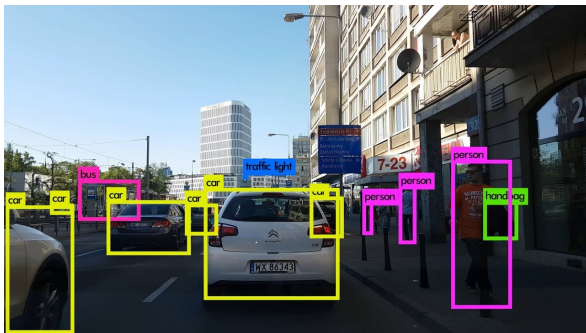


Video Object Segmentation

Video Object Segmentation



Object Detection



Object Segmentation



Object Tracking



Video Object Segmentation

This lecture

Video Object Segmentation

- Goal: Generate accurate and temporally consistent pixel masks for objects in a video sequence.



VOS: some challenges

- Strong viewpoint/appearance changes



VOS: some challenges

- Strong viewpoint/appearance changes
- Occlusions



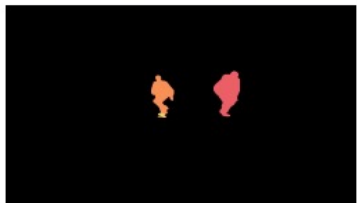
VOS: some challenges

- Strong viewpoint/appearance changes
- Occlusions
- Scale changes



VOS: tasks

Semi-supervised
(one-shot) video
object
segmentation



We get the first frame
ground truth mask, we know
what object to segment

Unsupervised
(zero-shot) video
object
segmentation

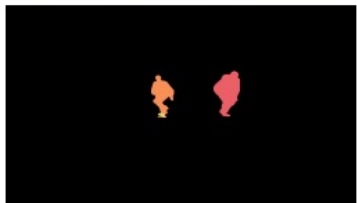


We have to find the
objects as well as their
masks

VOS: tasks

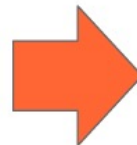
Motion segmentation,
salient object detection..

Semi-supervised
(one-shot) video
object
segmentation



We get the first frame
ground truth mask, we know
what object to segment

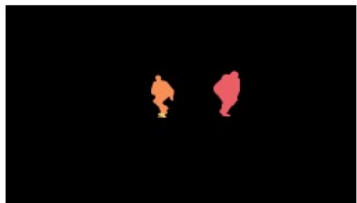
Unsupervised
(zero-shot) video
object
segmentation



We have to find the
objects as well as their
masks

This lecture

Semi-supervised
(one-shot) video
object
segmentation



We get the first frame
ground truth mask, we know
what object to segment

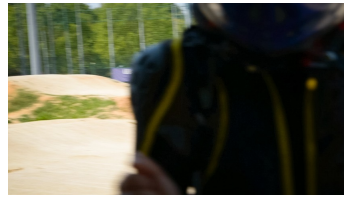
VOS: tasks

Unsupervised
(zero-shot) video
object
segmentation



We have to find the
objects as well as their
masks

Supervised Video Object Segmentation



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 the entire video
 - Currently a major testing ground for segmentation-based tracking

VOS Datasets

- Remember that large-scale datasets are needed for learning-based methods



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)

<https://davischallenge.org>
<https://youtube-vos.org>

Before we get started...

- Pixel-wise output
- If we talk about pixel-wise outputs and motion, there is a concept in Computer Vision that we need to know first

Optical flow

Optical flow

- Input: 2 consecutive images (e.g. from a video)
- Output: displacement of every pixel from image A to image B
- Results in the “perceived” 2D motion, not the real motion of the object

Optical flow

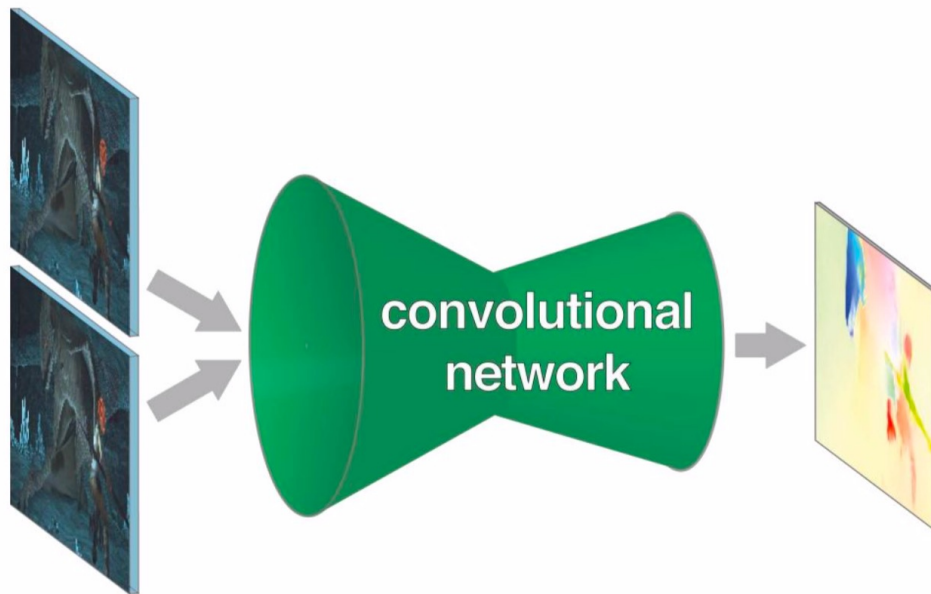


Optical flow



Optical flow with CNNs

- End-to-end supervised learning of optical flow



P. Fischer et al. „FlowNet: Learning Optical Flow With Convolutional Networks“. ICCV 2015

Optical flow with CNNs

FlowNet: Learning Optical Flow with Convolutional Networks

FlowNet
P. Fischer,
A. Dosovitskiy,
E. Ilg,
P. Häusser,
C. Hazırbaş,
V. Golkov,
P. v.d. Smagt,
D. Cremers,
T. Brox

