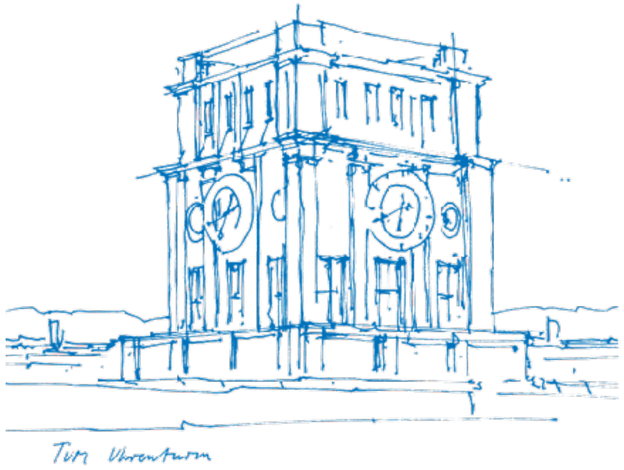
Computer Vision III:
Video object segmentation

Nikita Araslanov
10.01.2023

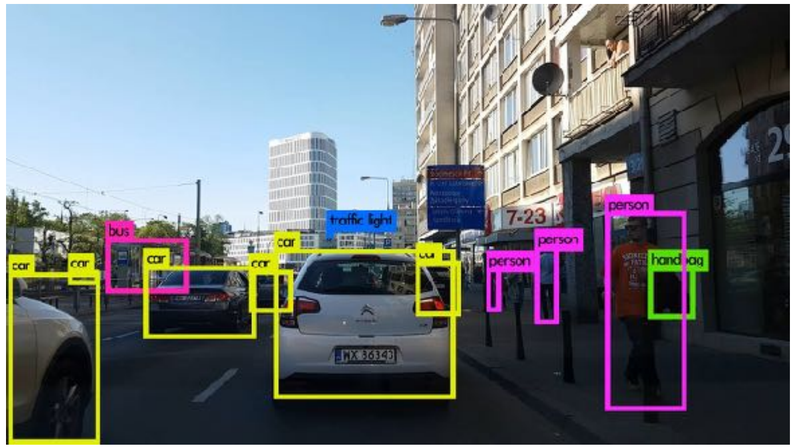
Content credit:
Prof. Laura Leal-Taixé
https://dvl.in.tum.de



Course progress

1. Introduction
2. Object detection 1
3. Object detection 2
4. Multiple object tracking 1
5. Multiple object tracking 2
6. Semantic segmentation
7. Instance segmentation
8. Panoptic segmentation
9. Video object segmentation ← we are here
10. Transformers (17.01)
11. Unsupervised and semi-supervised scene understanding (24.01)
12. Q&A (31.01)
13. Exam: 03.03 (register this week)

Video Object Segmentation



Object Detection

Object Tracking

Object Segmentation

Video Object Segmentation

Video Object Segmentation

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

DAVIS 2017

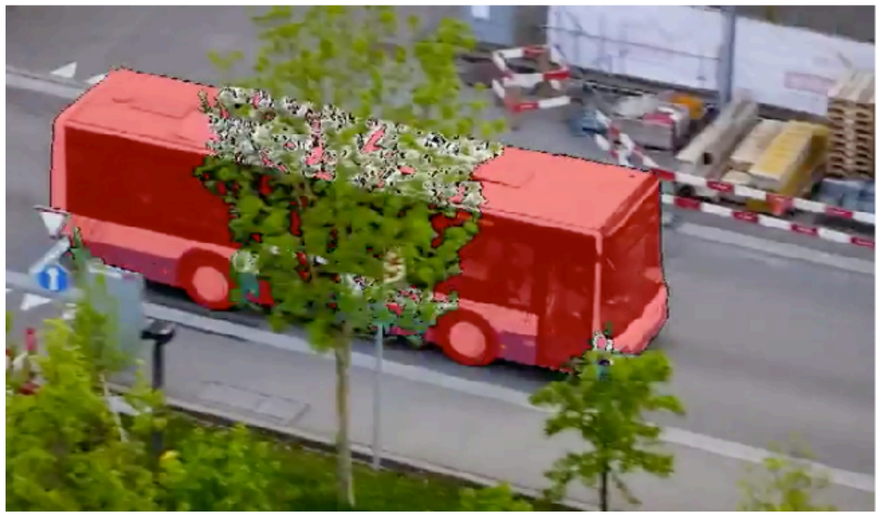
VOS: Some challenges

- Strong viewpoint/appearance changes

DAVIS 2017

VOS: Some challenges

- Strong viewpoint/appearance changes
- Occlusions



DAVIS 2017

VOS: Some challenges

- Strong viewpoint/appearance changes
- Occlusions
- Scale changes

DAVIS 2017

VOS: Some challenges

- Strong viewpoint/appearance changes
- Occlusions
- Scale changes
- Illumination
- Shape
- …

VOS: Some challenges

- Strong viewpoint/appearance changes
- Occlusions
- Scale changes
- Illumination
- Shape
- …

We need:

Appearance model

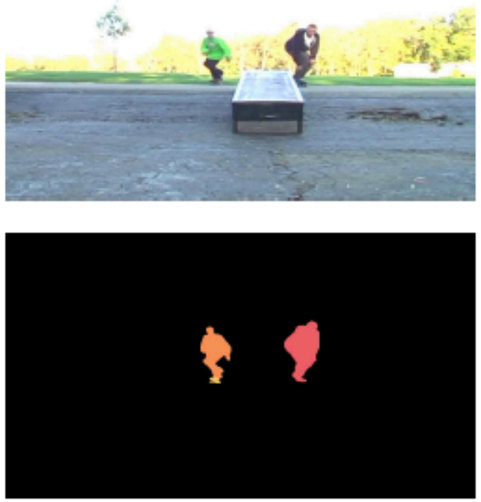
Motion model

VOS: Models

- **Appearance model**:
  - assumption: constant appearance
  - input: 1 frame;
  - output: segmentation mask.
- **Motion model**: ⟵ may be optional
  - assumption: smooth displacement; brightness constancy.
  - input: 2 frames;
  - output: motion (optical flow)
- Advanced models take advantage of both.

VOS: Tasks

“Semi-supervised” (one-shot) VOS



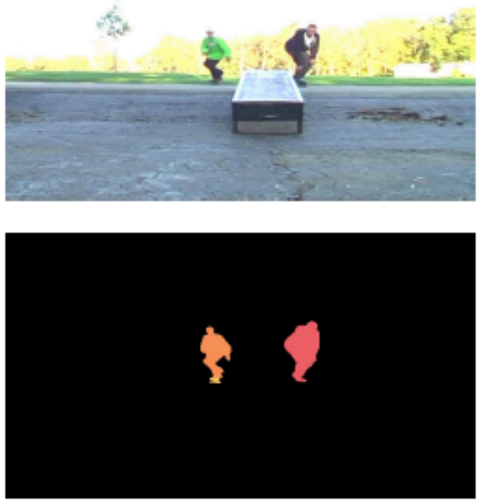
Inference time
- input: video + object mask in the first frame

“Unsupervised” (zero-shot) VOS

Inference time
- input: video (without any labels)

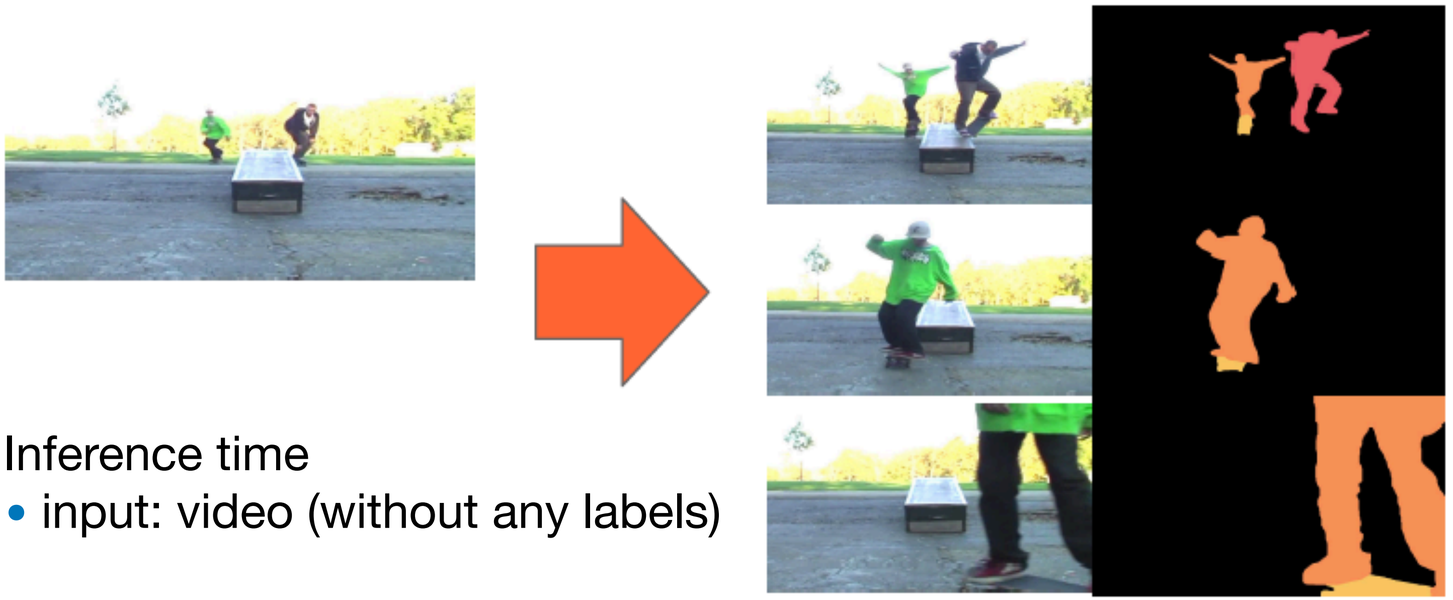
VOS: Tasks

“Semi-supervised” (one-shot) VOS



Inference time
- input: video + object mask in the first frame

“Unsupervised” (zero-shot) VOS



Inference time
- input: video (without any labels)

1. Identify objects of interest (e.g. using instance segmentation);
2. Establish temporal consistency.

Cons: Typically complex models (active research area).

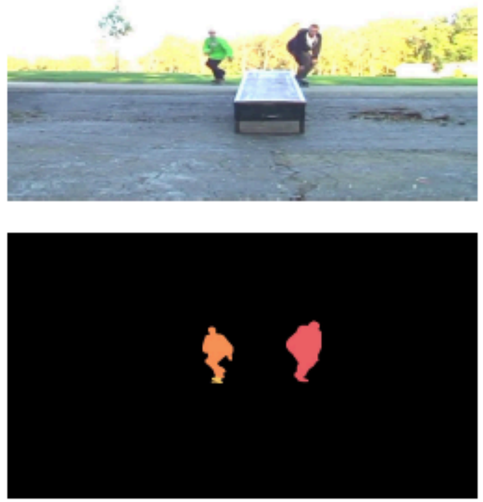
What to track?

- Choosing the objects to track can be subjective (esp. online):

- Offline tracking – considering the whole video – may provide a better clue (e.g. based on object permanence).

VOS: Tasks

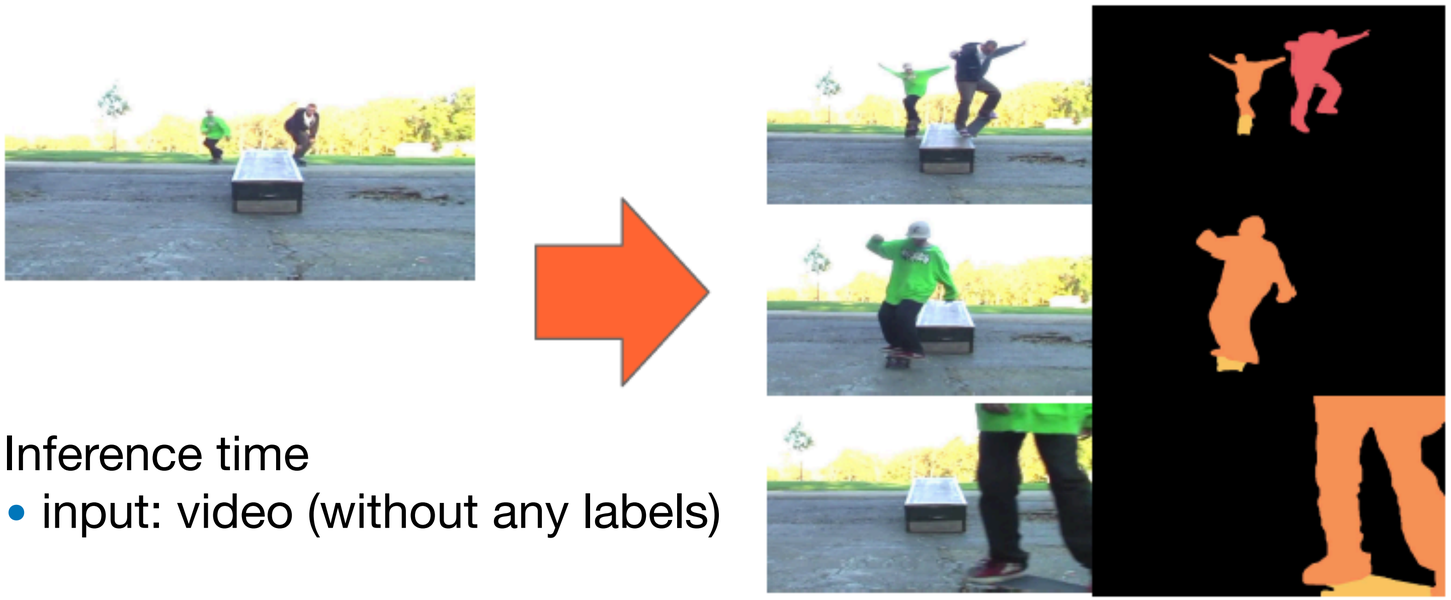
“Semi-supervised” (one-shot) VOS



Inference time
- input: video + object mask in the first frame

Focus on temporal consistency (this lecture)

“Unsupervised” (zero-shot) VOS



Inference time
- input: video (without any labels)

Semi-supervised VOS

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 dense (i.e. pixel-level) tracking

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)

