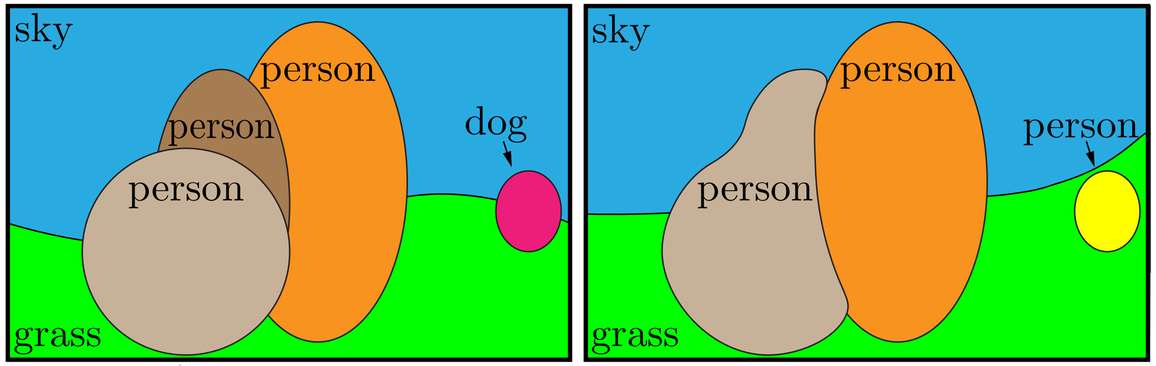
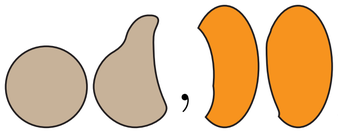
Evaluating panoptic segmentation

Panoptic quality (PQ)

- Example:



Ground truth Prediction

Person — TP: {  }; FN: {  }; FP: {  }

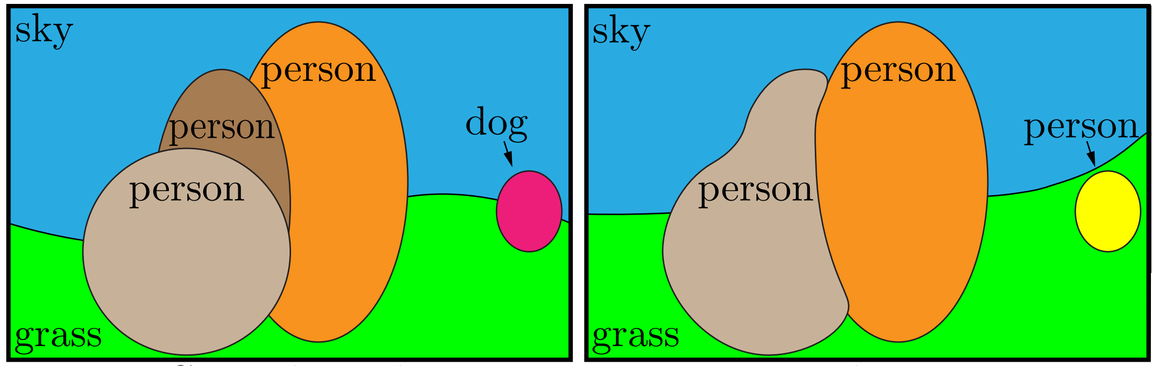
TP = True positive, FN = False negative, FP = false positive

- Wait, but don't we need to define an IoU threshold?

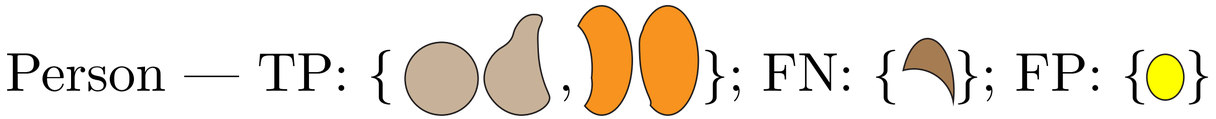
Kirillov et al., "Panoptic Segmentation". CVPR 2019.

Panoptic quality (PQ)

- To compute PQ we specify that a prediction and a ground truth match only if their IoU is greater than 0.5.
- This match, if found, is **unique**.
- Unique matching theorem:
  - A ground-truth segment has an IoU greater than 0.5 with at most **one** prediction.



Ground truth Prediction



proof hint

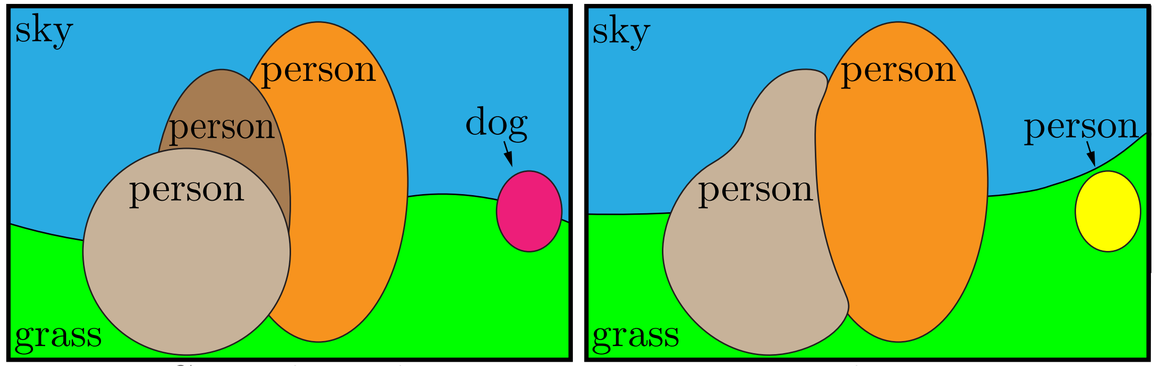
- With p_1 and p_2 being disjoint:

$$\mathrm{IoU}(p_1, g) + \mathrm{IoU}(p_2, g) \leq 1$$

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

1. Establish matches between the ground-truth and predictions;
2. Count TPs, FPs and FNs;
3. Compute PQ for each class:



Ground truth Prediction

Person — TP: { , }; FN: { }; FP: { }

$$PQ = \frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\ \mathrm{TP}\ |} \frac{|\ \mathrm{TP}\ |}{|\ \mathrm{TP}\ | + \frac{1}{2}|\ \mathrm{FP}\ | + \frac{1}{2}|\ \mathrm{FN}\ |}$$

...and then average.

