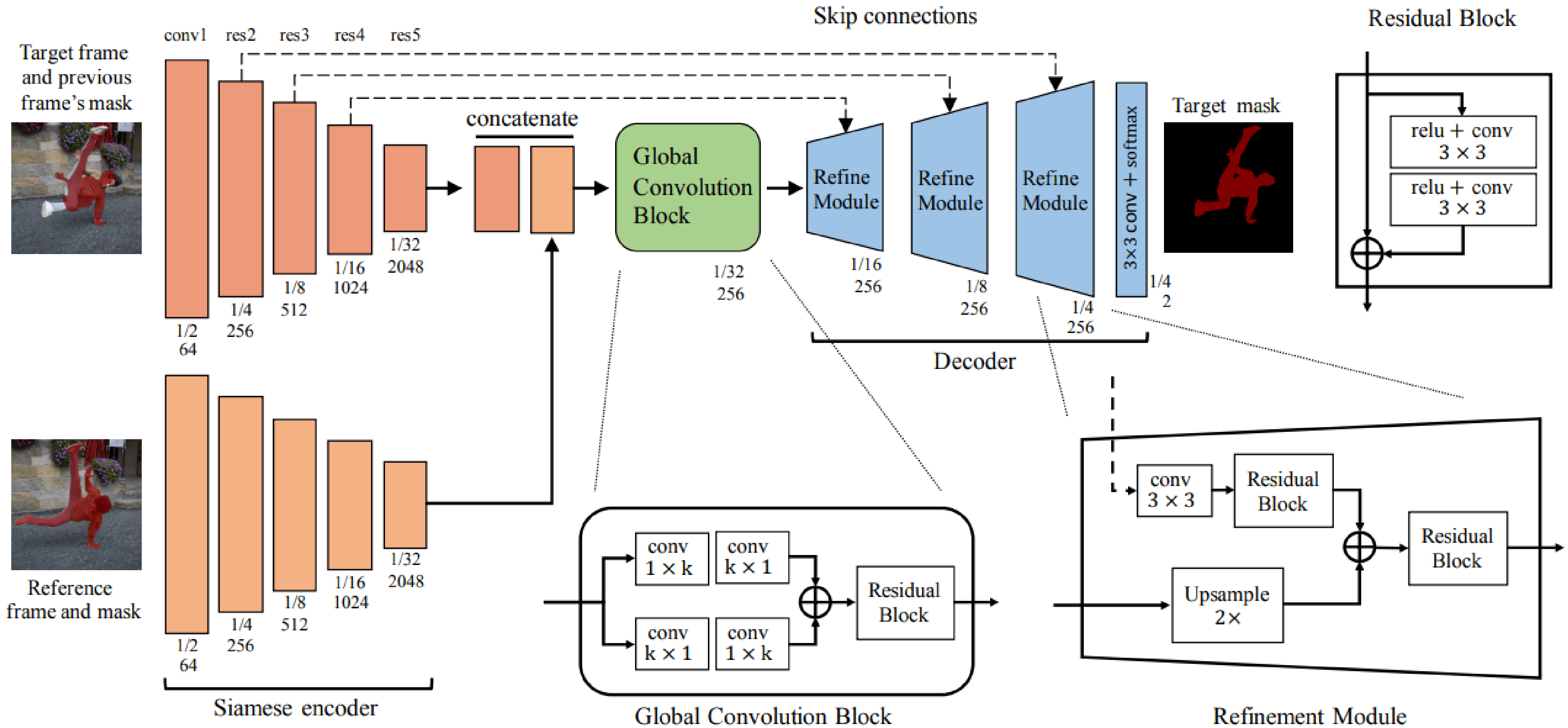
- Video Object Segmentation (VOS)
  - First-frame fine-tuning
  - Online Adaptation
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation
  - Further Approaches
- Multi-object Tracking and Segmentation (MOTS)
  - The future of segmentation based tracking

Combining Mask Propagation and Re-ID

- Idea
  - Mask propagation networks give segmentation dependent on previous frame prediction.
  - Re-ID networks try to match the appearance of the 1^{st} frame to the current frame.
  - We can combine both together by having input from the previous frame and the first frame and concatenating these together before decoding the output.