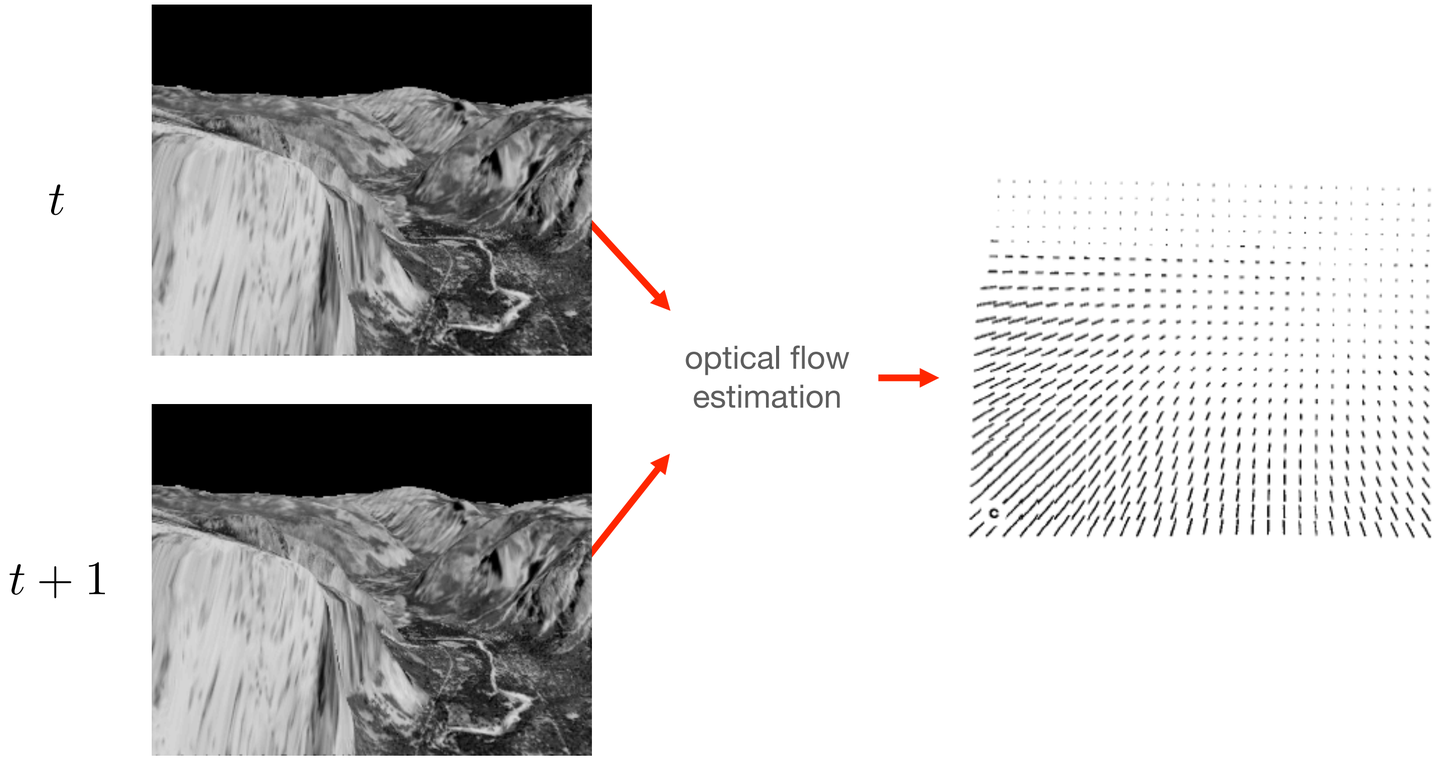
- We need large-scale (annotated) datasets for learning-based methods

Image credit: https://davischallenge.org; https://youtube-vos.org.

Motion-based VOS

Optical flow

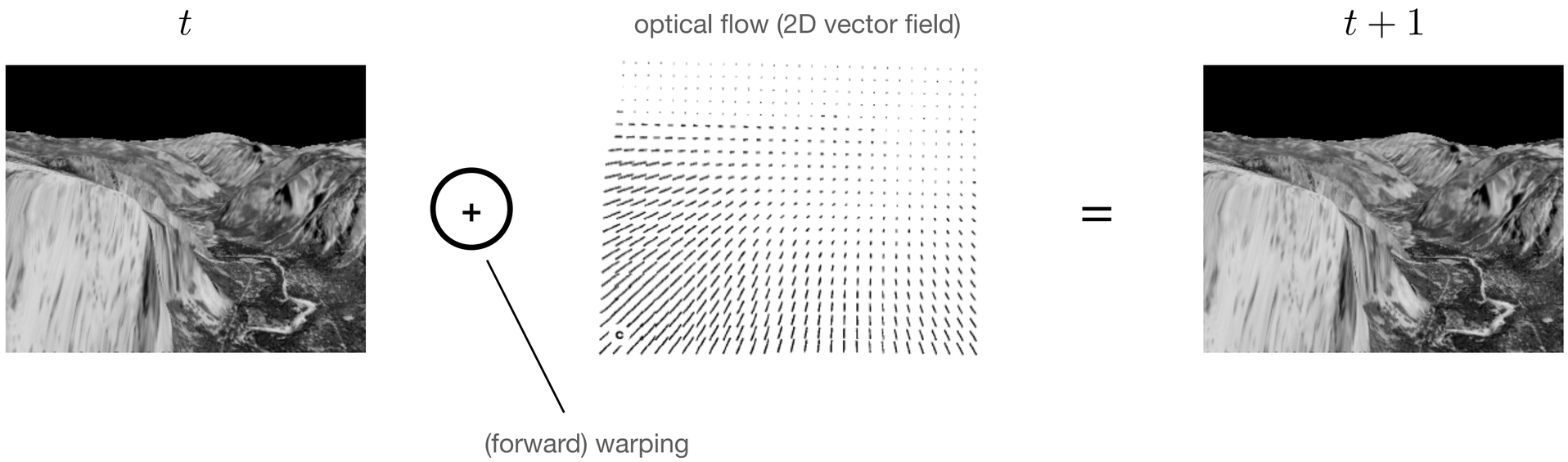
- Recall optical flow (from Computer Vision I):
  - a pattern of apparent motion (Lukas and Kanade, ’81; Horn & Schunk ’81);



[S. Roth; M. Black]

Optical flow

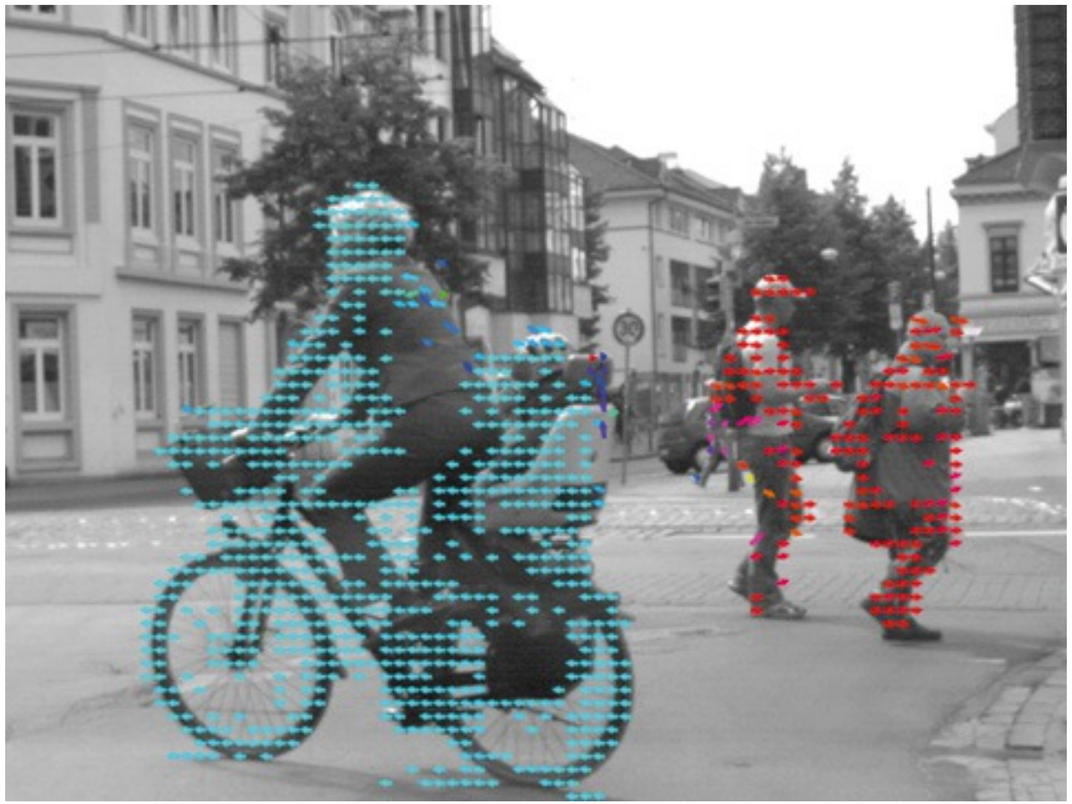
- Recall optical flow (from Computer Vision I):
  - a pattern of apparent motion (Lukas and Kanade, ’81; Horn & Schunk ’81);



[S. Roth; M. Black]

Motion segmentation

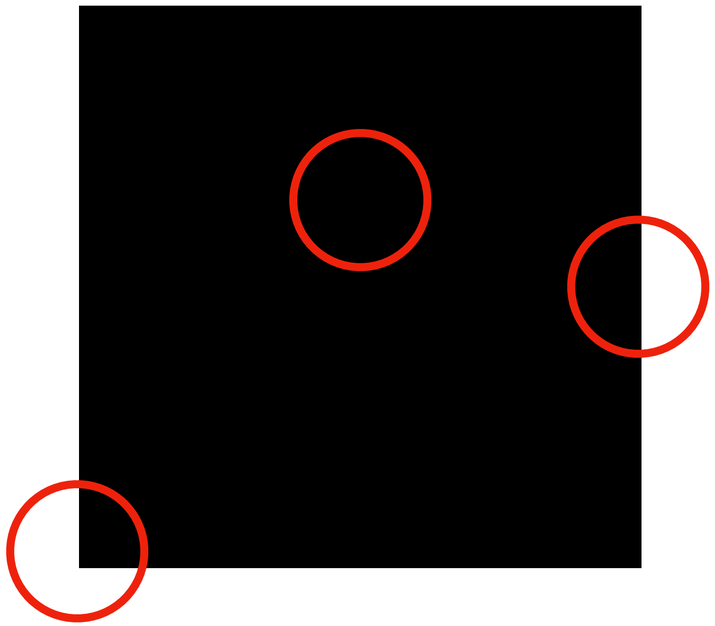
- We can segment some objects based on their motion:



Optical flow

- “perceived” 2D motion, not the real motion of the object.
- the aperture problem.

Consider the motion of the square
in 3 (fixed) observation areas:



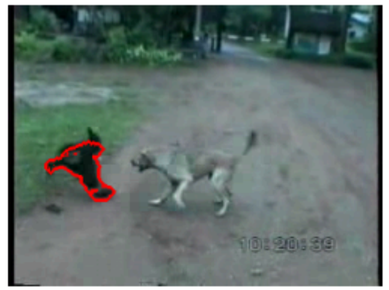
Optical flow

- Back in the day, optical flow used to be time-consuming.
  - ~80 seconds on a CPU for 640x480 image pair:

[Brox and Malik; 2010]

Can we do VOS with optical flow?

- Joint formulation of segmentation and optical flow estimation:



initial optical flow

update the object mask

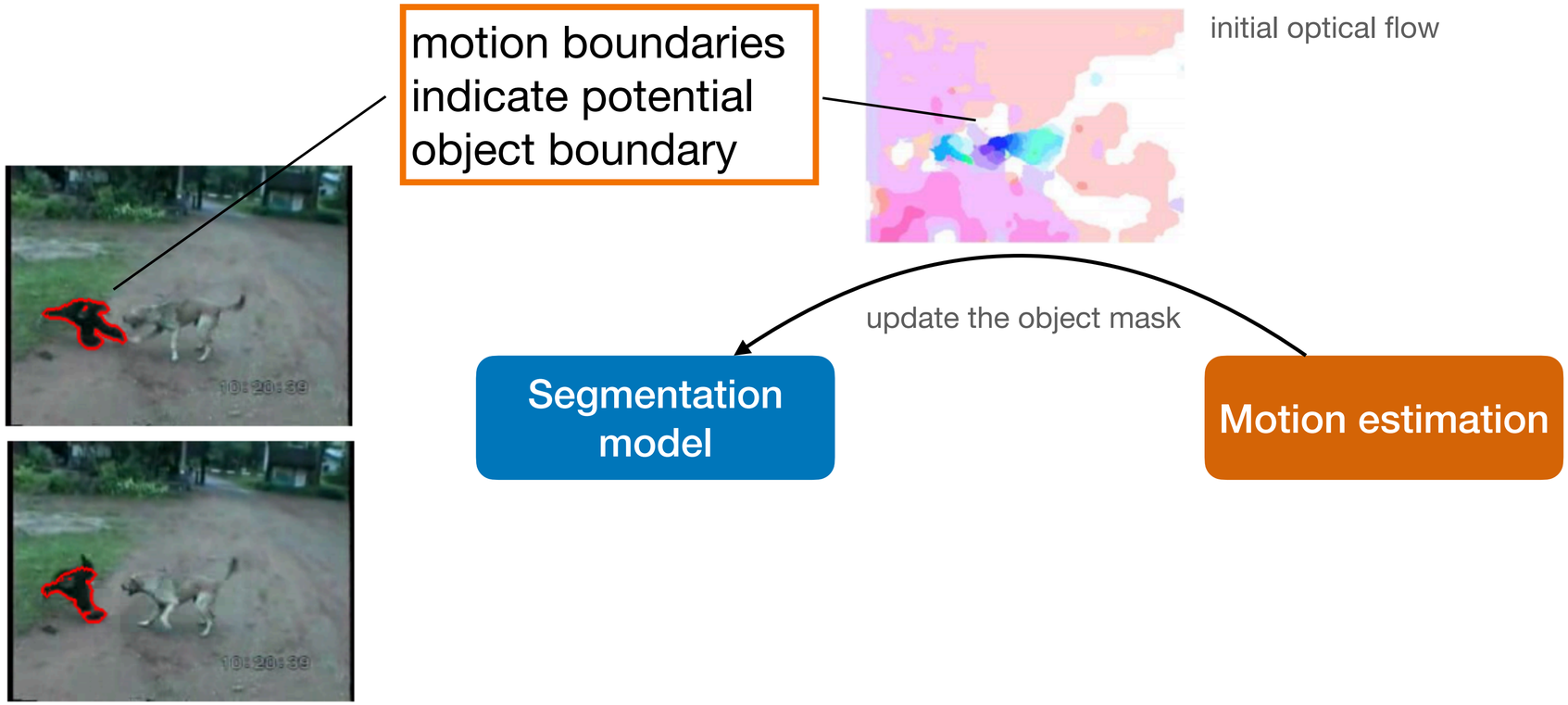
Segmentation model

Motion estimation

Y.H. Tsai et al. “Video Segmentation via Object Flow“. CVPR 2016

Can we do VOS with optical flow?