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

$$PQ = \frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\ \mathrm{TP}\ |} \frac{|\ \mathrm{TP}\ |}{|\ \mathrm{TP}\ | + \frac{1}{2}|\ \mathrm{FP}\ | + \frac{1}{2}|\ \mathrm{FN}\ |}$$

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

$$PQ = \underbrace{\frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\ \mathrm{TP}\ |}}_{\mathrm{SQ}} \frac{|\ \mathrm{TP}\ |}{|\ \mathrm{TP}\ | + \frac{1}{2}|\ \mathrm{FP}\ | + \frac{1}{2}|\ \mathrm{FN}\ |}$$

- SQ = “Segmentation Quality”:
  - Average mask IoU for true positives;
  - Measures pixel-level accuracy of predicted masks.
- Looks familiar (QUIZ)?

Recall multi-object tracking precision (MOTP):

$$\mathrm{MOTP} = \frac{\sum_{t,i} \mathrm{IoU}_{t,i}}{\sum_t \mathrm{TP}}$$

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

$$PQ = \frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\ \mathrm{TP}\ |} \boxed{\frac{|\ \mathrm{TP}\ |}{|\ \mathrm{TP}\ | + \frac{1}{2}|\ \mathrm{FP}\ | + \frac{1}{2}|\ \mathrm{FN}\ |}}$$

RQ

- RQ = “Recognition Quality”:
  - Object-level accuracy.
  - Does it look familiar? (QUIZ)

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

$$PQ = \frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\ \mathrm{TP}\ |} \boxed{\frac{|\ \mathrm{TP}\ |}{|\ \mathrm{TP}\ | + \frac{1}{2}|\ \mathrm{FP}\ | + \frac{1}{2}|\ \mathrm{FN}\ |}}$$

RQ

- RQ = “Recognition Quality”:
  - Object-level accuracy.
  - Does it look familiar? This is F-score (F_1)
  - F-score is the harmonic mean of precision and recall.

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

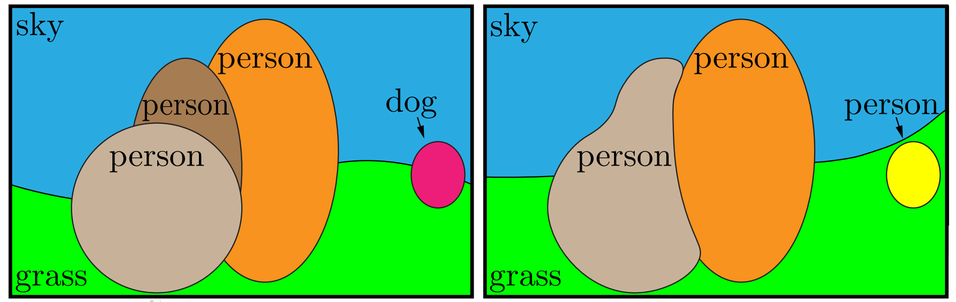
$$PQ = \frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\ \mathrm{TP}\ |} \frac{|\ \mathrm{TP}\ |}{|\ \mathrm{TP}\ | + \frac{1}{2}|\ \mathrm{FP}\ | + \frac{1}{2}|\ \mathrm{FN}\ |}$$

- Observation 1: $PQ, RQ, SQ \in [0, 1]$

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

$$PQ = \frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\ \mathrm{TP}\ |} \frac{|\ \mathrm{TP}\ |}{|\ \mathrm{TP}\ | + \frac{1}{2}|\ \mathrm{FP}\ | + \frac{1}{2}|\ \mathrm{FN}\ |}$$



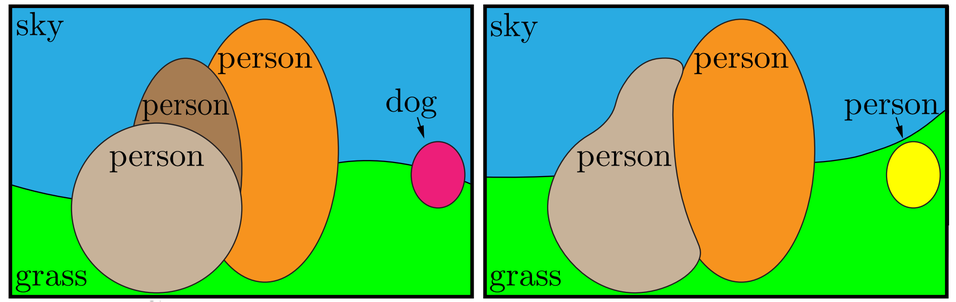
Ground truth Prediction

- Observation 2: What effect does missing one object have on PQ (e.g. “dog” above)?
  - Increment FN for that class (e.g. “dog”) AND FP for another class (e.g. “person”).

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

$$PQ = \frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\ \mathrm{TP}\ |} \frac{|\ \mathrm{TP}\ |}{|\ \mathrm{TP}\ | + \frac{1}{2}|\ \mathrm{FP}\ | + \frac{1}{2}|\ \mathrm{FN}\ |}$$



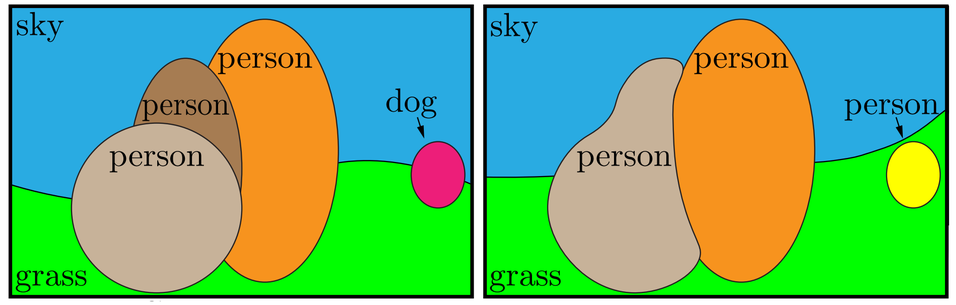
Ground truth Prediction

- Observation 2: What effect does missing one object have on PQ (e.g. “dog” above)?
  - Increment FN for that class (e.g. “dog”) AND FP for another class (e.g. “person”).
  - This reduces PQ of **two** classes.

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Panoptic quality (PQ)

$$PQ = \frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\ \mathrm{TP}\ |} \frac{|\ \mathrm{TP}\ |}{|\ \mathrm{TP}\ | + \frac{1}{2}|\ \mathrm{FP}\ | + \frac{1}{2}|\ \mathrm{FN}\ |}$$



Ground truth Prediction

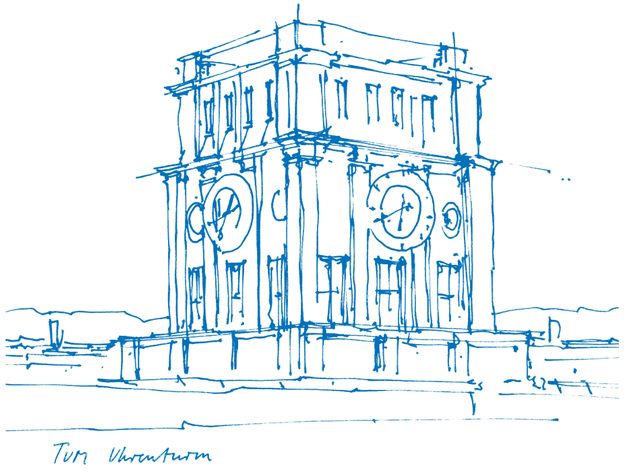
- Observation 2: What effect does missing one object have on PQ (e.g. “dog” above)?
  - Increment FN for that class (e.g. “dog”) AND FP for another class (e.g. “person”).
  - This reduces PQ of **two** classes.
  - Common trick: Predict as “unknown” class instead. FP count will not affect another class.