RGMP [Oh et al. CVPR2018]

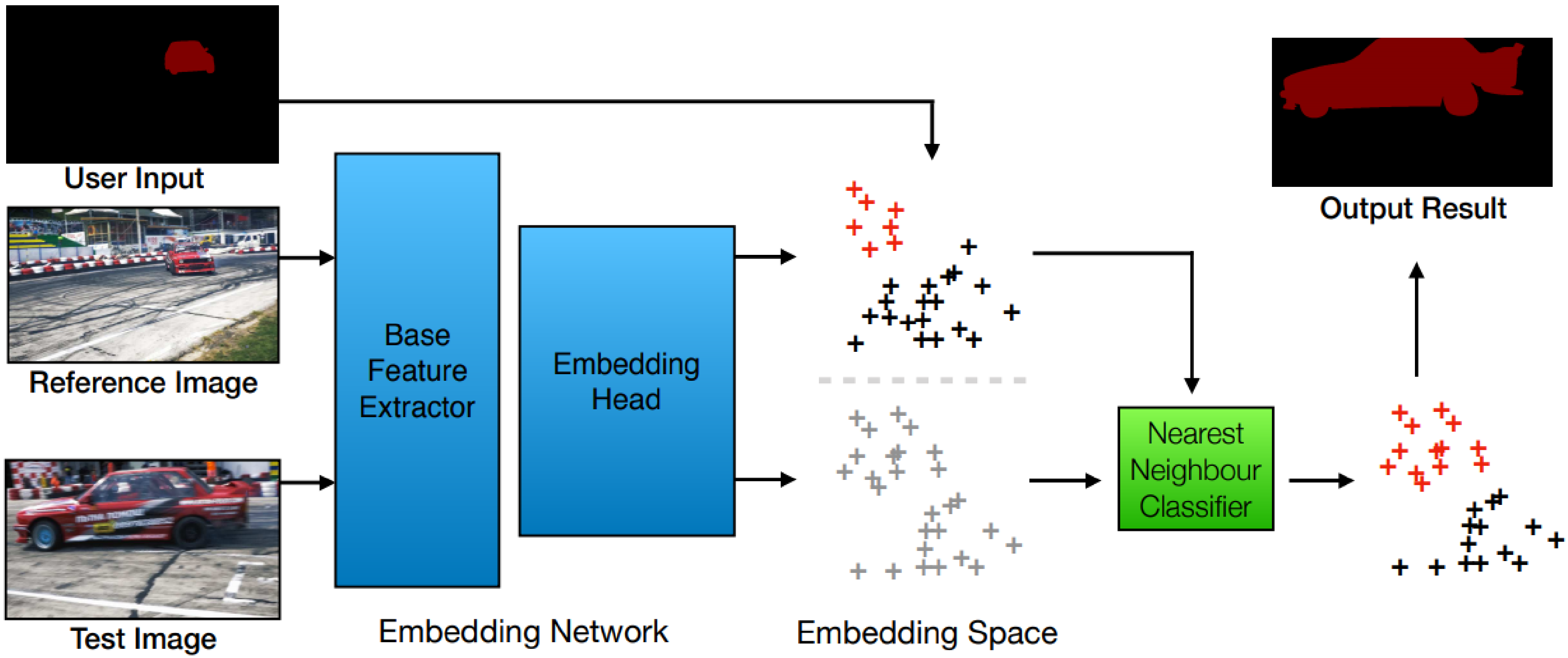
- Region Guided Mask Propagation



Instance Embedding Vectors

- Idea
  - Re-Identification networks based on bounding-box region proposals work really well.
  - This idea can be extended to a Re-Identification embedding for every pixel.
  - This pixel-wise Re-ID embedding vectors can then be used to directly extract a mask by taking the pixel with an embedding similar to the first frame embeddings.