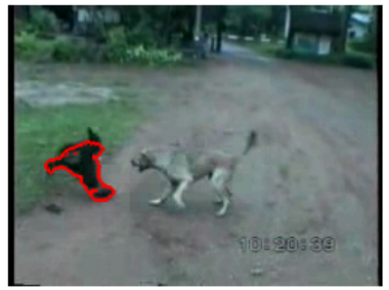
- Joint formulation of segmentation and optical flow estimation:



Y.H. Tsai et al. “Video Segmentation via Object Flow“. CVPR 2016

Can we do VOS with optical flow?

- Joint formulation of segmentation and optical flow estimation:



initial optical flow

Segmentation model

Motion estimation

update optical flow

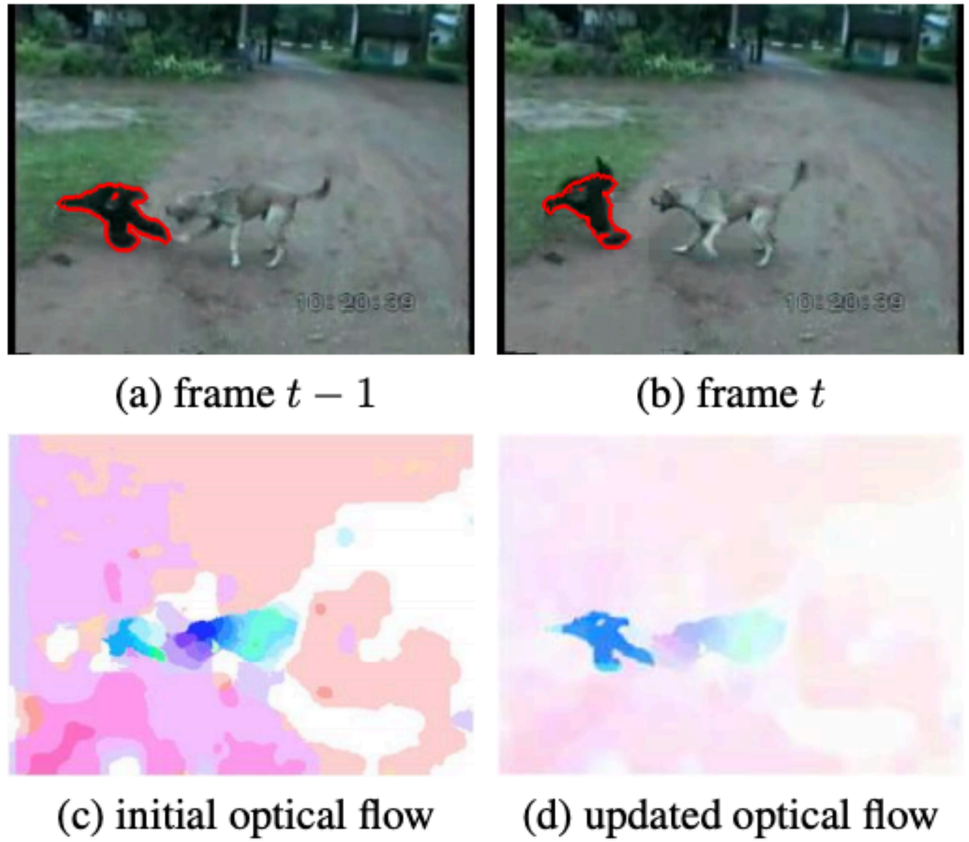
updated optical flow

object boundaries indicate motion discountinuities

Y.H. Tsai et al. “Video Segmentation via Object Flow“. CVPR 2016

Can we do VOS with optical flow?

- Joint formulation:
  - iteratively improving segmentation and motion estimation.
- Slow to optimise:
  - runtime: up to 20s (excluding OF).
- Initialisation matters:
  - we need (somewhat) accurate initial optical flow.
- DL to the rescue?



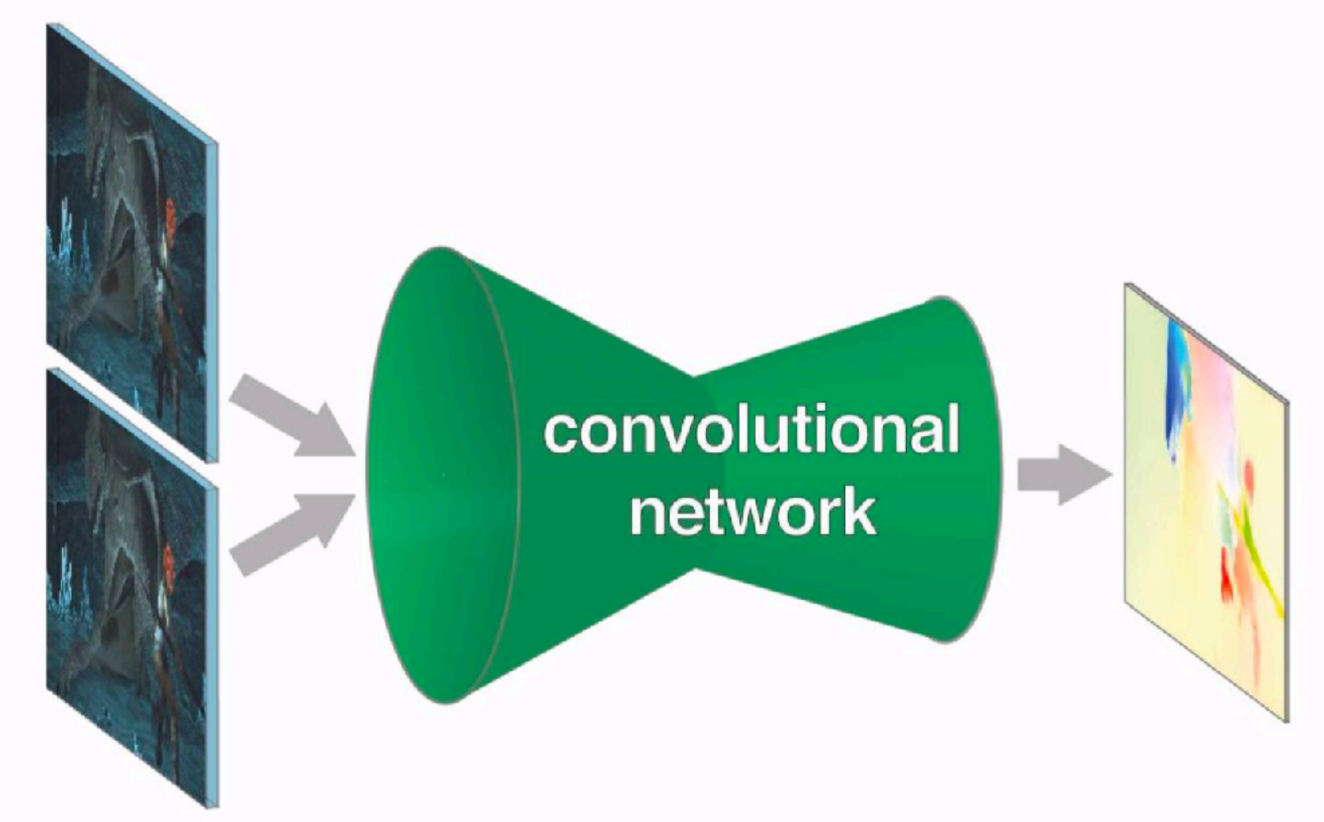
(a) frame $t-1$ (b) frame t

(c) initial optical flow (d) updated optical flow

Y.H. Tsai et al. “Video Segmentation via Object Flow“. CVPR 2016

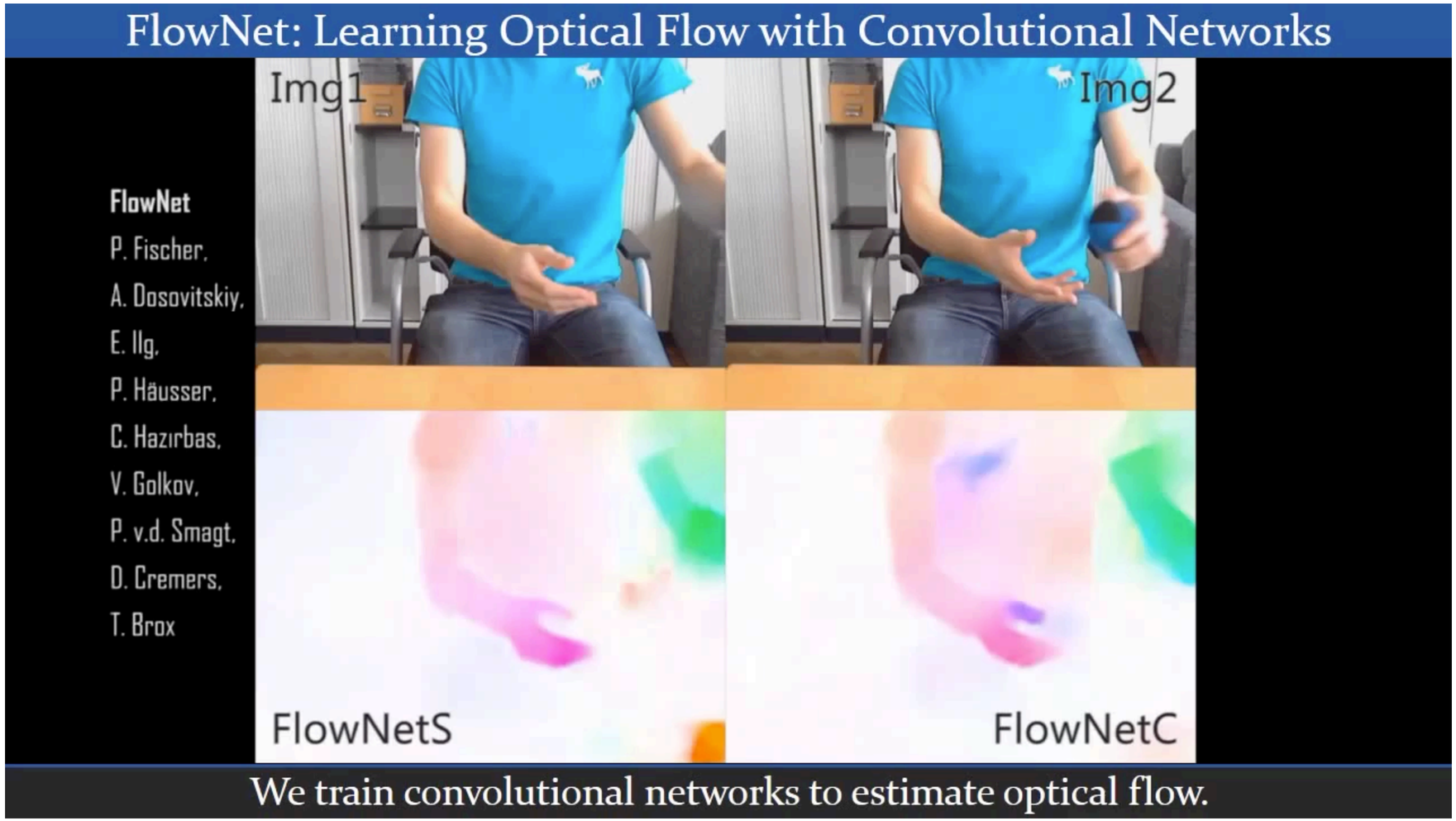
Optical flow with CNNs

- End-to-end supervised learning of optical flow



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

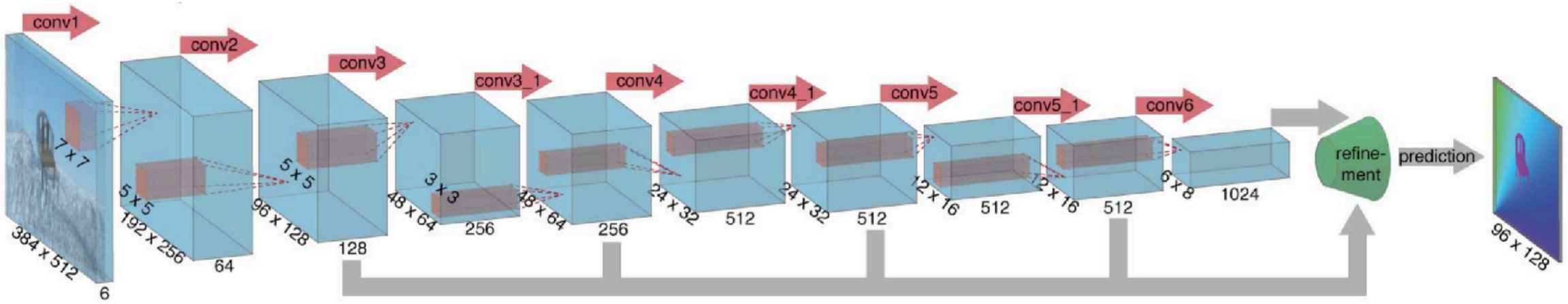
Optical flow with CNNs



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

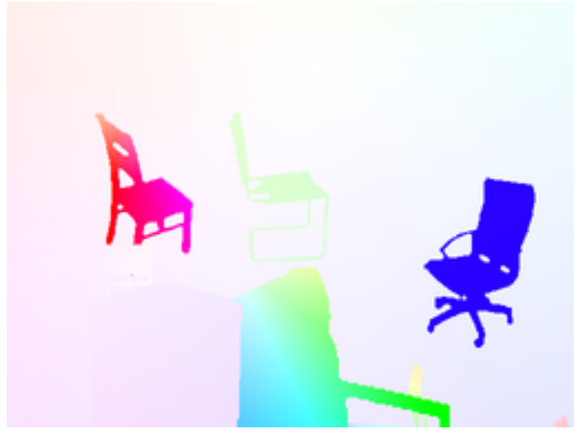
FlowNet: Architecture 1

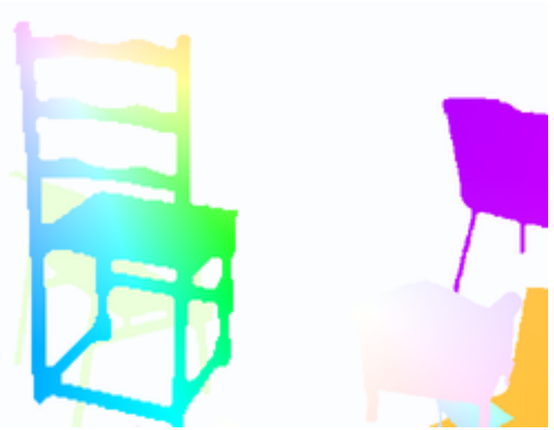
- Stack both images → input is now 2 x RGB = 6 channels