Kirillov et al., “Panoptic Segmentation”. CVPR 2019.

Computer Vision III:
Video object segmentation

Nikita Araslanov
12.12.2023

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



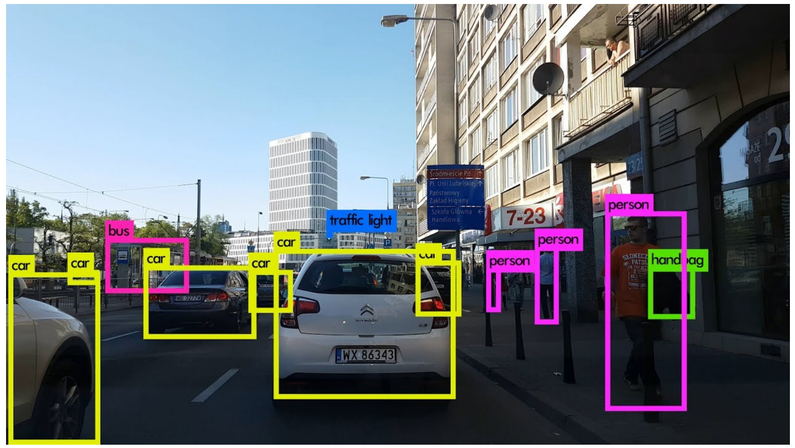
Course progress

1. Introduction
2. Object detection 1
3. Object detection 2
4. Single object tracking
5. Multiple object tracking
6. Semantic segmentation
7. Instance & panoptic segmentation
8. **Video object segmentation** ← We are here
9. Transformers

Christmas break

10. Semi-supervised DST
11. Unsupervised DST

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

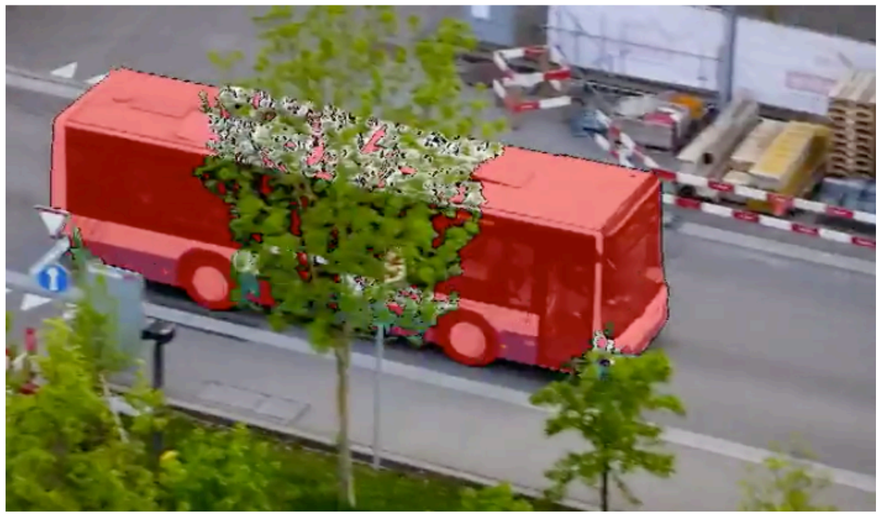
VOS: Some challenges

- 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

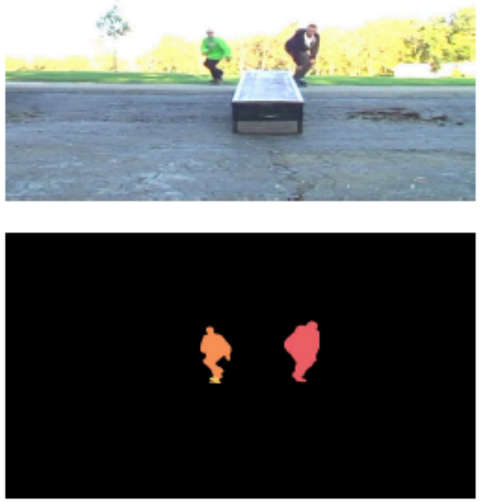
Motion model

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.