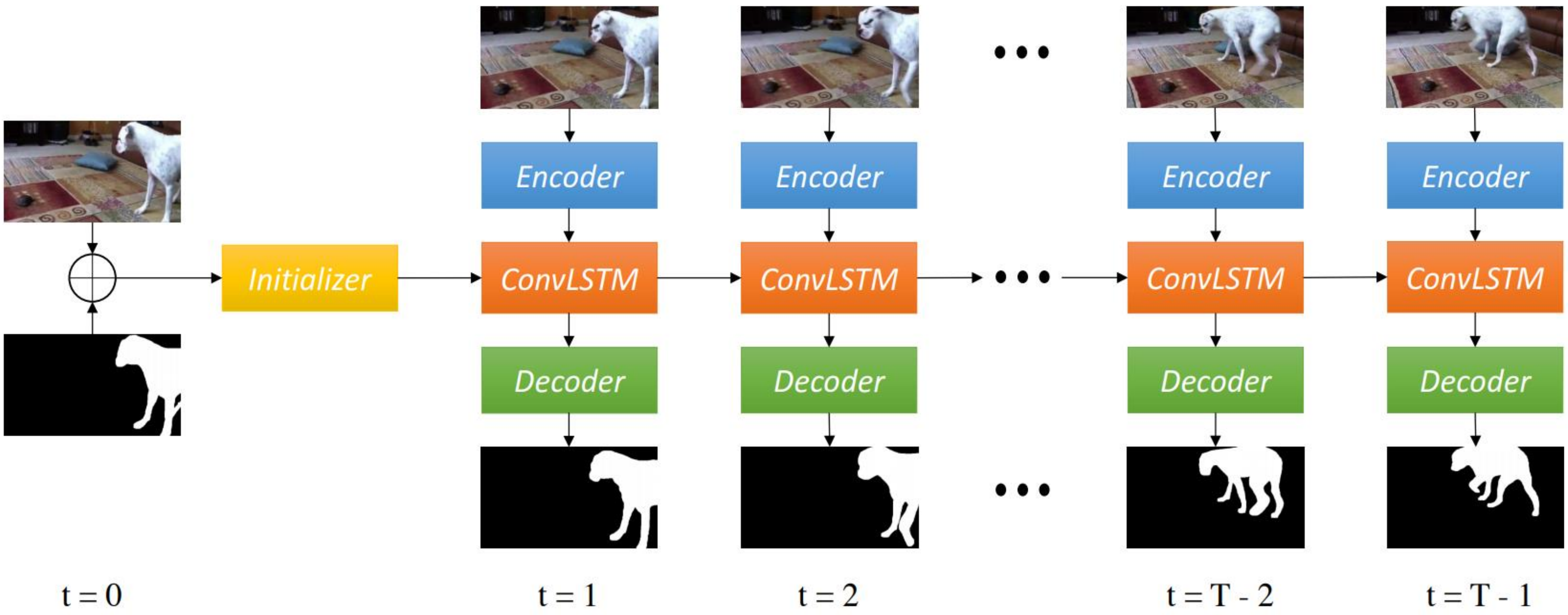
Blazingly Fast [Chen et al. CVPR18]



Using Recurrent Neural Networks

- Idea
  - Most of the approaches use neural networks trained to output results based on either only the current frame, or maybe the previous and/or first frames.
  - Using recurrent neural networks we can directly train our neural networks to learn to produce the results based on the entire sequence of images in a video in an end-to-end manner.

Seq2Seq [Xu et al. ECCV18]



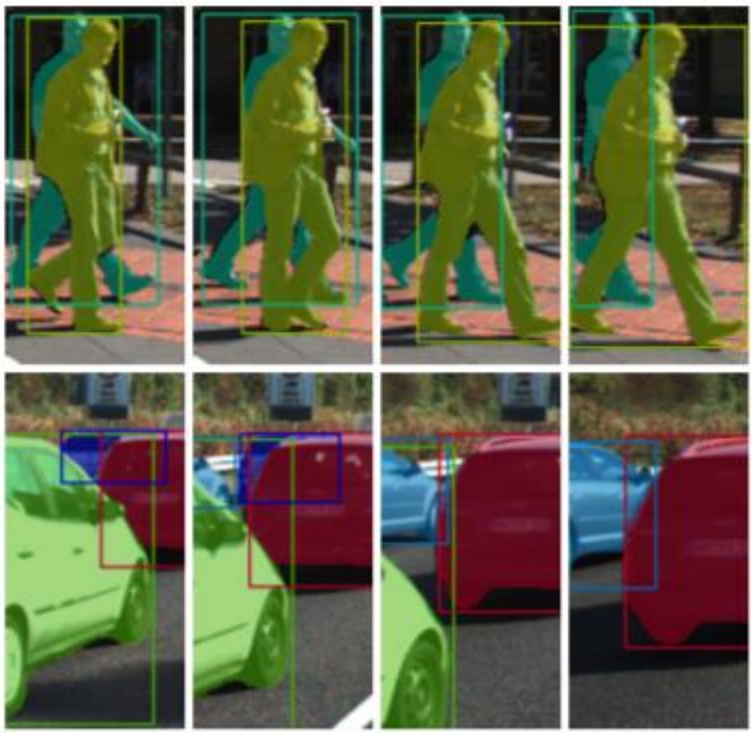
Topics of This Lecture

- Video Object Segmentation (VOS)
  - First-frame fine-tuning
  - Online Adaptation
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation
  - Further Approaches
- Multi-object Tracking and Segmentation (MOTS)
  - The future of segmentation based tracking

VOS -> MOTS

- Video Object Segmentation (VOS) is restricted in a number of ways.
  - First frame mask given
  - Short video clips with objects present in almost all frames
  - Few objects to track (max around 7 per video)
- Multi-Object Tracking and Segmentation (MOTS) is an extension of VOS that deals with all of these short comings.