- Training with L2 loss from synthetic data

Optical flow from synthetic data

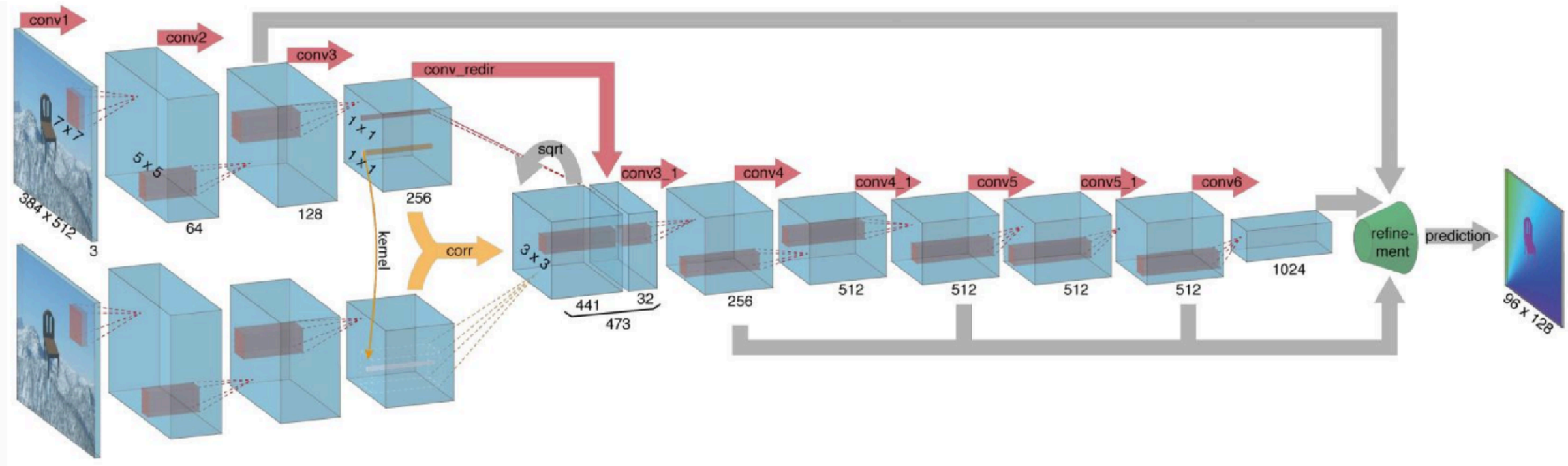




- Why chairs?
  - ...because we have large collections of 3D models.

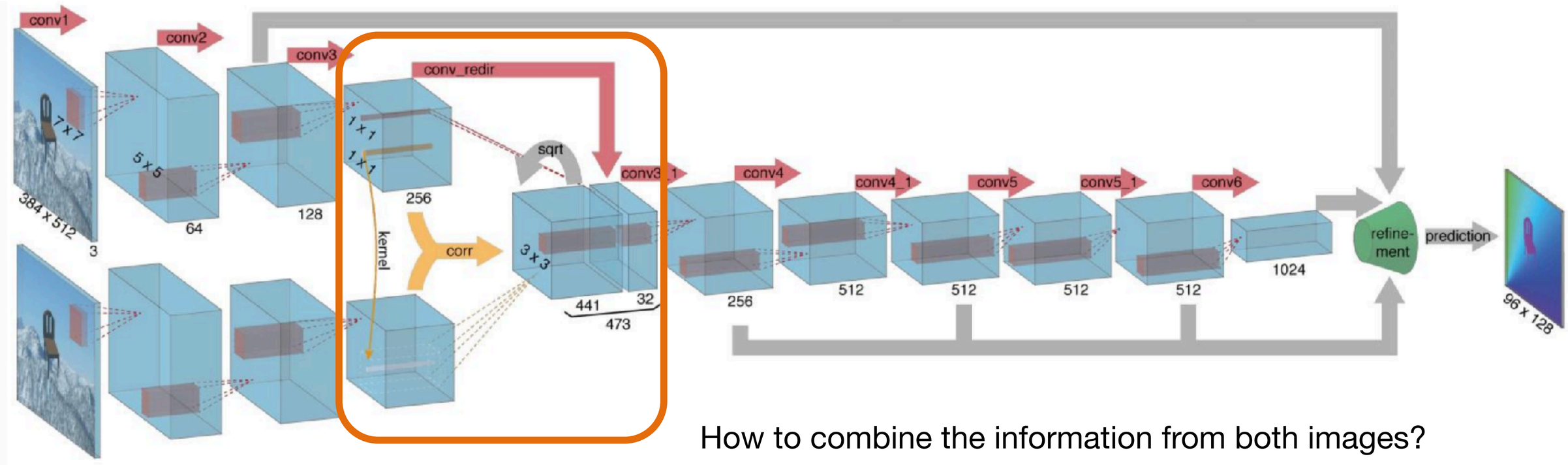
FlowNet: Architecture 2

- Siamese architecture



FlowNet: Architecture 2

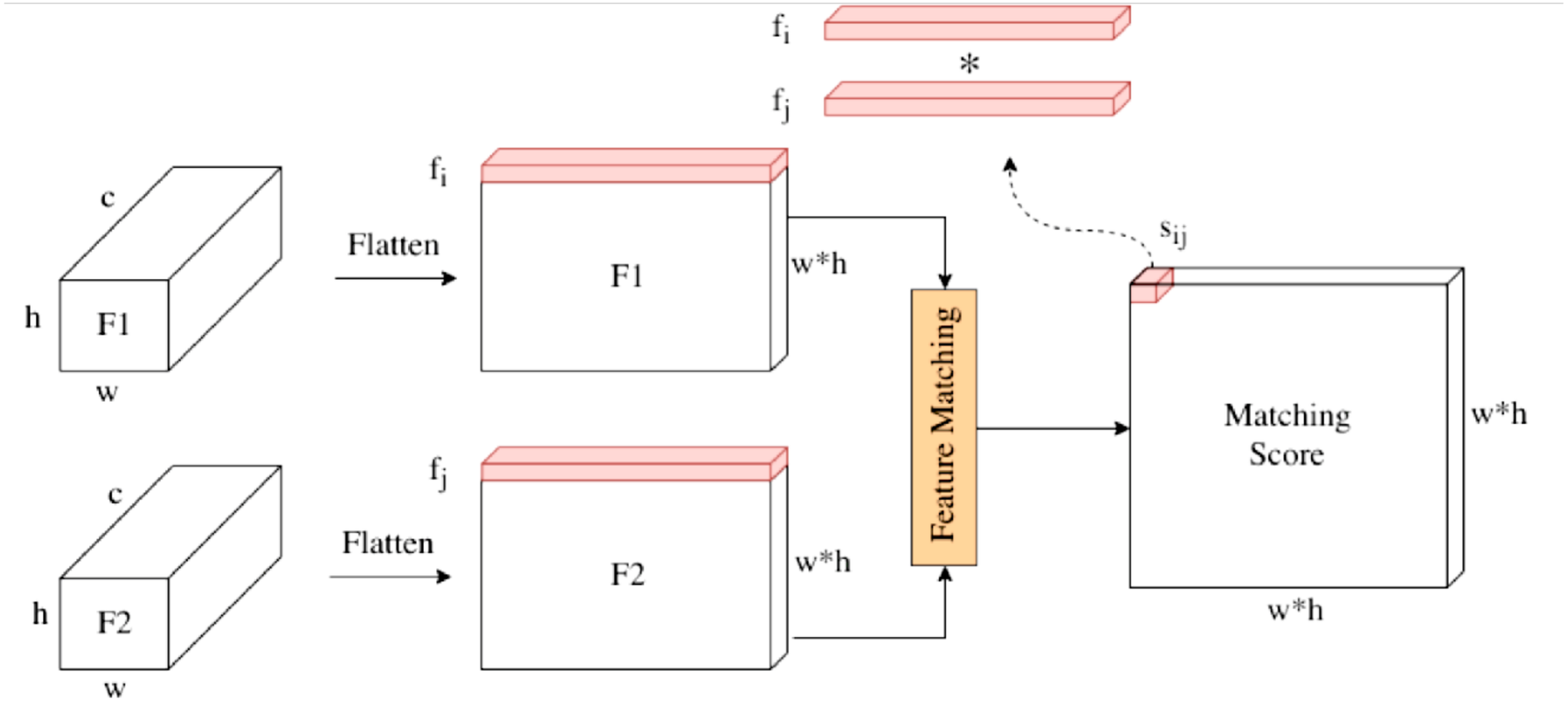
- Two key design choices



How to combine the information from both images?

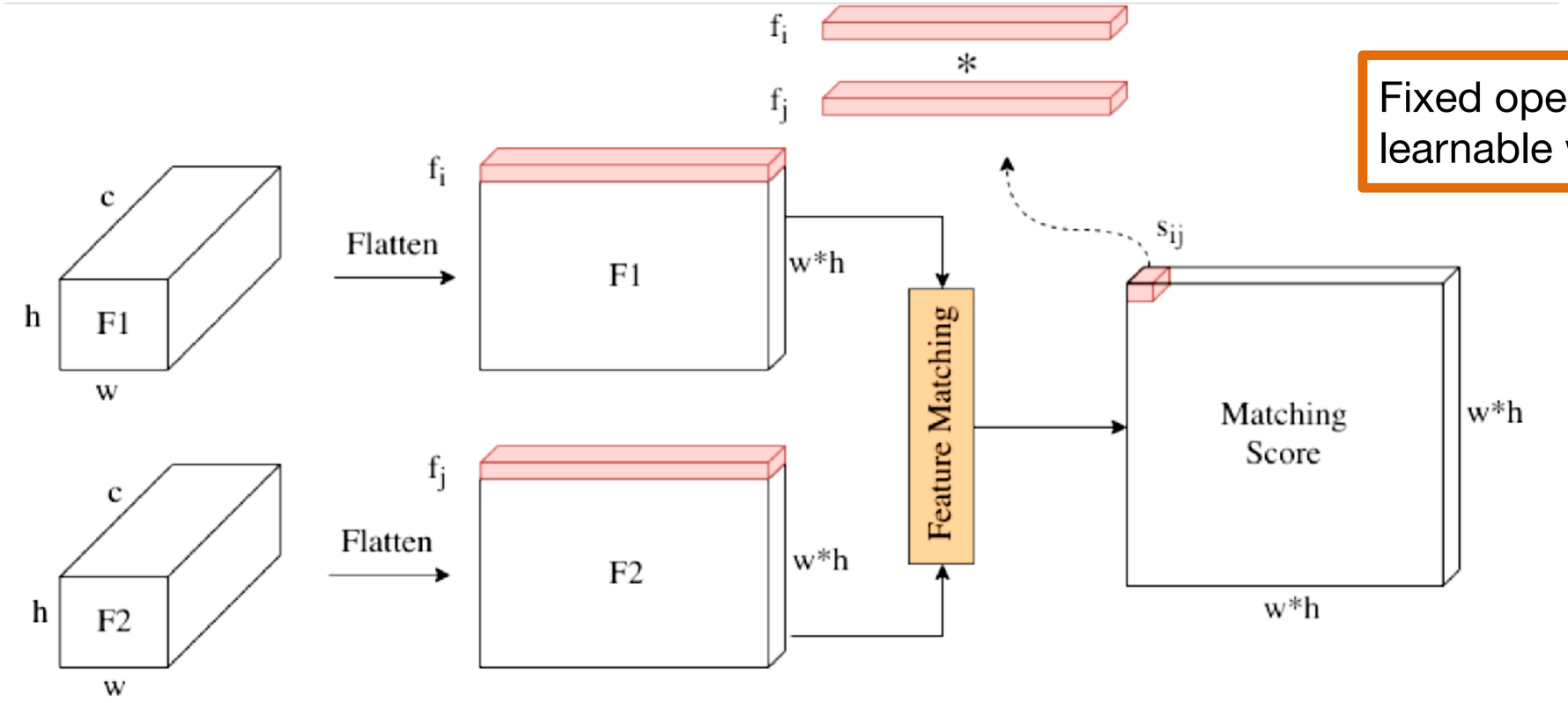
Correlation layer

- Given two feature tensors F1 and F2 compute pairwise dot-product



Correlation layer

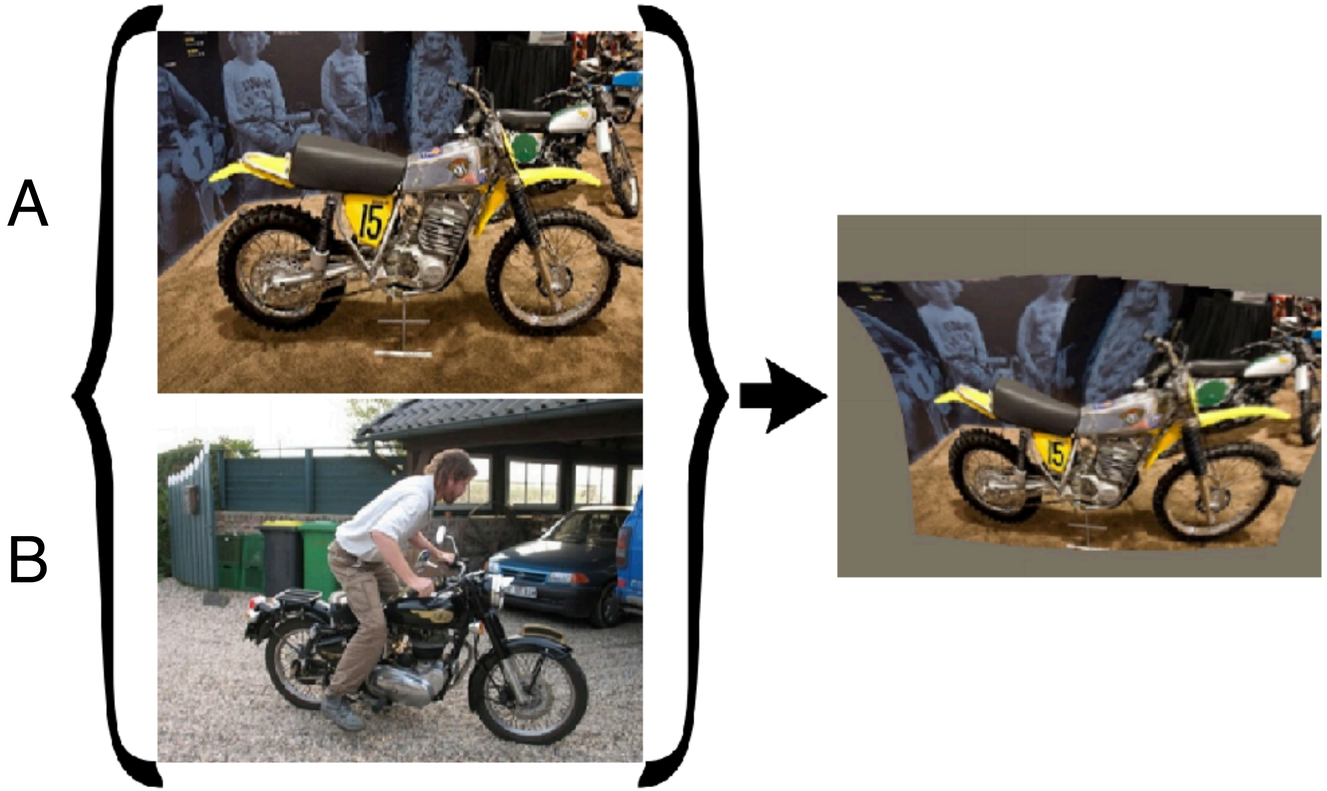
- The dot product measures similarity of two features



Fixed operation. No learnable weights!