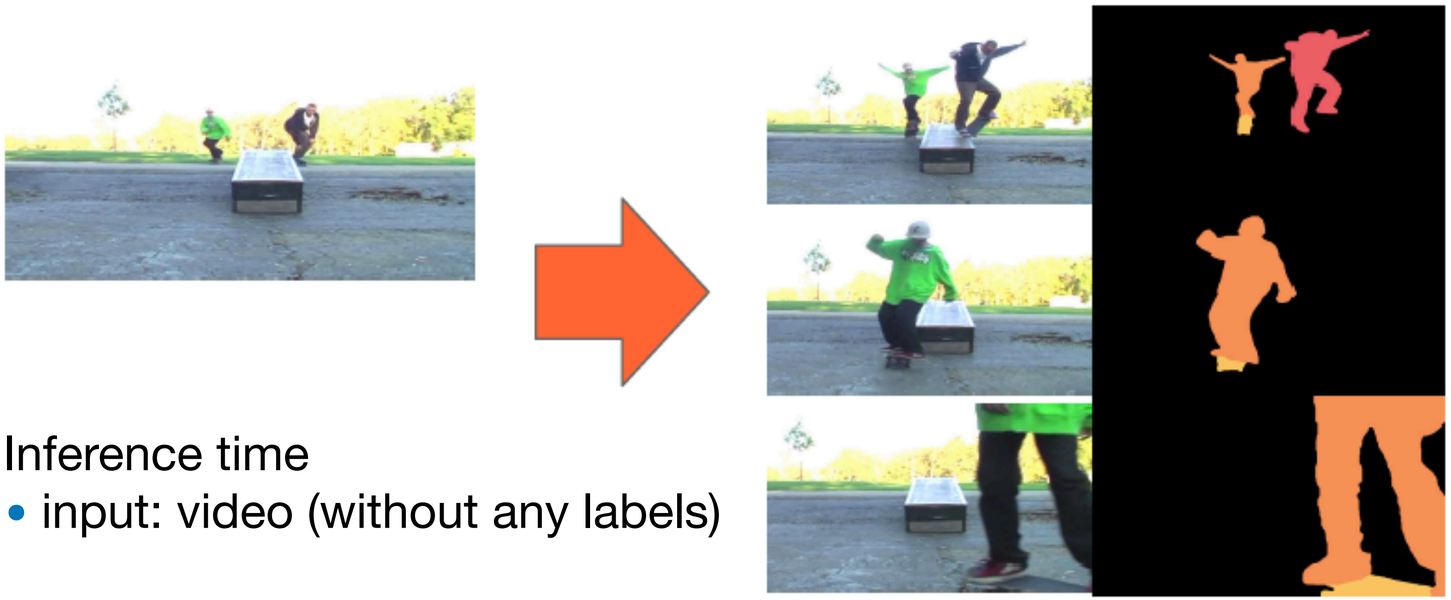
VOS: Tasks

One-shot (“semi-supervised”) VOS



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

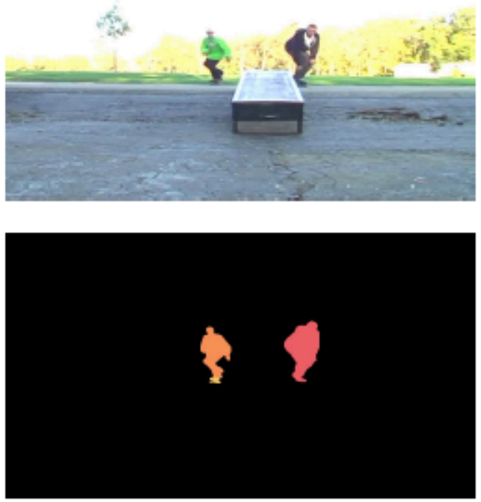
Zero-shot (“unsupervised”) VOS



Inference time
- input: video (without any labels)

VOS: Tasks

One-shot (“semi-supervised”) VOS



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

Zero-shot (“unsupervised”) VOS

Inference time
- input: video (without any labels)

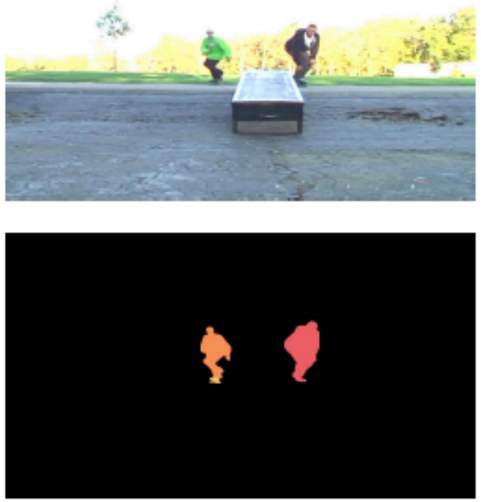
What to track?