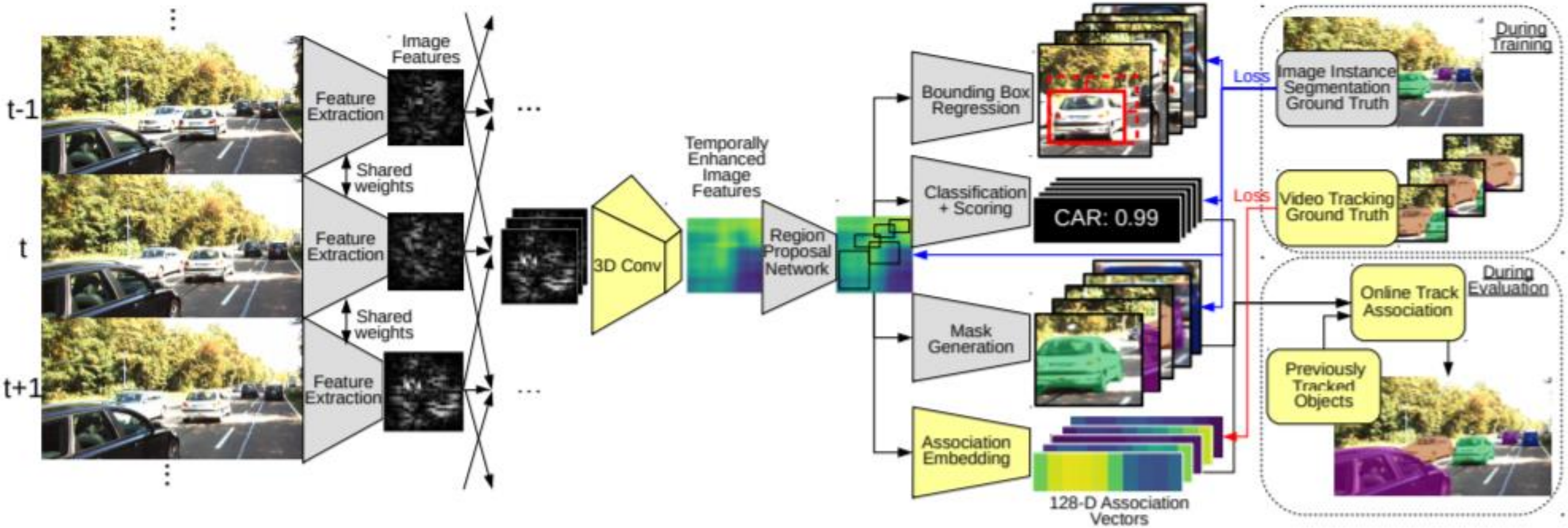
MOTS dataset



Solving MOTS

- Idea
  - Very similar approach to PReMVOS.
  - Proposal-generation followed by merging using optical flow and Re-ID vector.
  - 3D Convolutions for temporally consistent object proposals.
  - Re-ID vector built into the proposal network.
  - New tracks started by confident proposals that don't match well to previous tracks.

MOTS [Voigtlaender et al. sub.]



Topics of This Lecture

- Video Object Segmentation (VOS)
  - First-frame fine-tuning
  - Online Adaptation
  - Mask Refinement
  - Optical Flow Mask Propagation
  - Data Augmentation
  - Object Appearance Re-Identification
  - Proposal Generation
  - Further Approaches
- Multi-object Tracking and Segmentation (MOTS)
  - The future of segmentation based tracking

References and Further Reading

- Caelles, Sergi, et al. "One-shot video object segmentation." CVPR 2017. IEEE, 2017.
- Maninis, Kevis-Kokitsi, et al. "Video Object Segmentation Without Temporal Information." arXiv preprint arXiv:1709.06031 (2017).
- Voigtlaender, Paul, and Bastian Leibe. "Online adaptation of convolutional neural networks for video object segmentation." arXiv preprint arXiv:1706.09364 (2017).
- Perazzi, Federico, et al. "Learning video object segmentation from static images." Computer Vision and Pattern Recognition. 2017.
- Li, Xiaoxiao, et al. "Video object segmentation with re-identification." arXiv preprint arXiv:1708.00197 (2017).
- Khoreva, Anna, et al. "Lucid Data Dreaming for Multiple Object Tracking." arXiv preprint arXiv:1703.09554 (2017).
- Li, Xiaoxiao, and Chen Change Loy. "Video Object Segmentation with Joint Re-identification and Attention-Aware Mask Propagation." arXiv preprint arXiv:1803.04242 (2018).

References and Further Reading

- Oh, Seoung Wug et al. “Fast Video Object Segmentation by Reference-Guided Mask Propagation”. CVPR 2018.
- Chen, Yuhua et al. “Blazingly Fast Video Object Segmentation with Pixel-Wise Metric Learning”. CVPR 2018.
- Xu, Ning et al. “YouTube-VOS: Sequence-to-Sequence Video Object Segmentation”. ECCV 2018.
- Luiten, Jonathon et al. “PReMVOS: Proposal Generation, Refinement and Merging for Video Object Segmentation”. ACCV 2018.