Correlation layer

- Correlation layer is useful for finding image correspondences

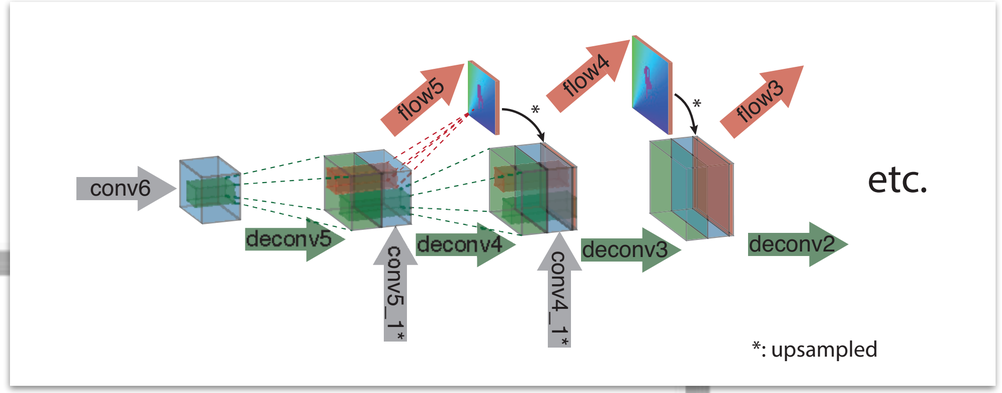


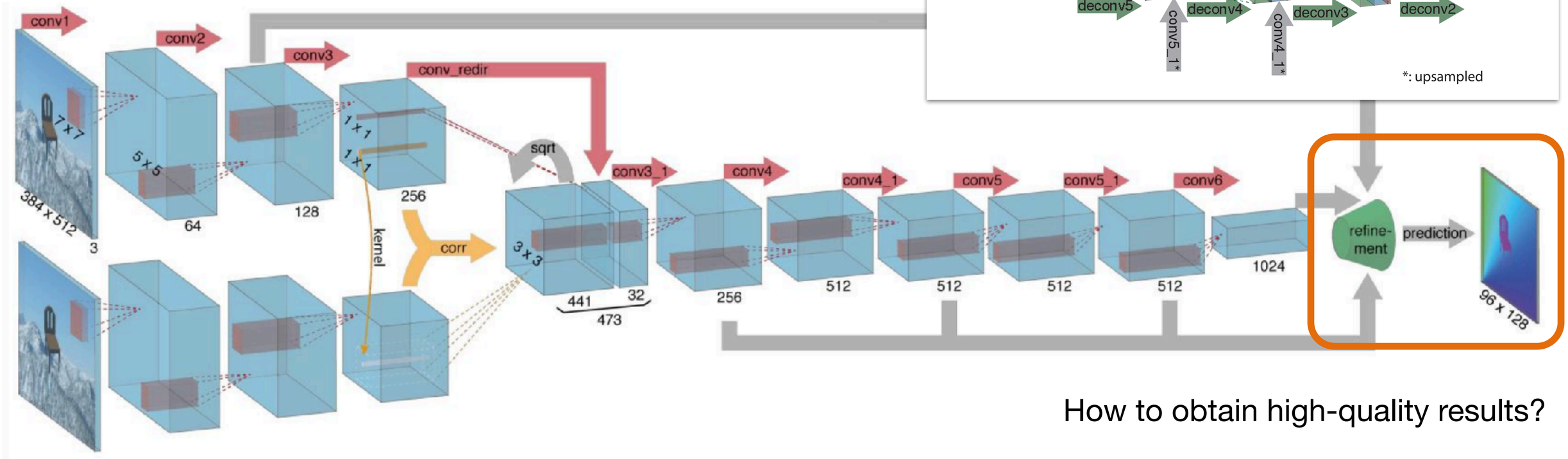
Find a transformation from image A to image B

I. Rocco et al. “Convolutional neural network architecture for geometric matching. CVPR 2017.

FlowNet: Architecture 2

- Two key design choices

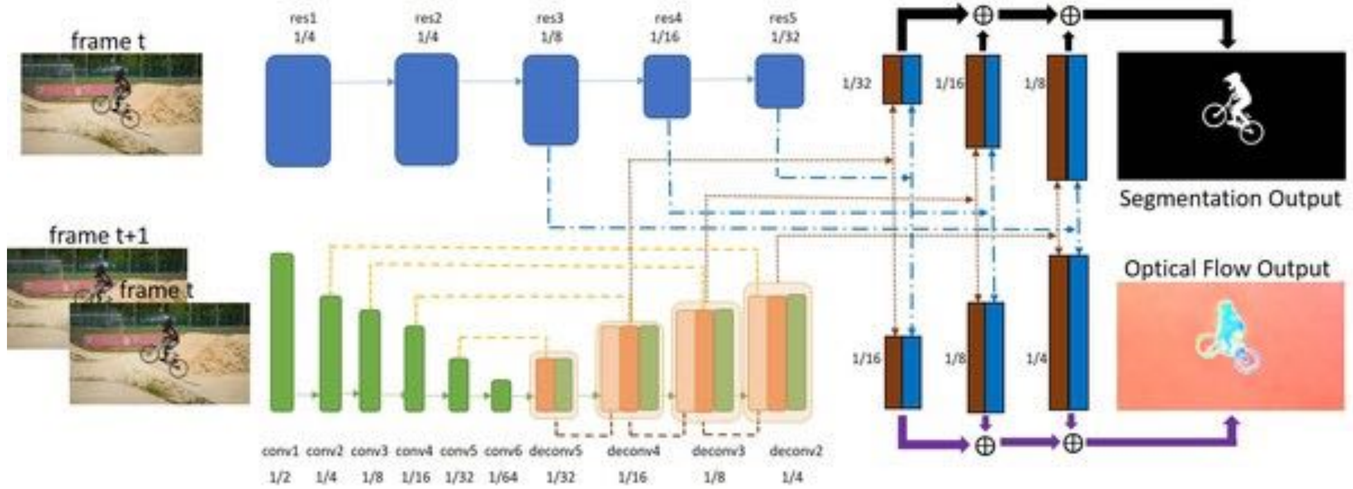




How to obtain high-quality results?

SegFlow

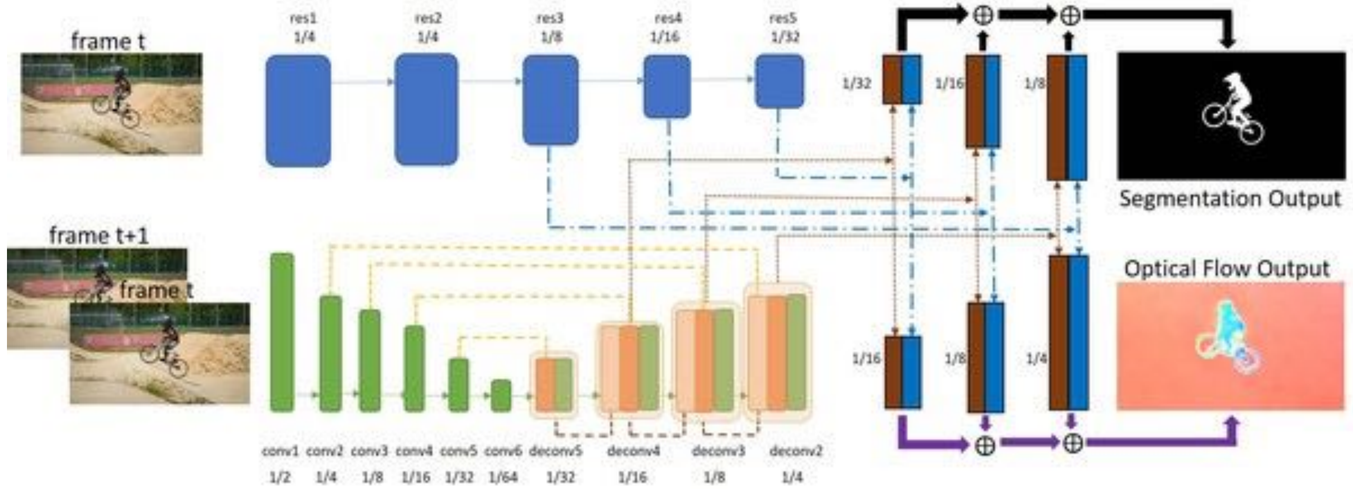
- Joint estimation of optical flow and object segment:



Cheng et al., “SegFlow: Joint Learning for Video Object Segmentation and Optical Flow”. ICCV 2017.

SegFlow

- Joint estimation of optical flow and object segment:

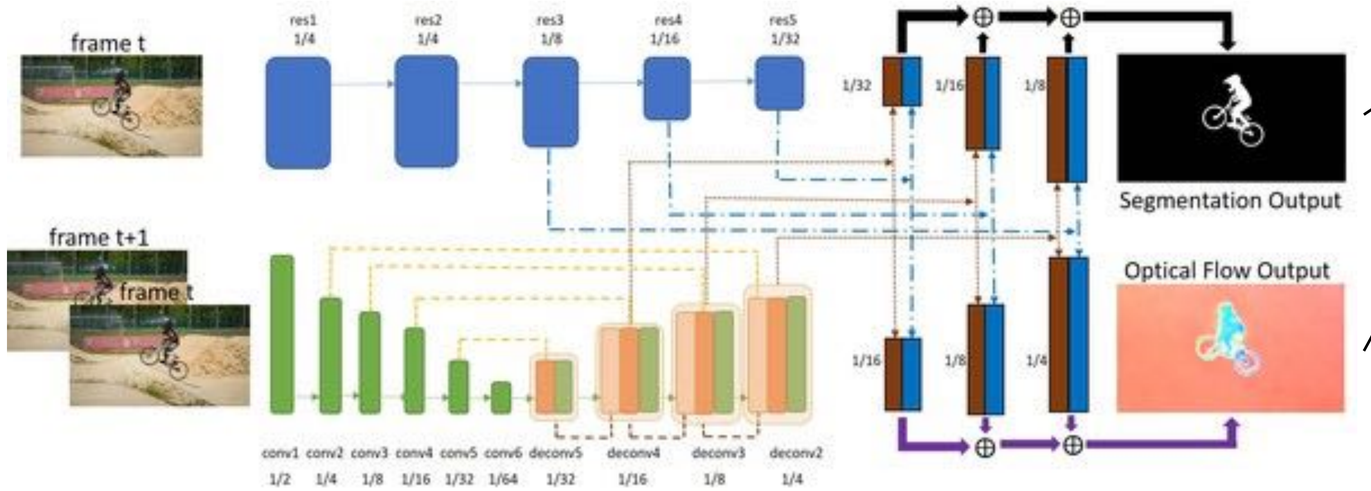


Joint feature representation at multiple scales

Cheng et al., “SegFlow: Joint Learning for Video Object Segmentation and Optical Flow”. ICCV 2017.

SegFlow

- Joint estimation of optical flow and object segment:



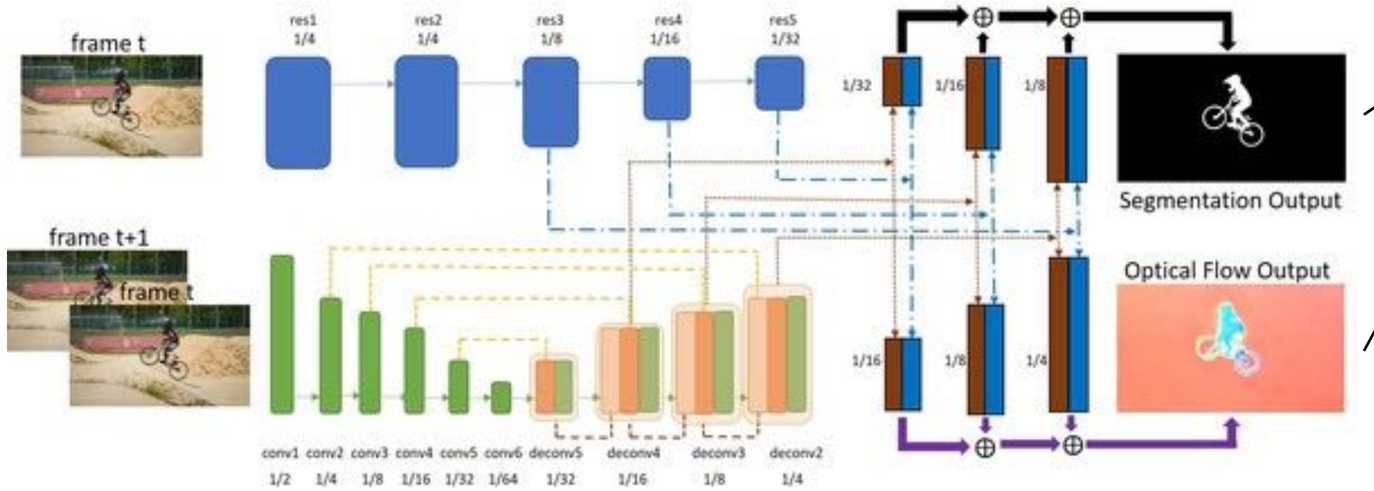
Supervision need not be synchronised

Joint feature representation at multiple scales

Cheng et al., “SegFlow: Joint Learning for Video Object Segmentation and Optical Flow”. ICCV 2017.

SegFlow

- Joint estimation of optical flow and object segment:



Supervision need not be synchronised

Joint feature representation at multiple scales

- Alternating optimisation:
  - fix one network to optimise the other.

Cheng et al., “SegFlow: Joint Learning for Video Object Segmentation and Optical Flow”. ICCV 2017.