- 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

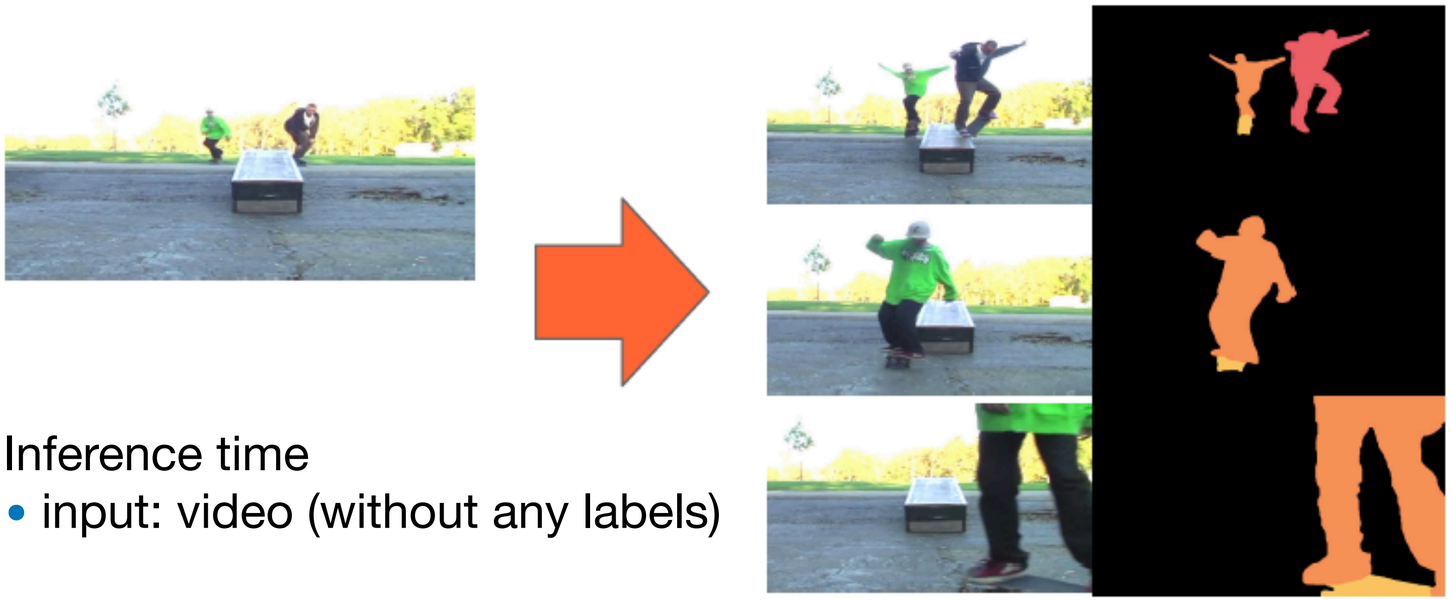
“Semi-supervised” (one-shot) VOS



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

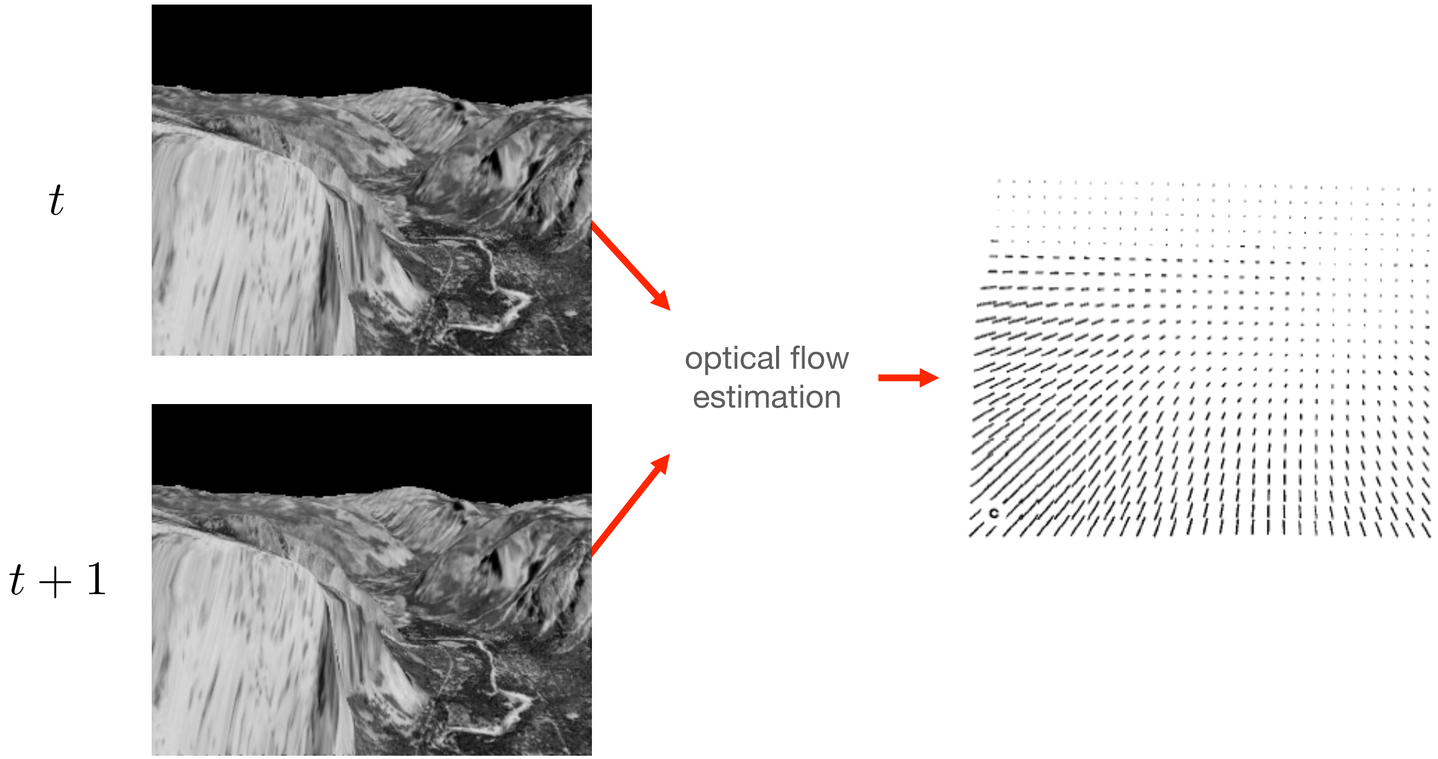
Goal: Complete video segmentation

- 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

Motion-based VOS

Optical flow

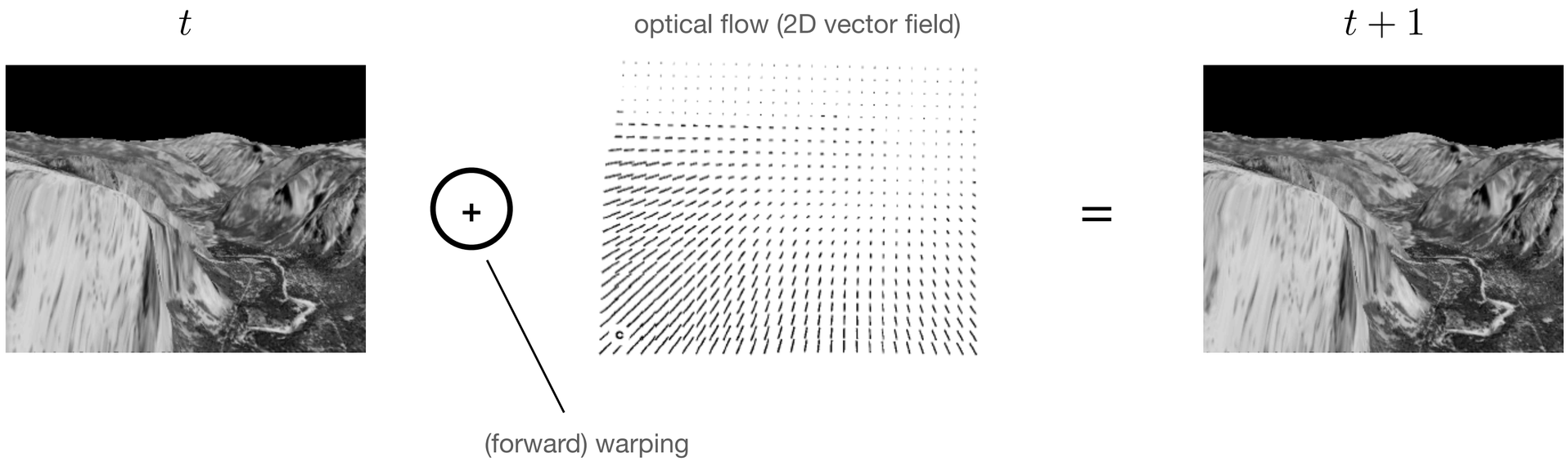
- Recall optical flow (from Computer Vision I):
  - a pattern of apparent motion (Lukas and Kanade, ’81; Horn & Schunk ’81);



[S. Roth; M. Black]

Optical flow

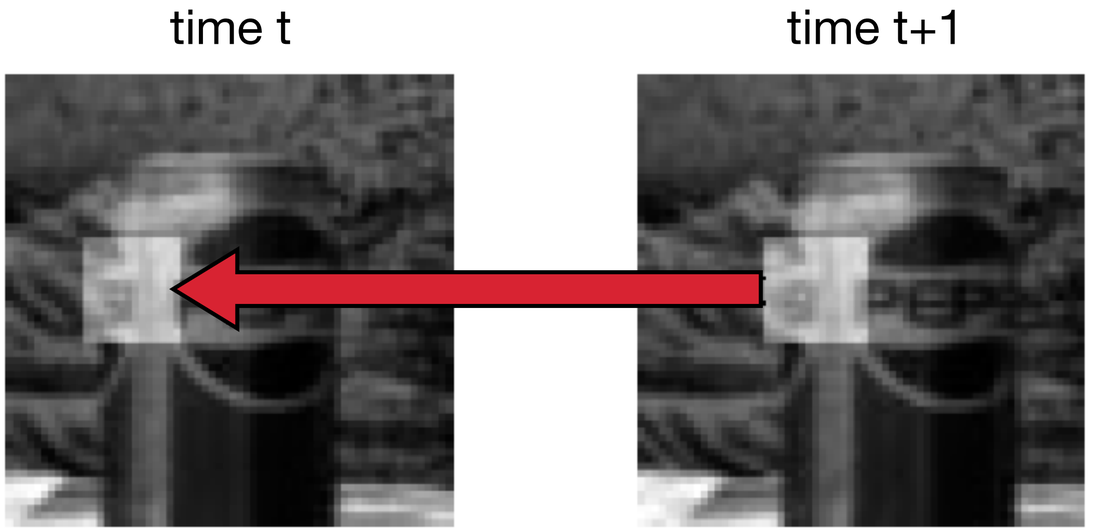
- Recall optical flow (from Computer Vision I):
  - a pattern of apparent motion (Lukas and Kanade, ’81; Horn & Schunk ’81);