Img1

Img2

FlowNetS

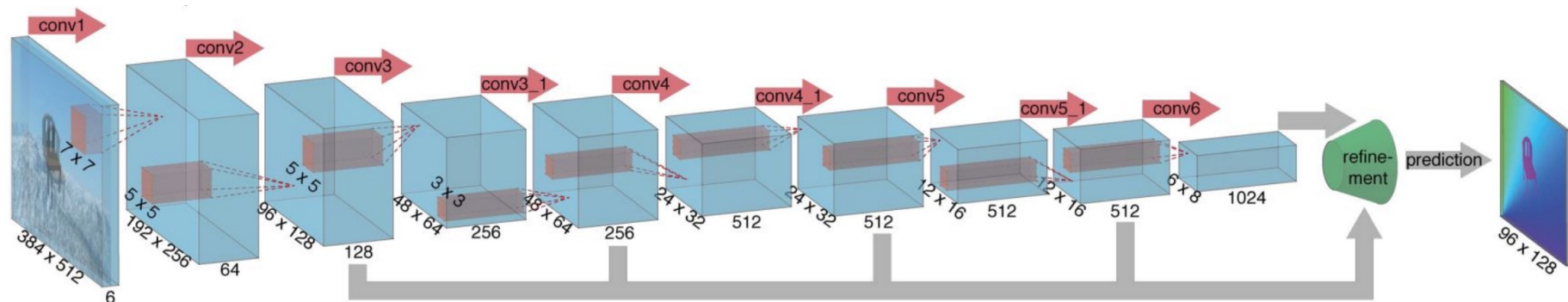
FlowNetC

We train convolutional networks to estimate optical flow.

P. Fischer et al. „FlowNet: Learning Optical Flow With Convolutional Networks“. ICCV 2015

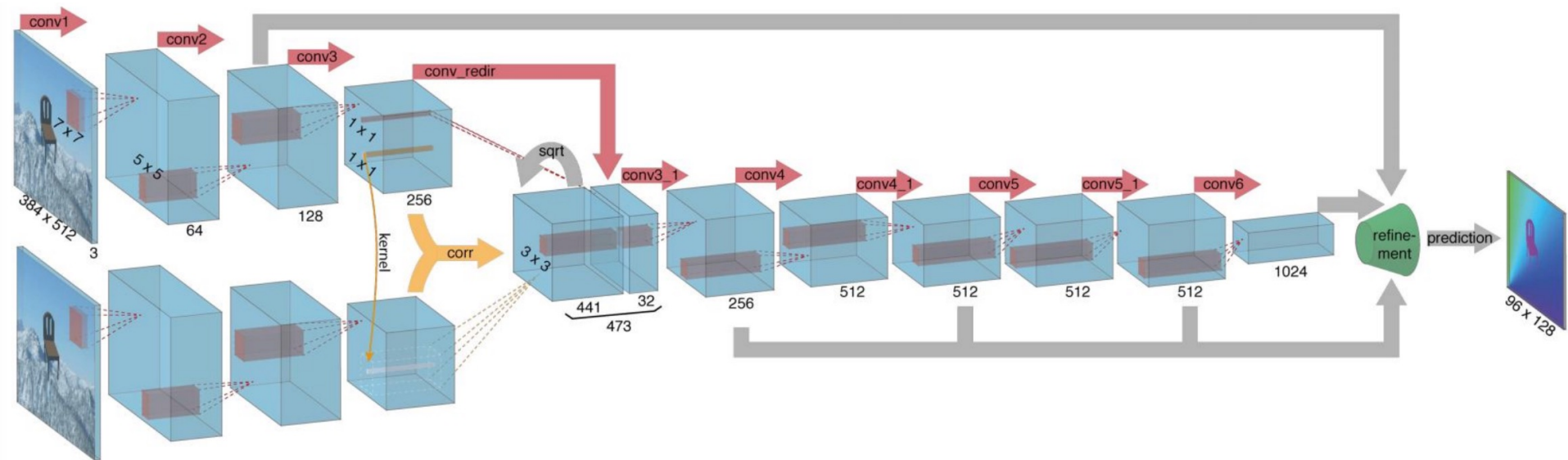
FlowNet: architecture 1

- Stack both images $\rightarrow$ input is now $2 \times \text{RGB} = 6$ channels