Motion-based VOS

- We can obtain accurate estimates of optical flow with low latency;
- Naively applying optical flow to dense tracking has limited benefits:
  - due to severe (self-)occlusions, illumination changes, etc.
  - still an active area of research in semi-supervised VOS (dense tracking).
- Emerging techniques in a completely unsupervised setting:

(Yang et al., 2019)

Appearance-only VOS

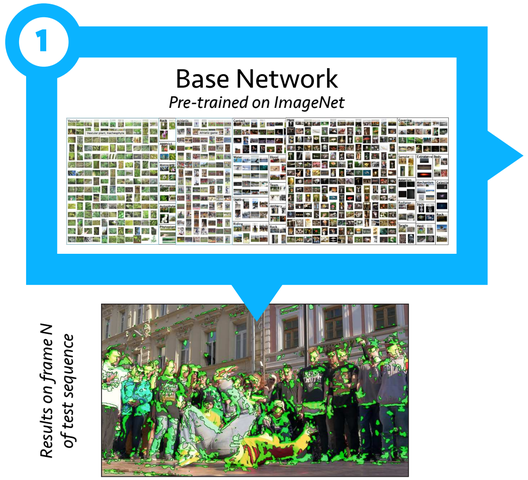
Appearance-only models

- Main idea:
  - Train a segmentation model from available annotation (including the first frame);
  - Apply the model to each frame independently;

- One-shot VOS (OSVOS): separate the training steps
  - Pre-training for ‘objectness’.
  - First-frame adaptation to specific object-of-interest using fine-tuning.

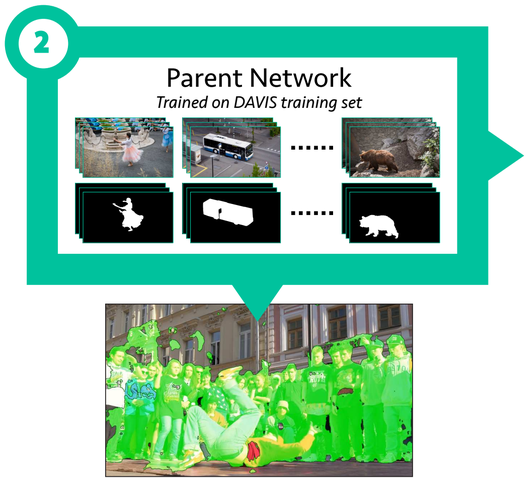
One-shot VOS

Pre-training



Edges and basic image features

Training



Learns how to do video segmentation

Finetuning

Learns which object to segment

S. Caelles et al. “One-shot video object segmentation”. CVPR 2017.

One-shot VOS

- One-shot: learning to segment sequence from one example (the first frame).
- This happens in the fine-tuning step:
  - the model learns the appearance of the foreground object.
- After fine-tuning, each frame is processed independently → no temporal information.
- The fine-tuned parameters are discarded before finetuning for the next video.

Experiments: Complex scenes

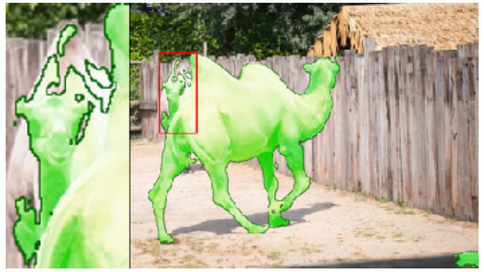
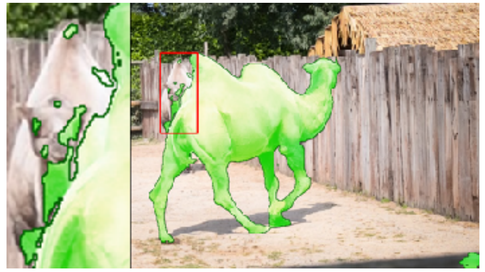
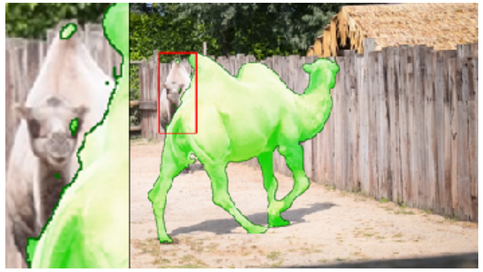
S. Caelles et al. “One-shot video object segmentation”. CVPR 2017.

Experiments: Accuracy vs Annotations

We can provide more examples...

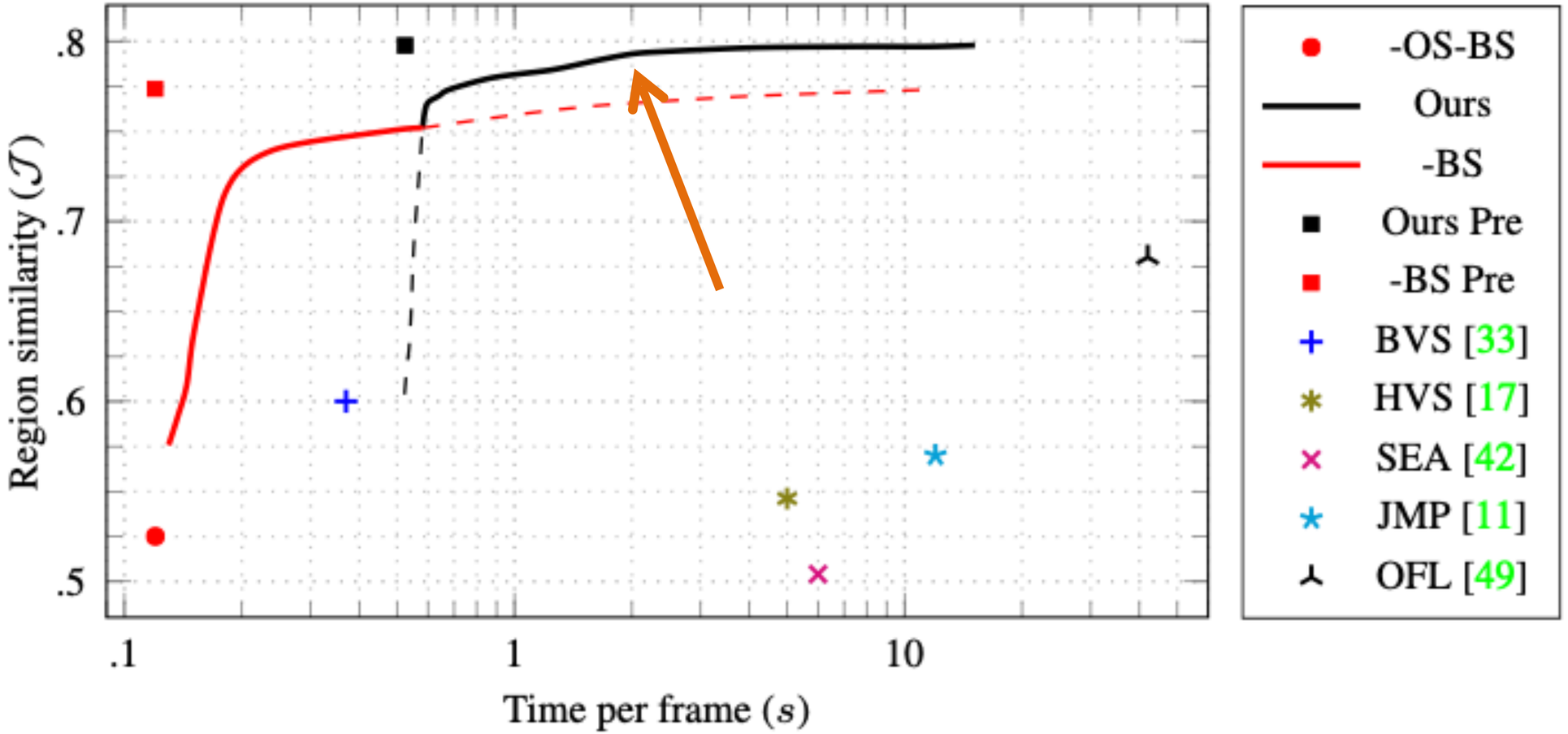


... to improve the accuracy:

S. Caelles et al. “One-shot video object segmentation”. CVPR 2017.

Runtime-accuracy trade-off



S. Caelles et al. “One-shot video object segmentation”. CVPR 2017.

Issues

- OSVOS has a static appearance model, hence it fails when
  - the background changes (ambiguity with the foreground)


SEEK
TRUTH


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

Issues

- OSVOS has a static appearance model, hence it fails when
  - the background changes (ambiguity with the foreground)
  - the object appearance changes significantly

Drifting problem

- The object appearance changes due to the changes in the object and camera pose:



Drifting problem

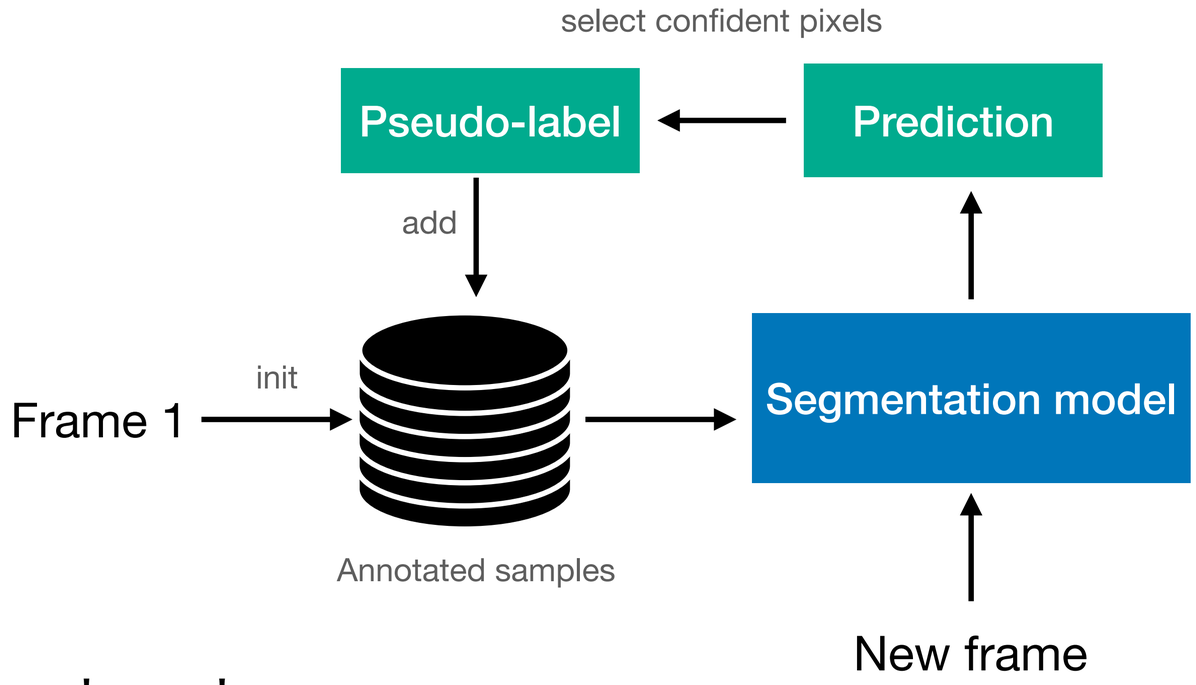
- The object appearance changes due to the changes in the object and camera pose:



One idea: adapt the model to the video using pseudo-labels

OnAVOS: Online Adaptation

- Online adaptation: adapt model to appearance changes in every frame, not just the first frame.



- Drawback: can be slow.

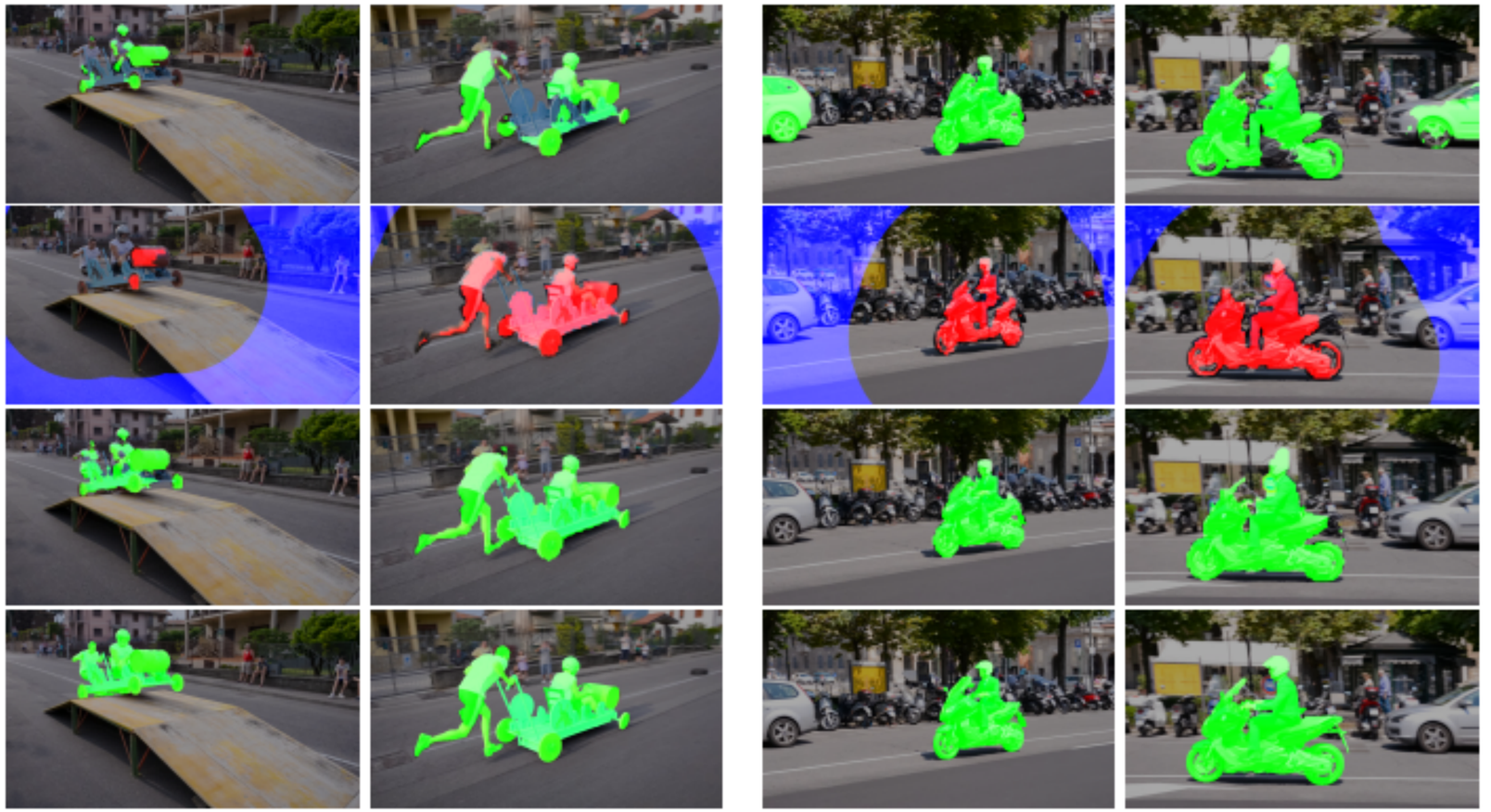
OnAVOS: Online Adaptation

Before adaptation

Pseudo labels

After adaptation

Ground truth



Blue = background samples
Red = foreground samples

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

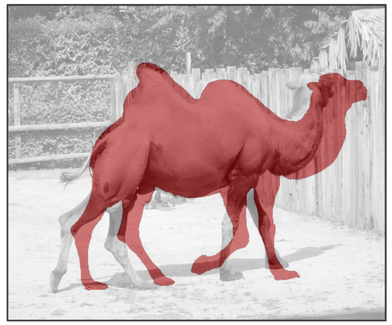
Issues

- OnAVOS is more accurate than One-Shot VOS;
- Instead of fine-tuning on a single sample, we fine-tune on a dynamic set of pseudo-labels;
- The pseudo-labels may be inaccurate, so their benefit is diminished over time.