[S. Roth; M. Black]

Brightness constancy

- Assumption 1:
  - Image measurements (brightness) in a small region remain the same



$$I(x + u(x, y), y + v(x, y), t + 1) = I(x, y, t)$$

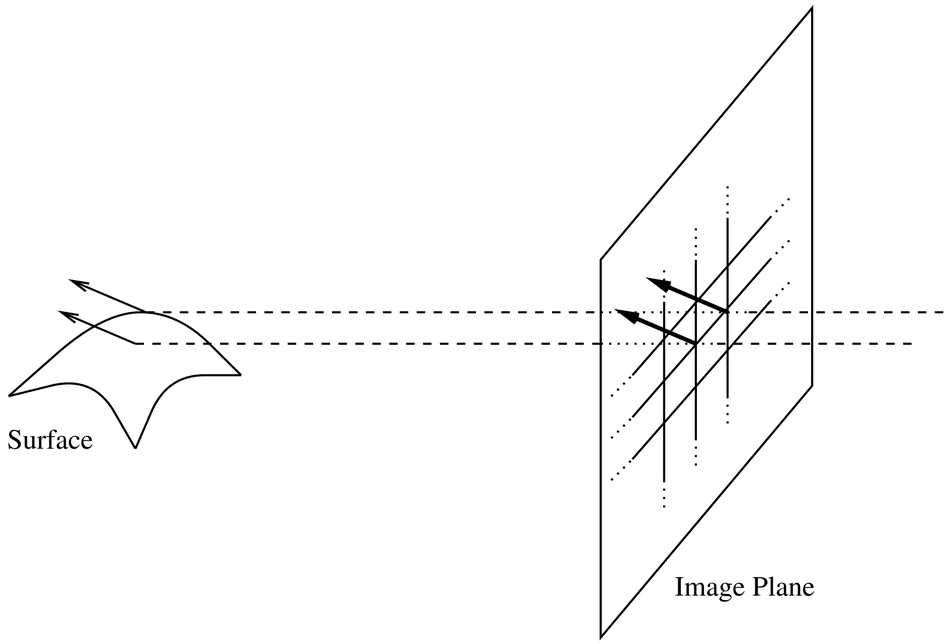
horizontal motion

vertical motion

[S. Roth; M. Black]

Spatial coherence

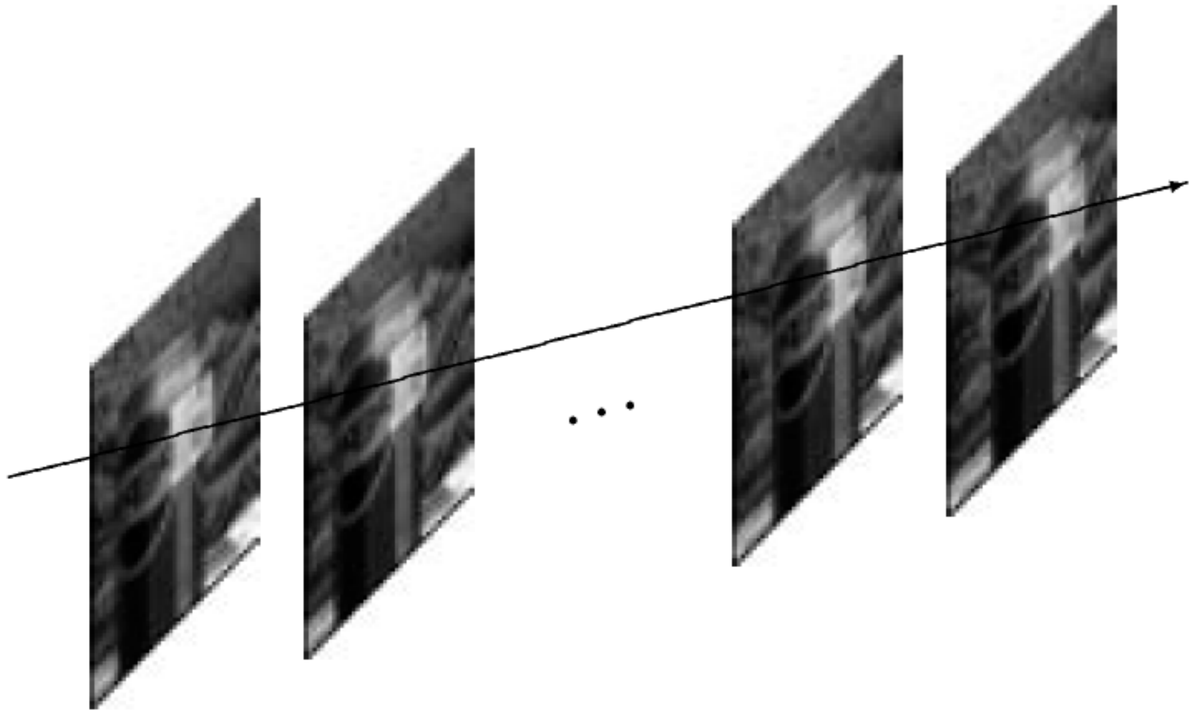
- Assumption 2:
  - Neighbouring points in the scene typically belong to the same surface and hence typically have similar 3D motions.



[S. Roth; M. Black]