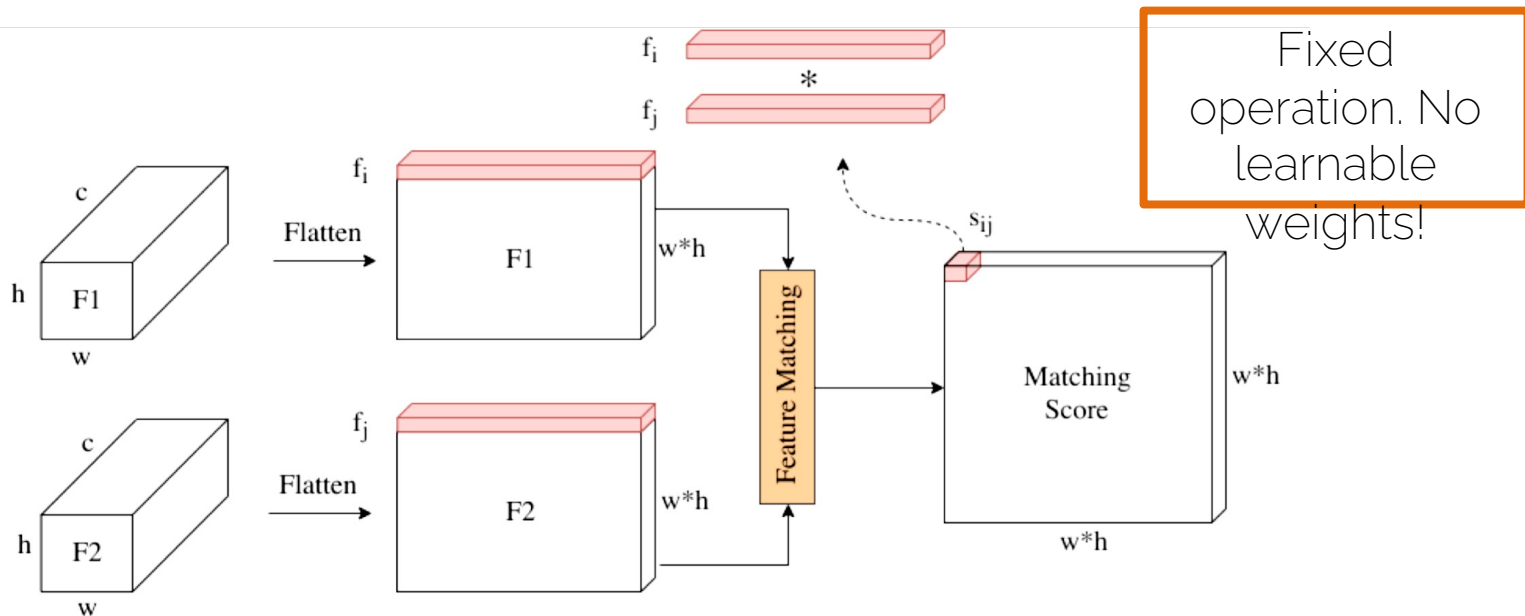
FlowNet: architecture 2

- Siamese architecture



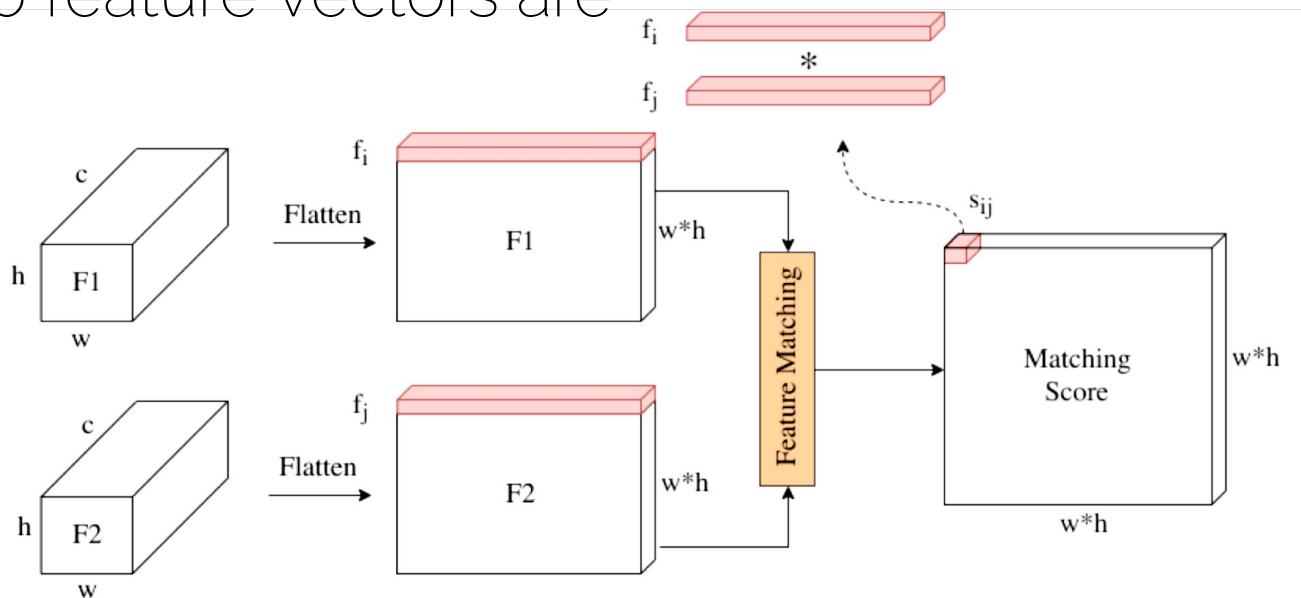
Correlation layer

- Multiplies a feature vector with another feature vector



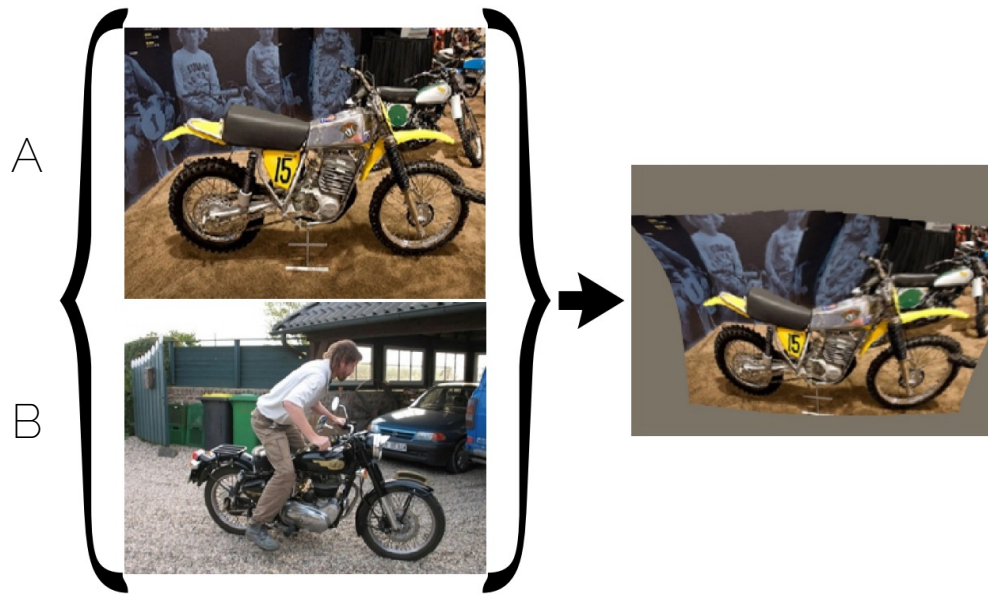
Correlation layer

- The matching score represents how correlated these two feature vectors are



Correlation layer

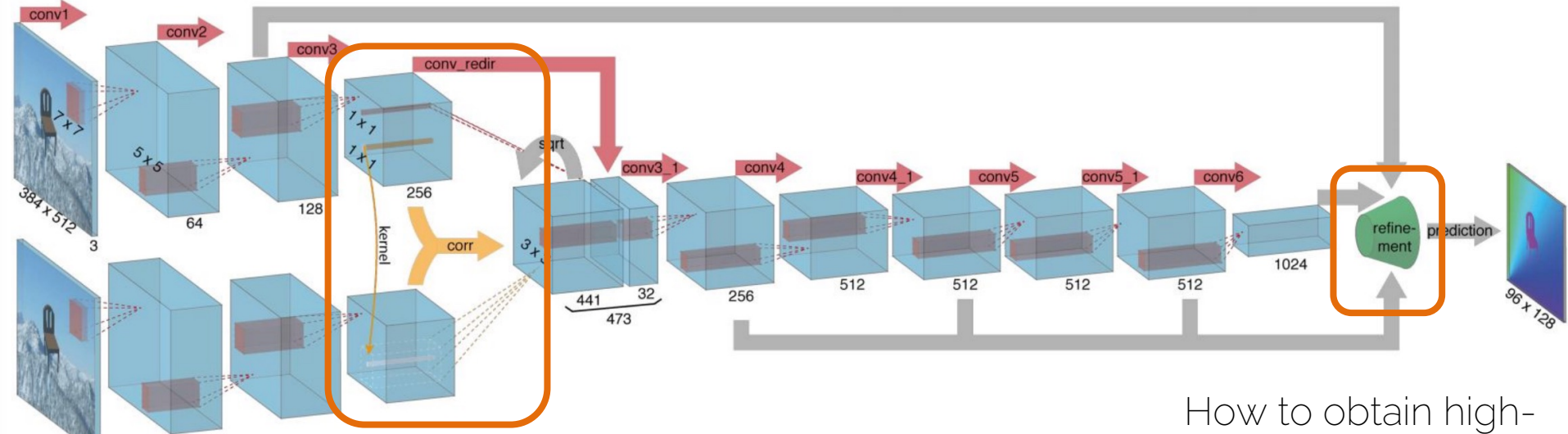
- Hint for anyone interested in 3D reconstruction:
Useful for finding image correspondences



Find a
transformation
from image A
to image B

FlowNet : architecture 2