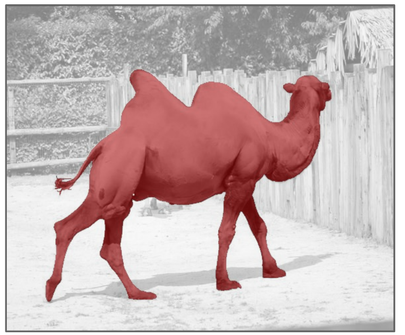
- Next: Can ensure we fine-tune the model with a correct signal?

MaskTrack

Input frame *t*



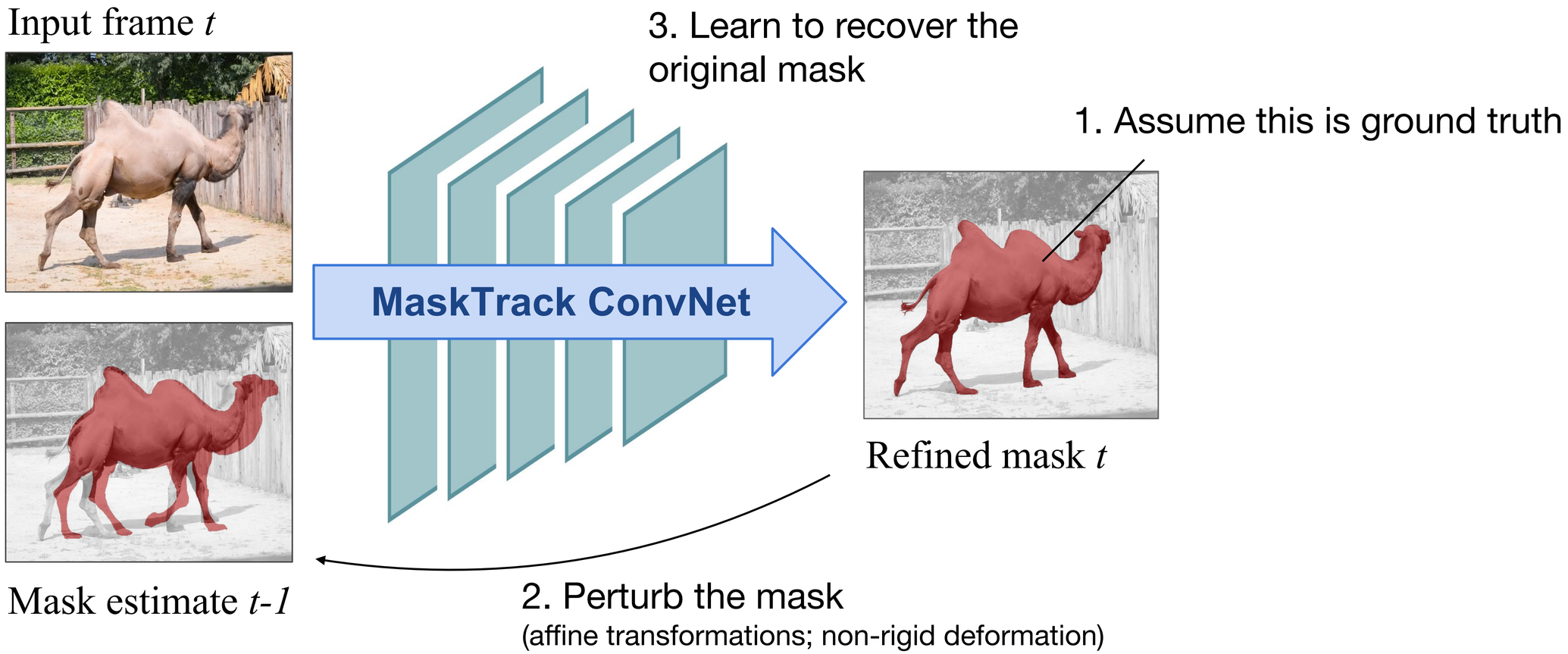
Mask estimate *t-1*



Refined mask *t*

Perazzi et al. „Learning Video Object Segmentation from Static Images“. CVPR 2017.

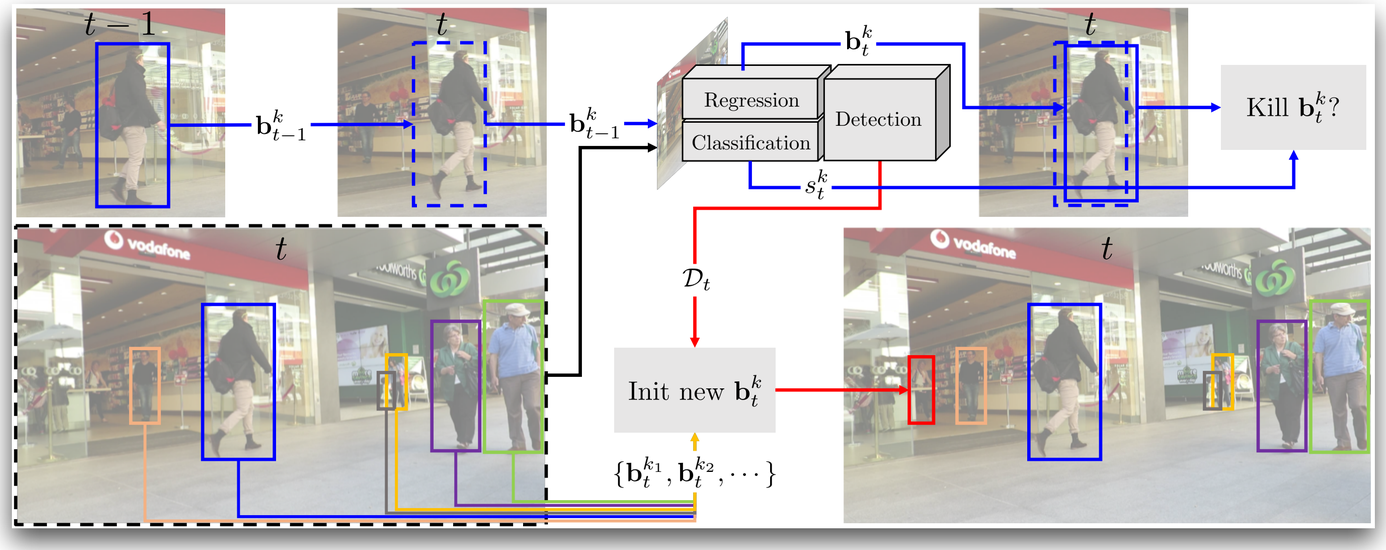
MaskTrack



Perazzi et al. „Learning Video Object Segmentation from Static Images“. CVPR 2017.

MaskTrack

- Training inputs can be simulated!
  - Like displacements to train the regressor of Faster R-CNN
  - Very similar in spirit to Tracktor (Lecture 4)



Perazzi et al. „Learning Video Object Segmentation from Static Images“. CVPR 2017.

MaskTrack

- Training inputs can be simulated!

(a) Annotated image

(b) Example training masks

Perazzi et al. „Learning Video Object Segmentation from Static Images“. CVPR 2017.

Summary

- Advantages of appearance-based models:
  - can be trained on static images;
  - can recover well from occlusions;
  - conceptually simple.

- Disadvantages:
  - no temporal consistency;
  - can be slow at test-time (need to adapt);

Proposal-based VOS

Proposal Generation

Until now:
- Input is the whole image
- Proposals could be used for refinement

Now:
- Generate proposals
- Link them temporally (similar to tracking-by-detection)

- Instance segmentation networks (e.g. Mask R-CNN) produce instance-level representation (masks with optional bounding boxes).
- One could link them over time to produce the tracking result.

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

J. Luiten et al. „PReMVOS: Proposal-generation, Refinement and Merging for Video Object Segmentation“. ACCV 2018.

PReMVOS

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

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

PReMVOS: Summary

- Very complex but a winner on multiple challenges.

- The techniques we have learned so far are complementary.
- We need simpler models → future research

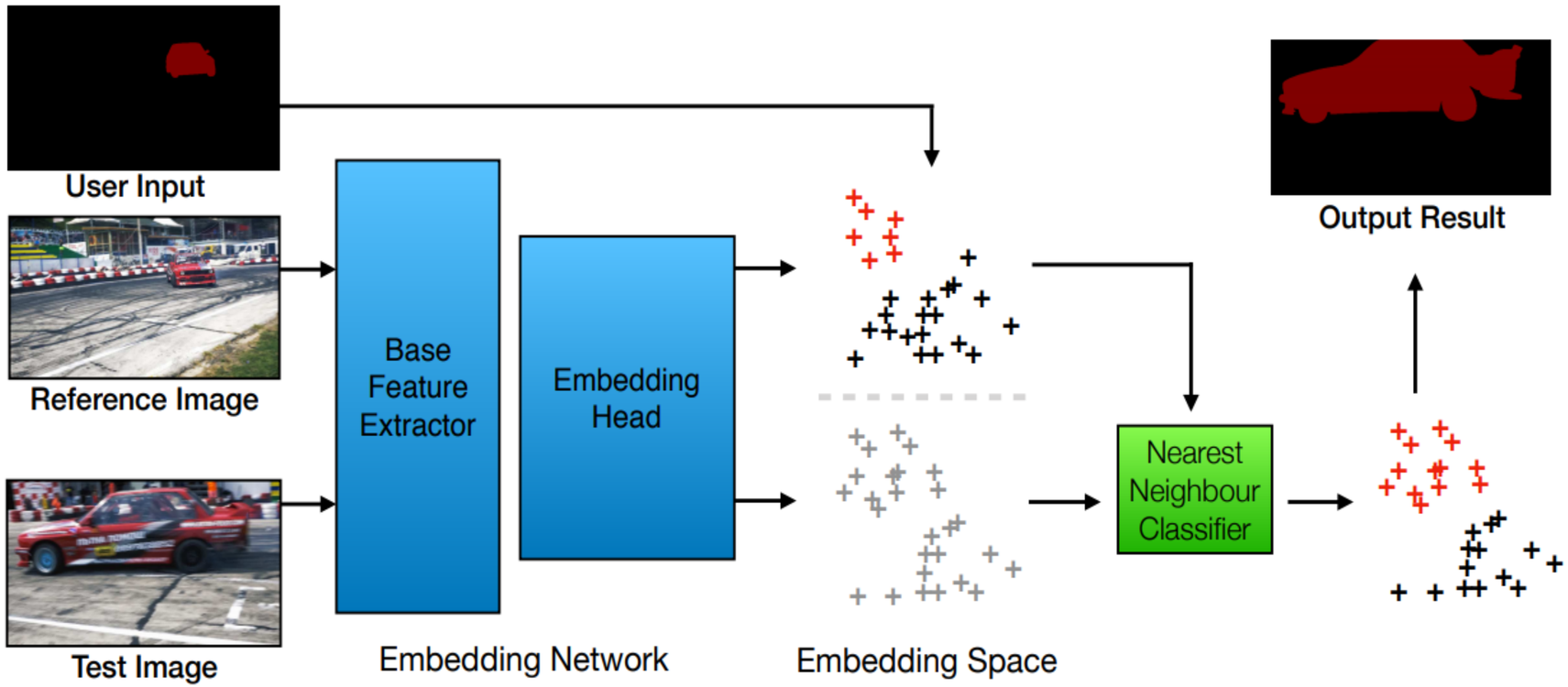
Metric-based approaches

Pixel-wise retrieval

- Idea: Learn a pixel-level embedding space where proximity between two feature vectors is semantically meaningful

Pixel-wise retrieval

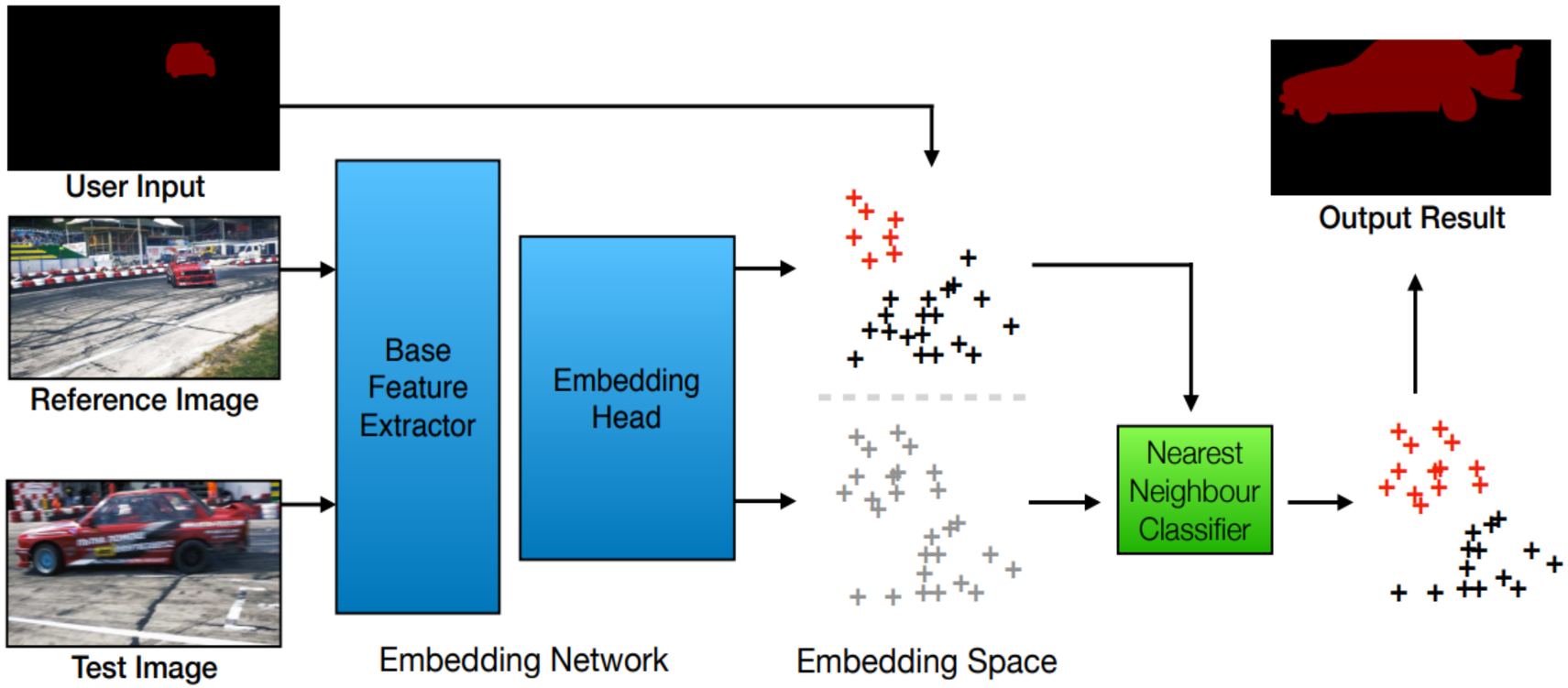
- The user input can be in any form, first-frame ground-truth mask, scribble...