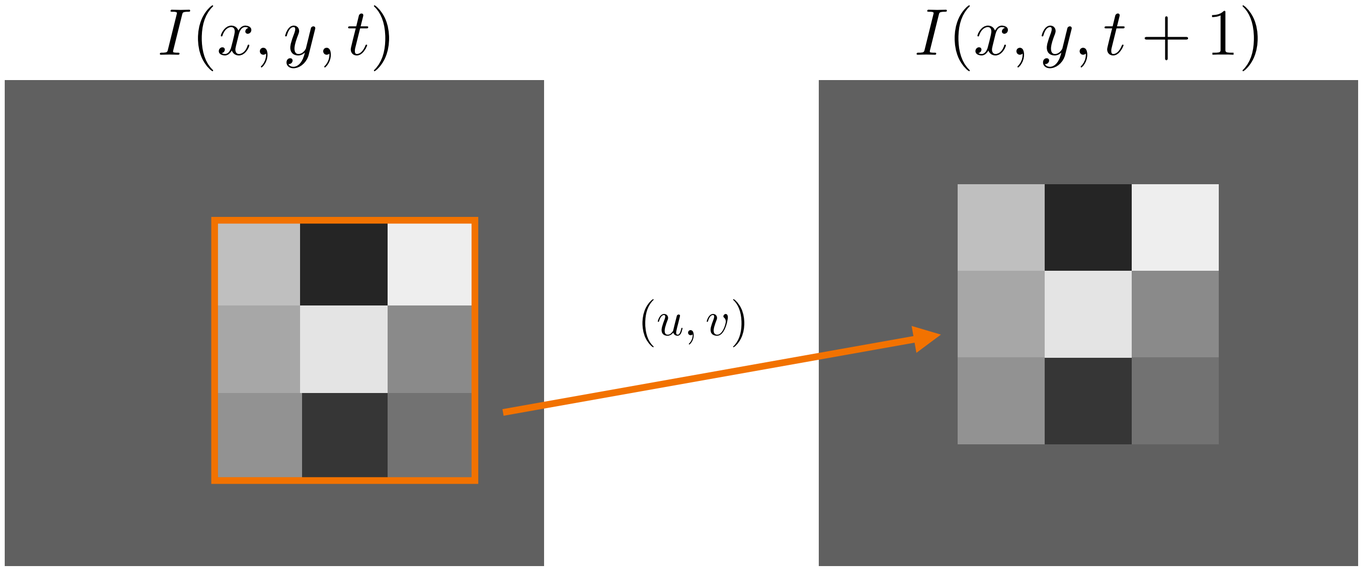
Temporal persistence

- Assumption 3:
  - The image motion of a surface patch changes gradually over time.



[S. Roth; M. Black]

Minimise brightness difference



$$E_{\mathrm{SSD}}(u, v) = \sum_{(x,y)\in R} \big(I(x + u, y + v, t + 1) - I(x, y, t)\big)^2$$

Quiz: What assumptions did we incorporate here?

[S. Roth; M. Black]

Minimise brightness difference

- Goal: find (u, v) minimising

$$E_{\mathrm{SSD}}(u, v) = \sum_{(x,y)\in R} \big(I(x+u, y+v, t+1) - I(x, y, t)\big)^2$$

- First-order Taylor approximation of $I(x+\Delta_x, y+\Delta_y, t+\Delta_t)$:

$$I(x, y, t) + \Delta_x \frac{\partial}{\partial x} I(x, y, t) + \Delta_y \frac{\partial}{\partial x} I(x, y, t) + \Delta_t \frac{\partial}{\partial t} I(x, y, t) + \epsilon(\Delta_x^2, \Delta_y^2, \Delta_t^2)$$

Approximation error

Minimise brightness difference

- Plugging in our approximation into E_{SSD}, we get:

$$E_{\mathrm{SSD}} \approx \sum_{(x,y)\in R} \big(u \cdot I_x(x,y,t) + v \cdot I_y(x,y,t) + I_t(x,y,t)\big)^2$$

- Differentiate w.r.t. u and v:

$$\frac{\partial E_{\mathrm{SSD}}}{\partial u} \approx 2 \sum_R (u \cdot I_x + v \cdot I_y + I_t) I_x = 0$$

$$\frac{\partial E_{\mathrm{SSD}}}{\partial v} \approx 2 \sum_R (u \cdot I_x + v \cdot I_y + I_t) I_y = 0$$

Minimise brightness difference

- By rearranging the terms, we get a system of 2 equations with 2 unknowns:

$$\begin{bmatrix} \sum_R I_x^2 & \sum_R I_x I_y \\ \sum_R I_x I_y & \sum_R I_y^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} -\sum_R I_x I_t \\ -\sum_R I_y I_t \end{bmatrix}$$

structure tensor

- Structure tensor is positive definite, hence invertible, so

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_R I_x^2 & \sum_R I_x I_y \\ \sum_R I_x I_y & \sum_R I_y^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_R I_x I_t \\ -\sum_R I_y I_t \end{bmatrix}$$

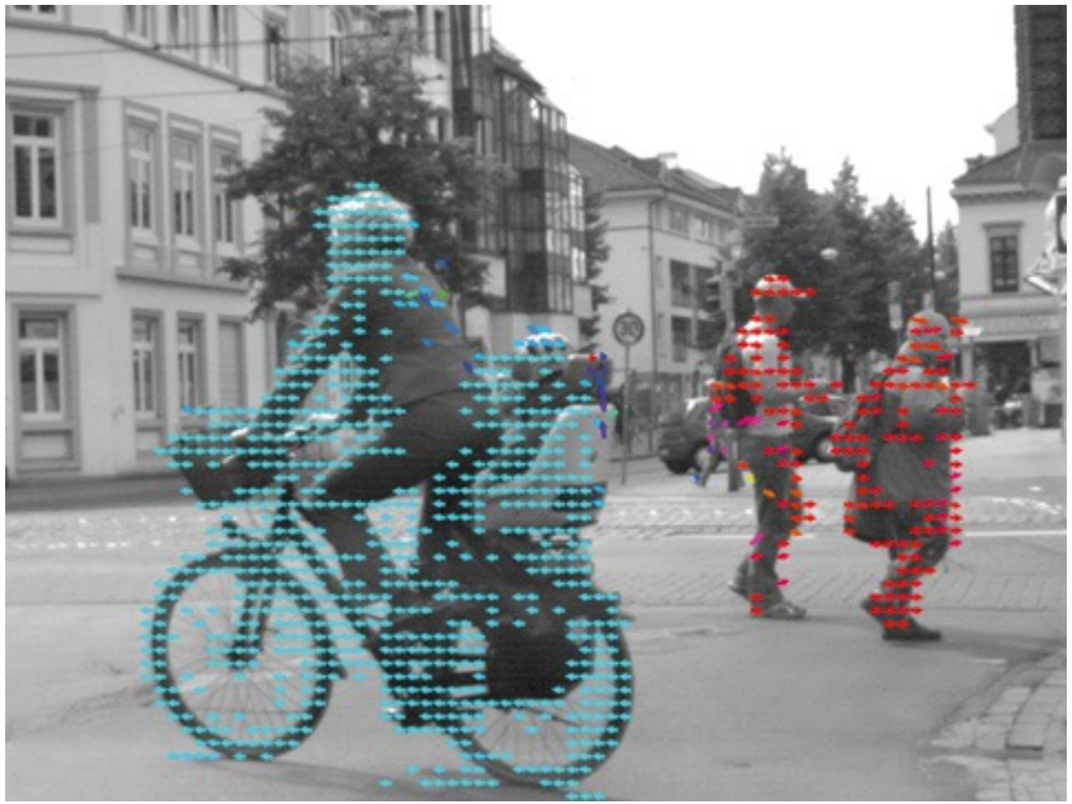
Lucas-Kanade

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_R I_x^2 & \sum_R I_x I_y \\ \sum_R I_x I_y & \sum_R I_y^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_R I_x I_t \\ -\sum_R I_y I_t \end{bmatrix}$$

- This is a classical flow technique – Lucas-Kanade method:
  - B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. IJCAI, pp. 674–679, 1981.
- Many extensions (e.g. Horn–Schunck introduces global smoothness).