- Two key design choices

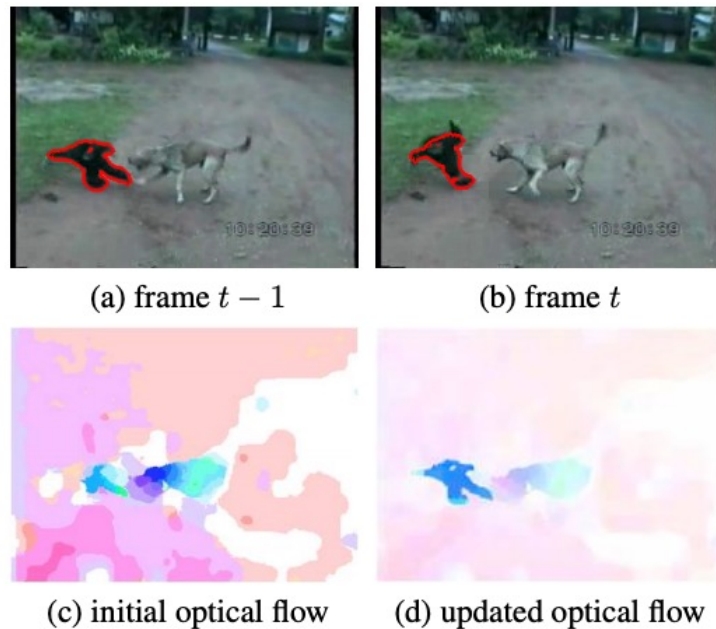


How to combine the information from both images?

How to obtain high-quality results?

Can we do VOS with OF?

- Indeed!
- Better if we focus on the flow of the object
- We can improve segmentation and OF iteratively (no DL yet)



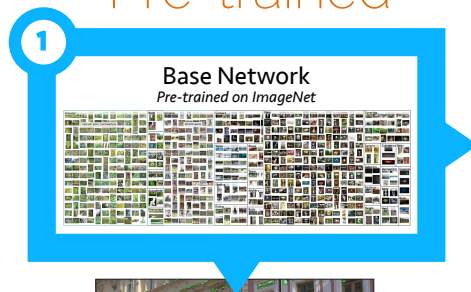
OSVOS

First-frame fine-tuning

- Goal: Learn the **appearance** of the object to track
- Main contribution: separate training steps
 - Pre-training for 'objectness'.
 - First-frame adaptation to specific object-of-interest using fine-tuning.