Y. Chen et al. „Blazingly Fast Video Object Segmentation with Pixel-Wise Metric Learning“. CVPR 2018.

Pixel-wise retrieval

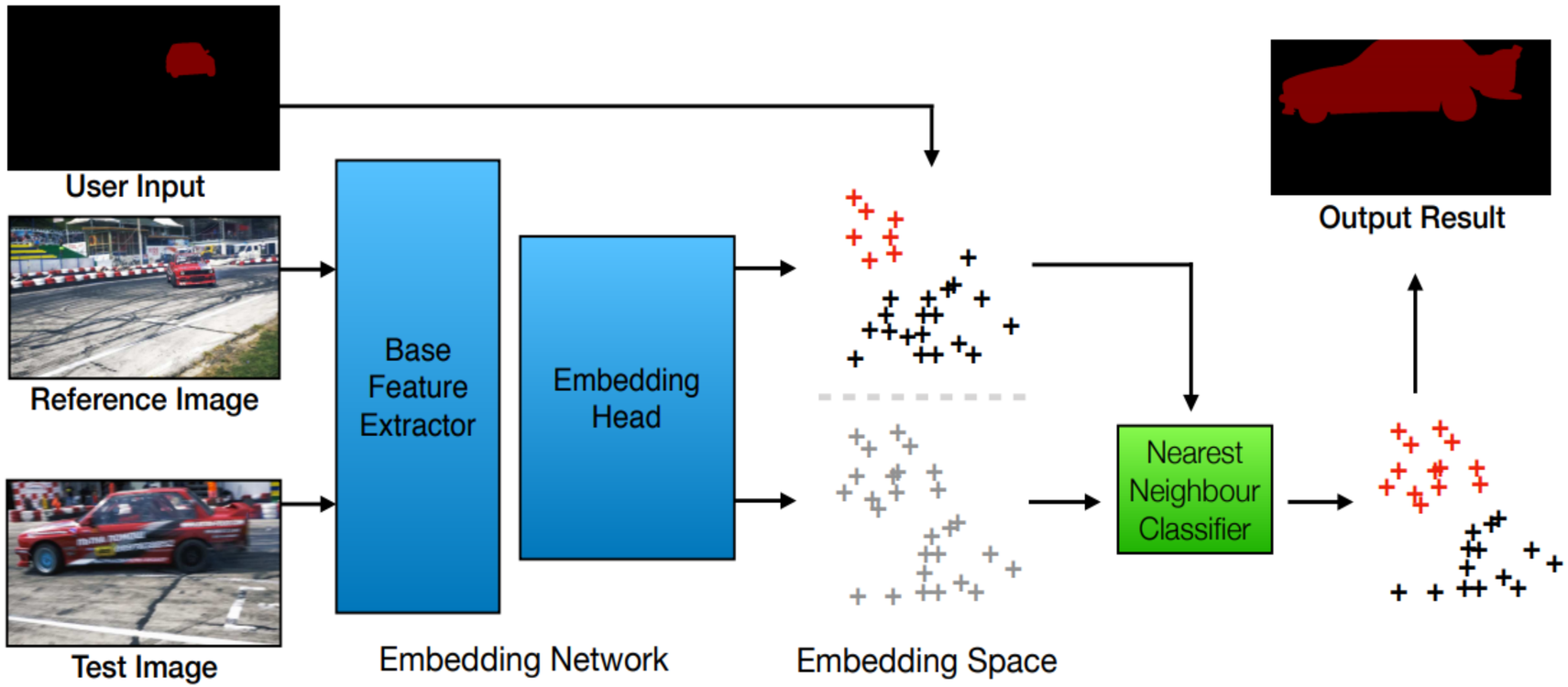
- Training: use the triplet loss to bring foreground pixels together and separate them from background pixels



Y. Chen et al. „Blazingly Fast Video Object Segmentation with Pixel-Wise Metric Learning“. CVPR 2018.

Pixel-wise retrieval

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



Y. Chen et al. „Blazingly Fast Video Object Segmentation with Pixel-Wise Metric Learning“. CVPR 2018.

Pixel-wise retrieval: Summary

Advantages:

- We do not need to retrain the model for each sequence, nor fine-tune;
- We can use unsupervised training to learn a useful feature representation

Y. Chen et al. „Blazingly Fast Video Object Segmentation with Pixel-Wise Metric Learning“. CVPR 2018.

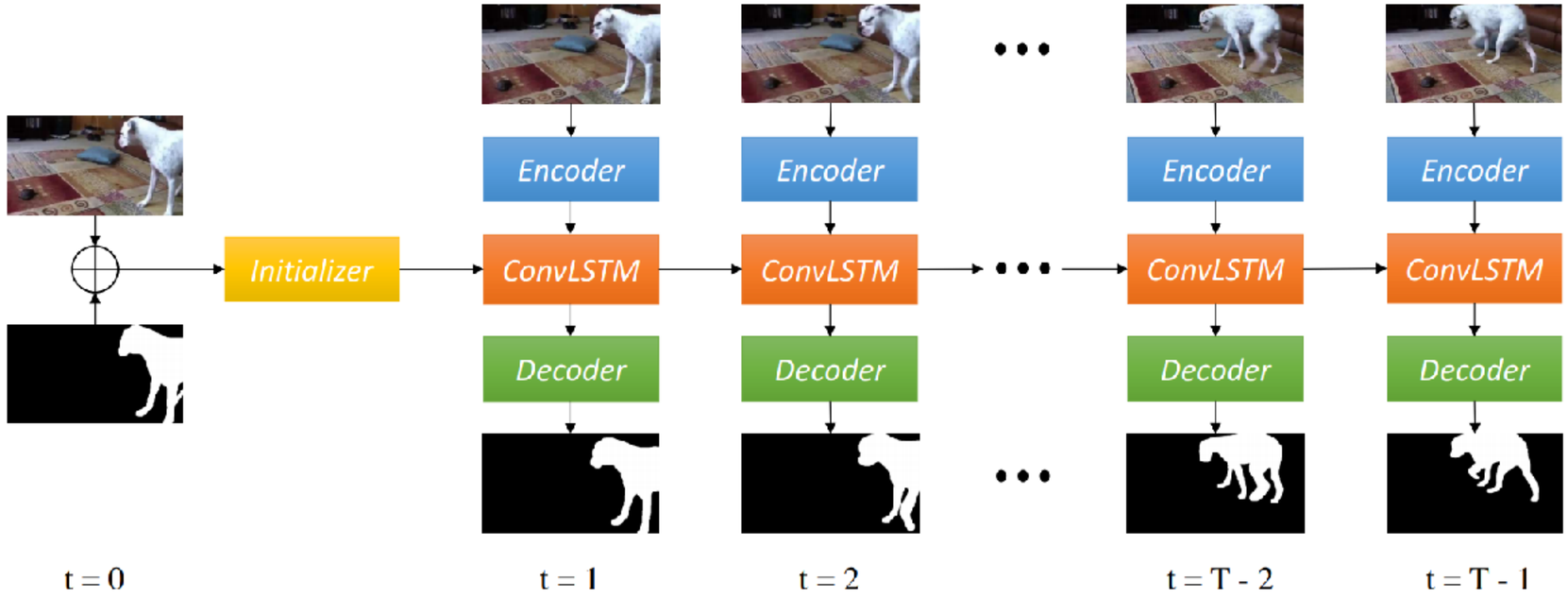
We are dealing with video

- A video is sequence of images.
- What category of models are specifically designed for sequences?

- Recurrent Neural Networks!

Temporal LSTM

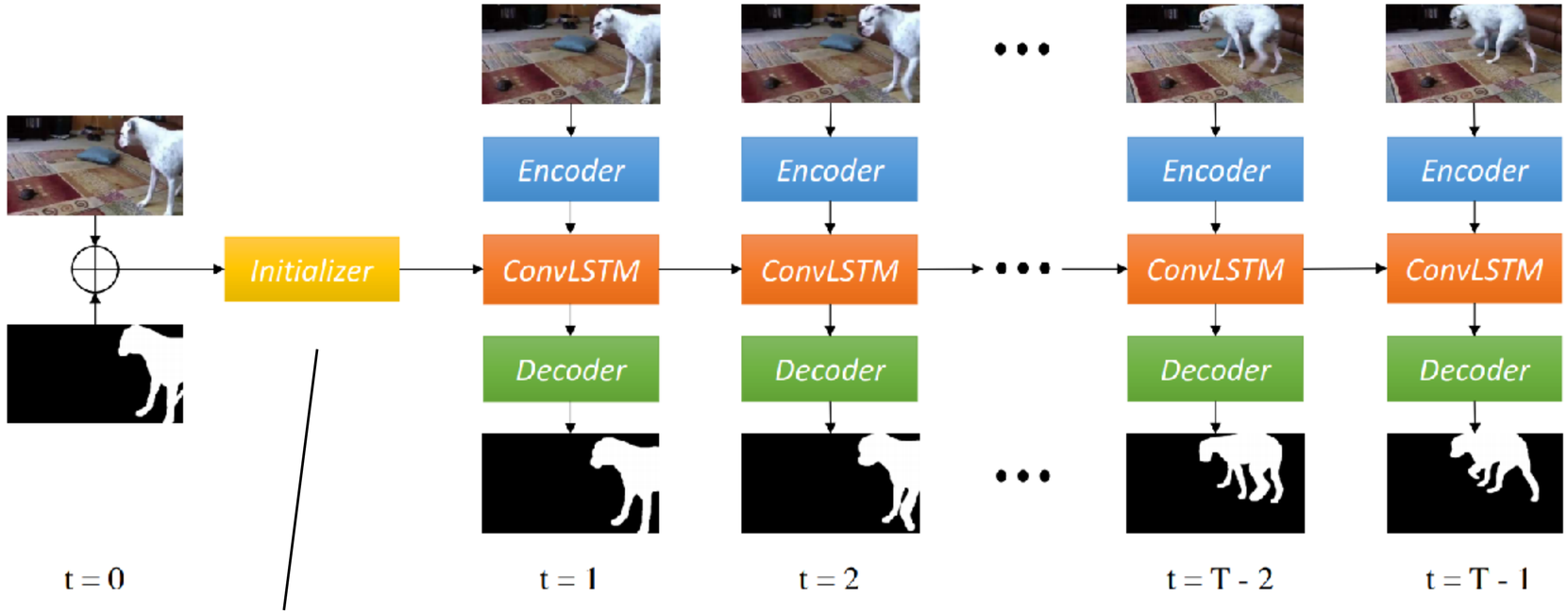
- One-shot video object segmentation
- If we have multiple objects, each of them is predicted independently



N. Xu et al. „YouTube-VOS: Sequence-to-sequence video object segmentation." ECCV 2018.

Temporal LSTM

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



initial mask is embedded into the initial state of the RNN

N. Xu et al. „YouTube-VOS: Sequence-to-sequence video object segmentation." ECCV 2018.

Temporal LSTM: Summary

- Advantage: Improved temporal coherence.

Disadvantages:

- Typically more computationally greedy.
- We have to limit the sequence length to contain the computational requirement.
- May struggle with large motions.
- Error accumulation.

Summary for today

- Motion-based models:
  - FlowNet; SegFlow;
  - Recurrent architectures (Seq2seq).
- Appearance-based models:
  - OSVOS: First-frame fine-tuning;
  - OnAVOS: Online Adaptation;
  - MaskTrack: Mask Refinement;
  - ReID-VOS: Object Appearance Re-Identification;
  - PReMVOS.

VOS vs MOTS

- “Semi-supervised” 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

P. Voigtlaender et al. „MOTS: Multi-Object Tracking and Segmentation“. CVPR 2019

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.

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

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

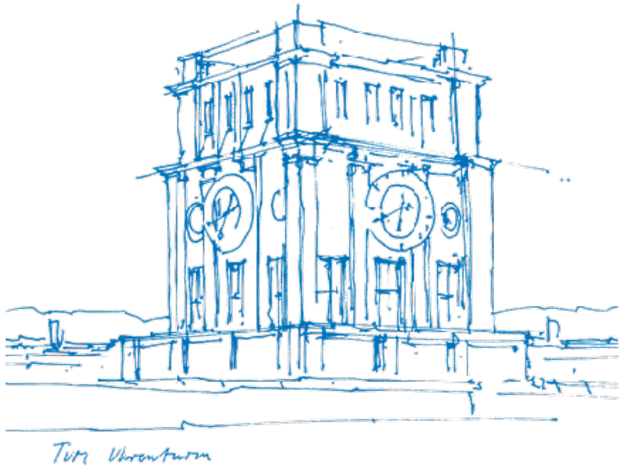
Metrics for VOS

- Region similarity: Jaccard index (IoU) of ground truth mask and predicted mask.
  - Mean: average for the dataset
  - Recall: fraction of sequences scoring higher than a threshold (e.g. 0.5 in DAVIS)
  - Decay: quantifies the performance loss (or gain) over time.
    - Can be F-Decay (contour), J-Decay (region);
    - In DAVIS: mean(25% first frames) - mean(25% last frames).

Computer Vision III:
Video object segmentation

Nikita Araslanov
10.01.2023

Content credit:
Prof. Laura Leal-Taixé
https://dvl.in.tum.de



Acknowledgments

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