One-shot VOS

Pre-trained



Results on frame N
of test sequence



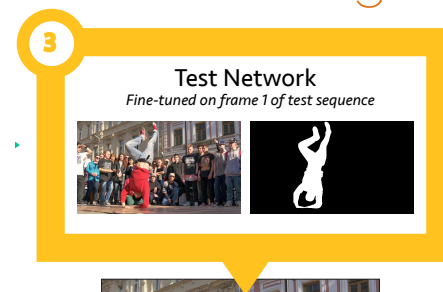
Edges and
basic image
features

Training



Learns how to
do video
segmentation

Finetuning

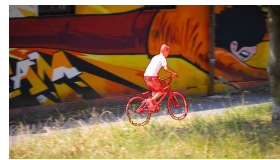


Learns which
object to
segment

One-shot VOS

- One-shot: we see the first frame ground truth
- Finetuning step: this is used to technically *overfit* to the test sequence first frame. Overfitting is therefor used to learn the appearance of the foreground object (and the background!)
- Test time: each frame is processed independently → no temporal information

Frame-based segmentation



- **PRO**: it recovers well from occlusions (unlike mask propagation or optical flow-based methods)



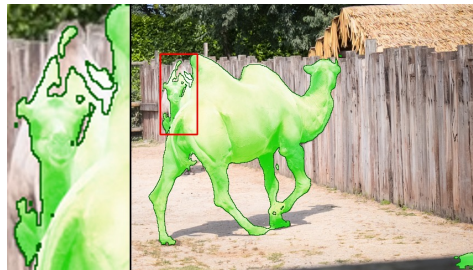
- **CON**: it is temporally inconsistent



Experiments: highly dynamic scenes

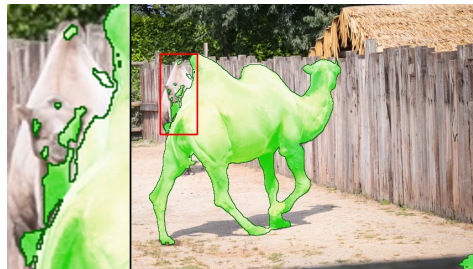
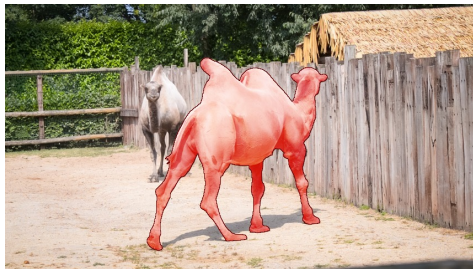


Experiments: accuracy vs annotations

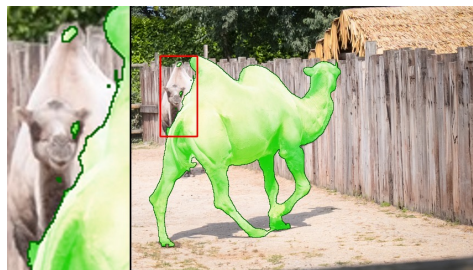


Two camels!

Another
annotation
where the 2nd
camel is
background



Another
annotation



Mask is
refined

Observations

- OSVOS does not have an object of object shape.
- It is a pure appearance-based method, if the foreground (or the background) appearance changes too much, the method fails

First frame

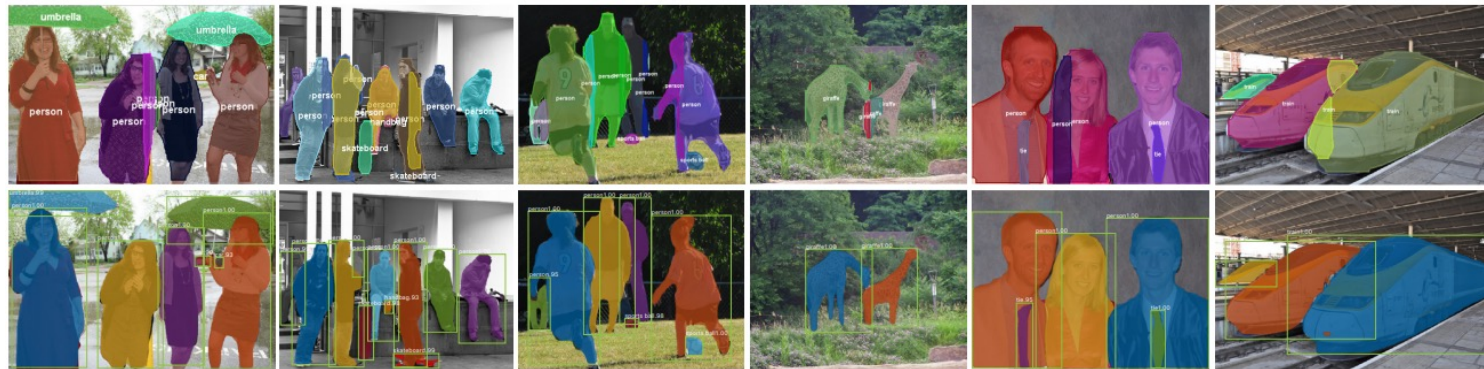


He was occluded in the first frame, therefore the network never learned he was background.



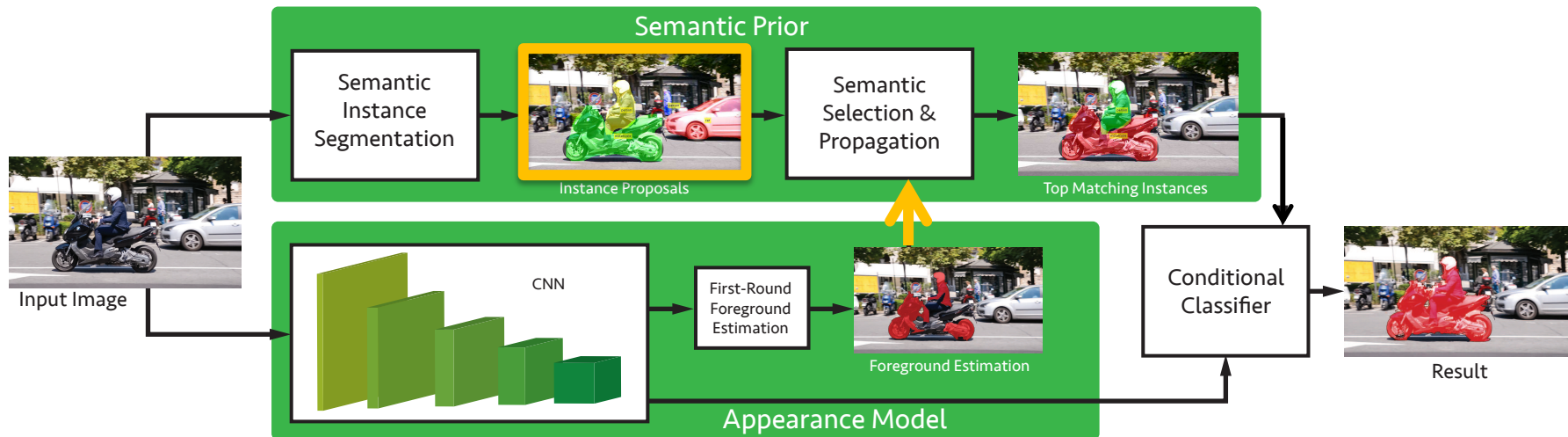
But wait....

- We have already seen models that have an idea of object shape..
- Instance segmentation methods!



OSVOS-S: Semantic propagation

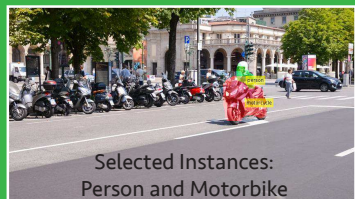
Semantic prior branch that gives us proposals to select from



Prior: semantics stay coherent throughout the sequence

OSVOS-S: Semantic propagation

Semantic Selection



Frame 0

Semantic Propagation

Instance Segmentation
Proposals



First-Round
Foreground Estimation



Top Person and
Motorbike



Frame 18

Frame 24

Frame 30

Frame 36

Drifting problem

- If the object greatly changes its appearance (e.g., though pose or camera changes), then the model is not powerful anymore



- But this change was gradual....

Drifting problem

- If the object greatly changes its appearance (e.g., though pose or camera changes), then the model is not powerful anymore



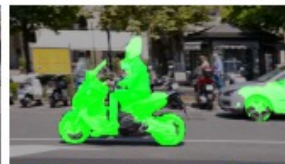
Why not gradually update the model?

OnAVOS: Online Adaptation

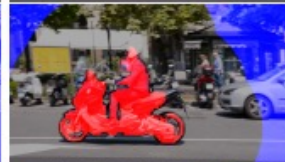
- Online adaptation: adapt model to appearance changes every frame – not just the first frame.
- Iteratively fine-tune the model on previous prediction every frame.
- **CON:** Extremely slow.

OnAVOS: Online Adaptation

un-adapted
baseline



adaptation
targets



Blue = background
samples

Red = foreground
samples

online
adapted



ground
truth



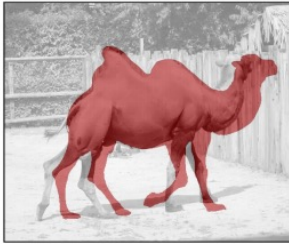
P. Voigtlander and B. Leibe. "Online adaptation of convolutional neural networks for video object segmentation". BMVC 2017

Mask Refinement

- Assumption: an object, i.e., a mask, does not move a lot from frame to frame.
- We can often start with an approximate mask (either from previous frame or from coarse estimate).
- We can then 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.

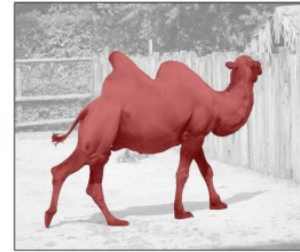
MaskTrack

Input frame t



Mask estimate $t-1$

MaskTrack ConvNet



Refined mask t

Why the name?



MaskTrack

- Training inputs can be simulated!
 - Like displacements to train the regressor of Faster-RCNN
 - Very similar in spirit to Tracktor



(a) Annotated image



(b) Example training masks

Worth reading

- S. Jain et al. "Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos." CVPR 2017. → Optical flow propagation
- A. Khoreva et al. "Lucid Data Dreaming for Video Object Segmentation," IJCV 2019 → clever data augmentation.
- X. Li et al. „Video object segmentation with re-identification" CVPRW 2017. → use reidentification techniques to recover from occlusions

Proposal-based approaches

Proposal Generation

Until now:

- Input is the whole image
- Proposals are put on top just to refine

Now:

Input are proposals
Goal is to "link" them (much like we did in tracking-by-detection)

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

PReMVOS

- 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
- Refinement
 - Fully-convolutional segmentation network trained to refine the segmentation given a proposal bounding box

PReMVOS: Overview



Proposal generation

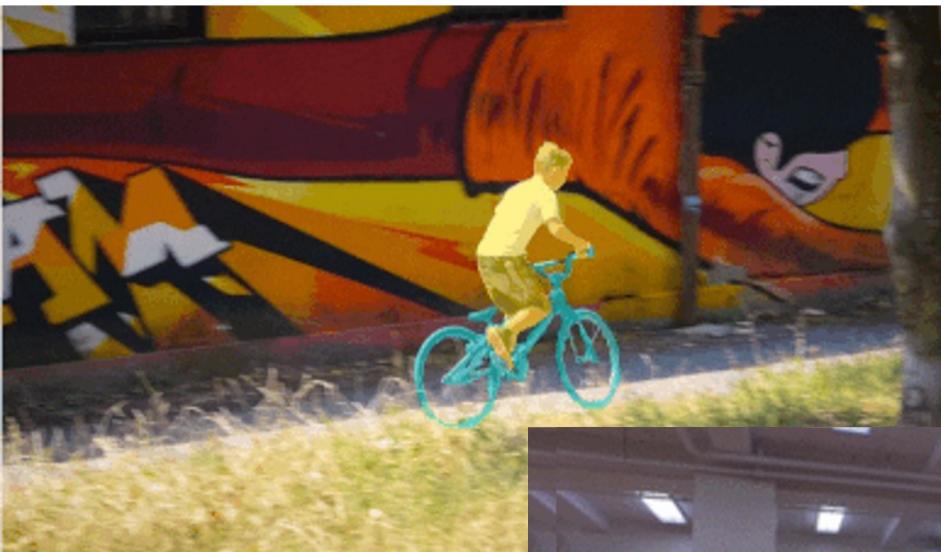
Refinement

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

PReMVOS: results

- Very complex but a winner
- DAVIS Challenge 2018 Winner
- Youtube-VOS Challenge 2018 Winner



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?

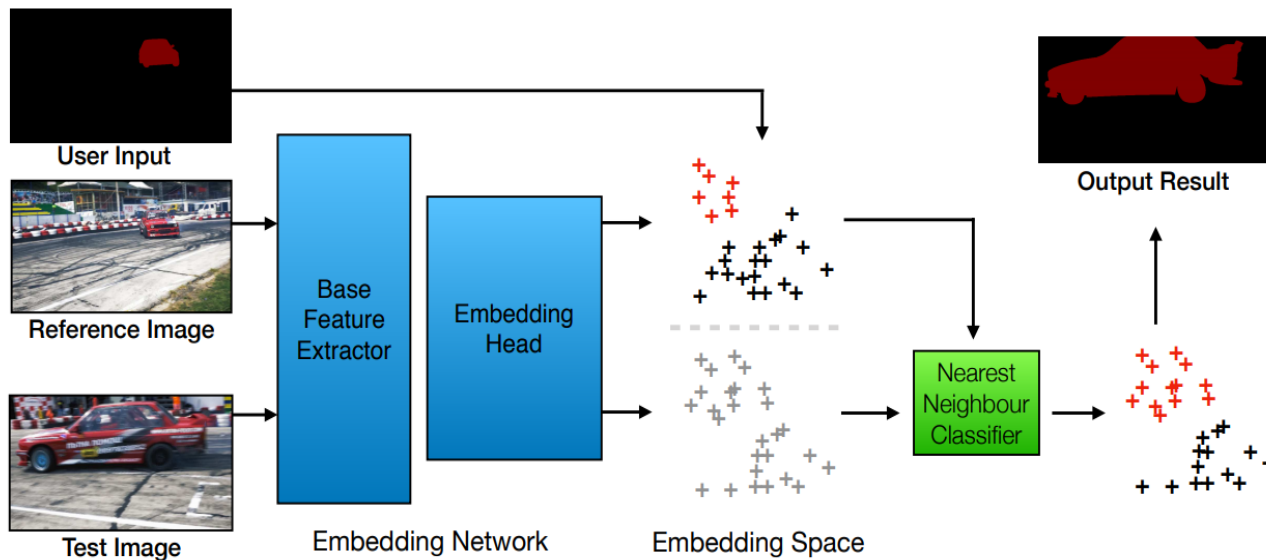
Retrieval approaches

Pixel-wise retrieval

- 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.

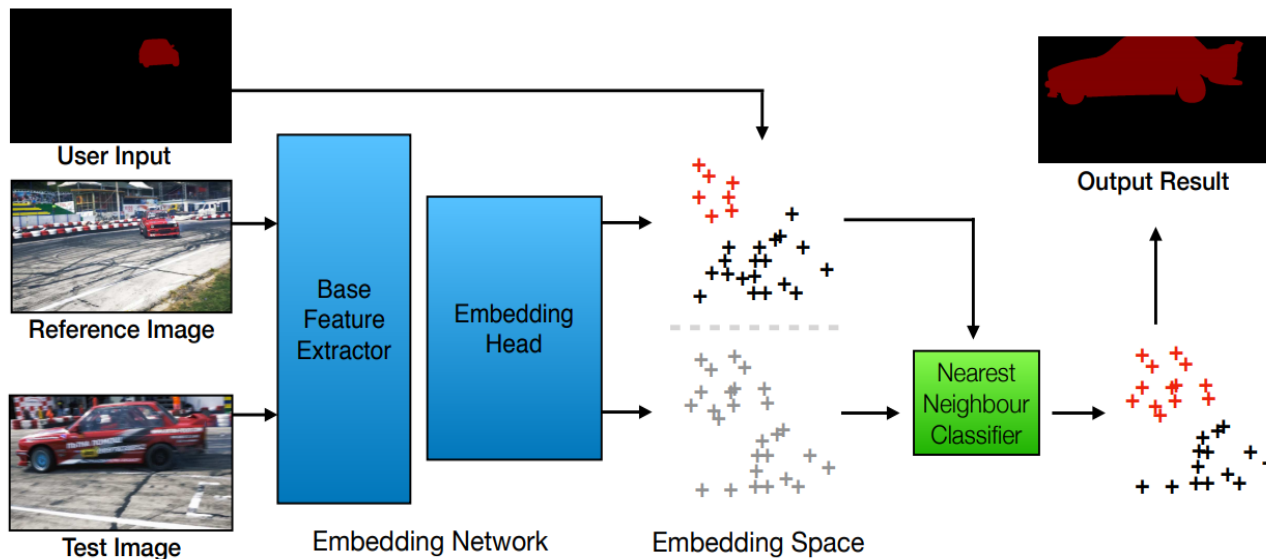
Pixel-wise retrieval

- Training: use the triplet loss to bring foreground pixels together and separate them from background pixels



Pixel-wise retrieval

- Test: embed pixels from both annotated and test frame, and perform a nearest neighbor search for the test pixels.



We do not need to retrain the model for each sequence, nor finetune

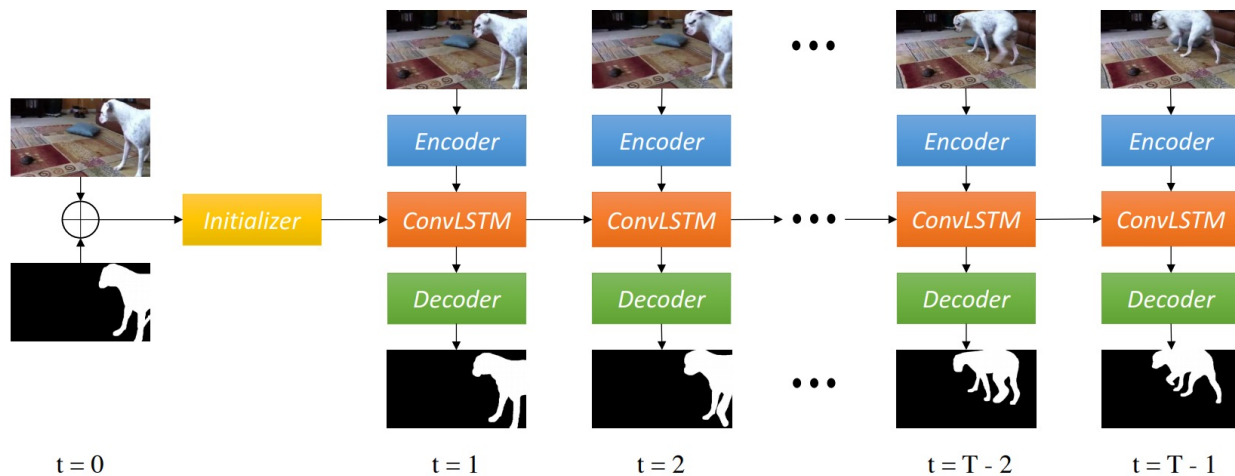
We are dealing with video

- Which is a sequence of images....
- And we have not talked about....
- Recurrent Neural Networks!

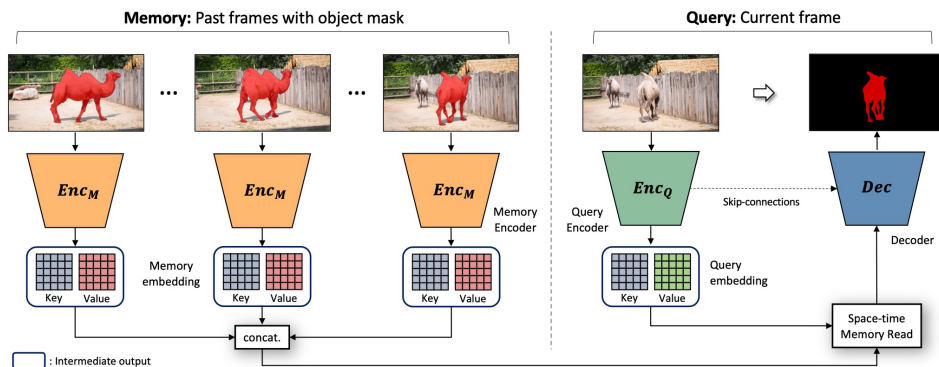
Spatio-temporal approaches

Temporal LSTM

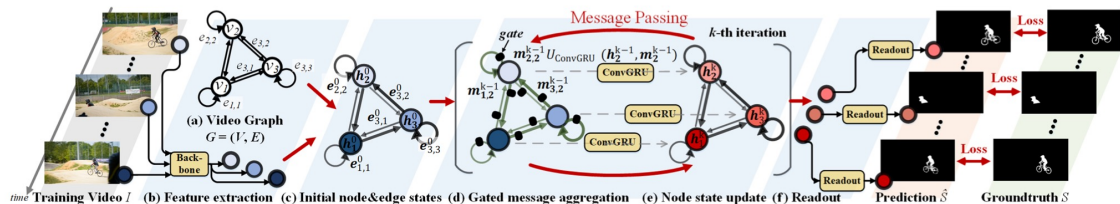
- One-shot video object segmentation
- If we have multiple objects, each of them is predicted independently



Transformers, GNNs



S. Oh "Video Object Segmentation using Space-Time Memory Networks". ICCV 2019



W. Wang et al. "Zero-Shot Video Object Segmentation via Attentive Graph Neural Networks". ICCV 2019

Overview of the methods

- Video Object Segmentation (VOS)
 - OSVOS: First-frame fine-tuning (appearance model)
 - OSVOS-S: + semantic guidance through proposals (shape)
 - OnAVOS: Online Adaptation (stronger appearance model)
 - MaskTrack: Mask Refinement
 - Lucid: clever data augmentation
 - ReID-VOS: Object Appearance Re-Identification
 - PReMVOS: putting it all together
 - Seq2seq recurrent architecture
- Shape
Appearance
Motion
Matching

Evaluation and metrics

Metrics for VOS

- **Region similarity:** Jaccard index (IoU) of ground truth mask and predicted mask.
- **Contour Accuracy:** measures the precision and recall of the boundary pixels. This is put together in the F-measure.

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Metrics for VOS

- You can use error measure statistics
- **Region similarity:** Jaccard index (IoU) of ground truth mask and predicted mask.
 - Mean: average for the dataset
 - Decay: quantifies the performance loss (or gain) over time. → This is currently used to judge temporal stability
 - Recall: fraction of sequences scoring higher than a threshold

Tracking and Segmentation

VOS -> MOTs

- Video Object Segmentation (VOS) is limited by:
 - First frame mask given (in the supervised case)
 - Short video clips with objects present in almost all frames
 - Objects in a video are (mostly) of different categories
 - Few objects to track (max around 7 per video)
- Multi-Object Tracking and Segmentation (MOTS)
 - Scenarios with a large number of objects (20-40), mostly of the same category (e.g., pedestrians)
 - Long sequences
 - No first frame annotation provided, one has to deal with appearing and disappearing objects.

MOTS dataset

- Segmentations coming to MOTChallenge pedestrian tracking dataset



Video Object Segmentation

Disclaimer

- This lecture was done borrowing material from:
 - Prof. Xavier Giró, Technical University of Catalonia (UPC)
 - Jonathon Luiten, RWTH Aachen