Motion segmentation

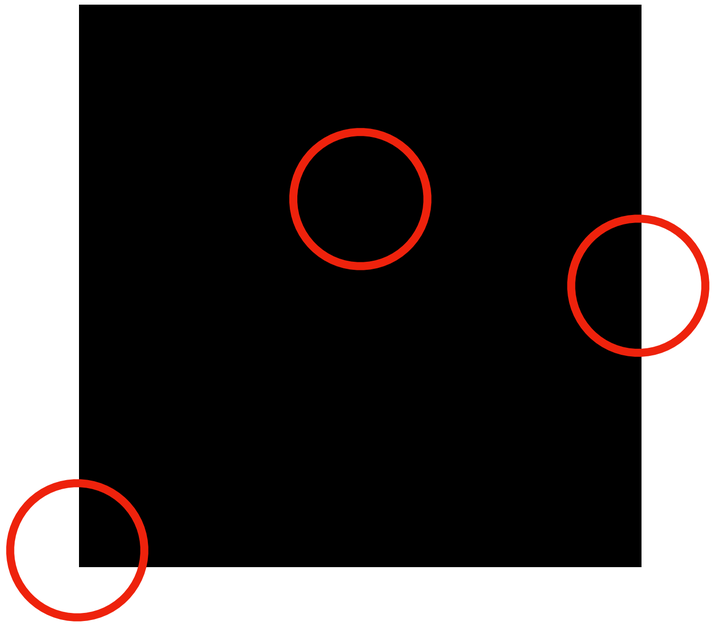
- We can segment some objects based on their motion:



Optical flow

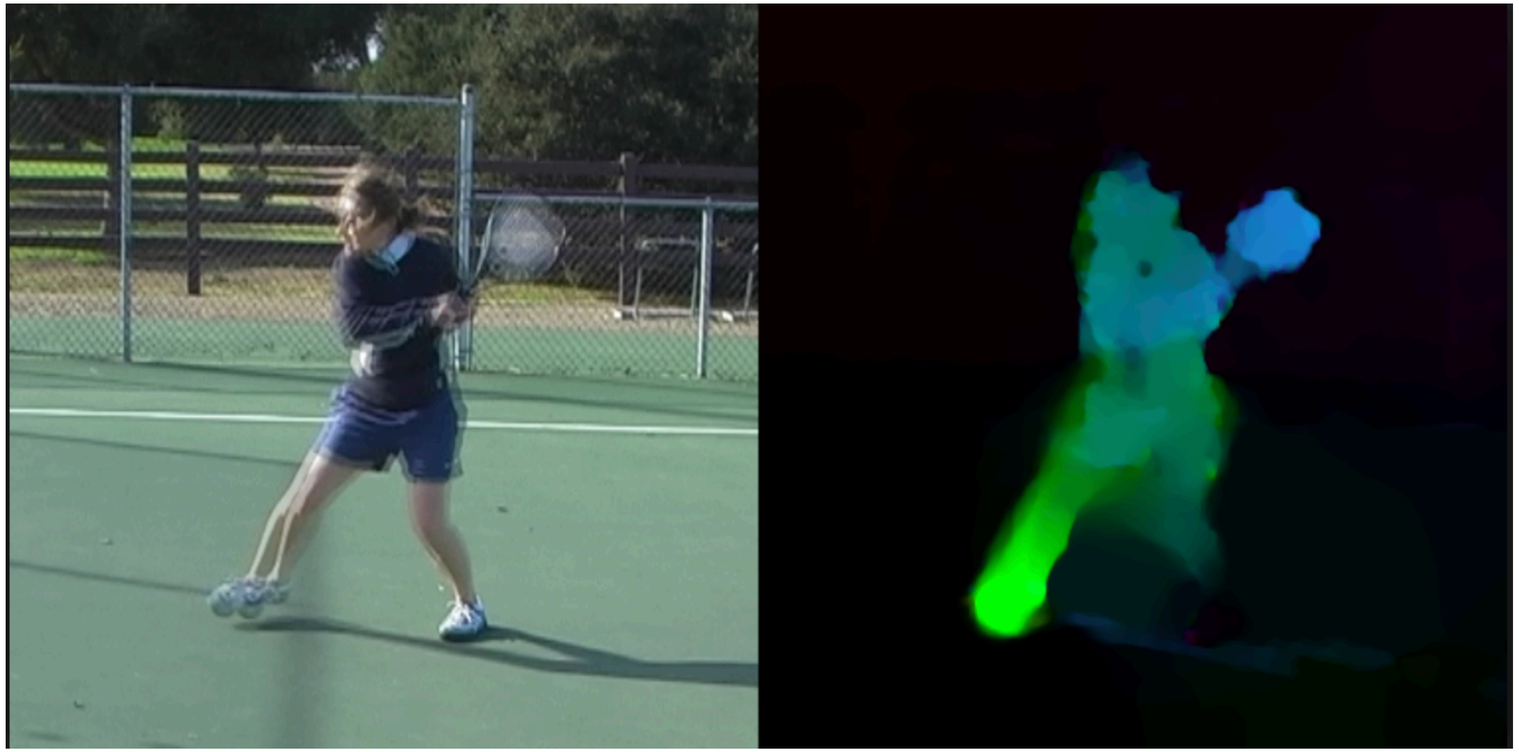
- “perceived” 2D motion, not the real motion of the object.
- the aperture problem.

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



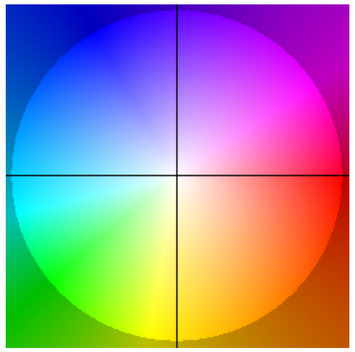
Optical flow

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



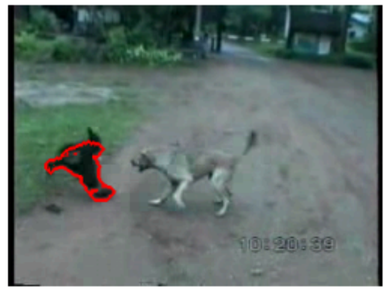
[Brox and Malik; 2010]

Optical flow colour wheel



Can we do VOS with optical flow?

- Joint formulation of segmentation and optical flow estimation:



initial optical flow

update the object mask

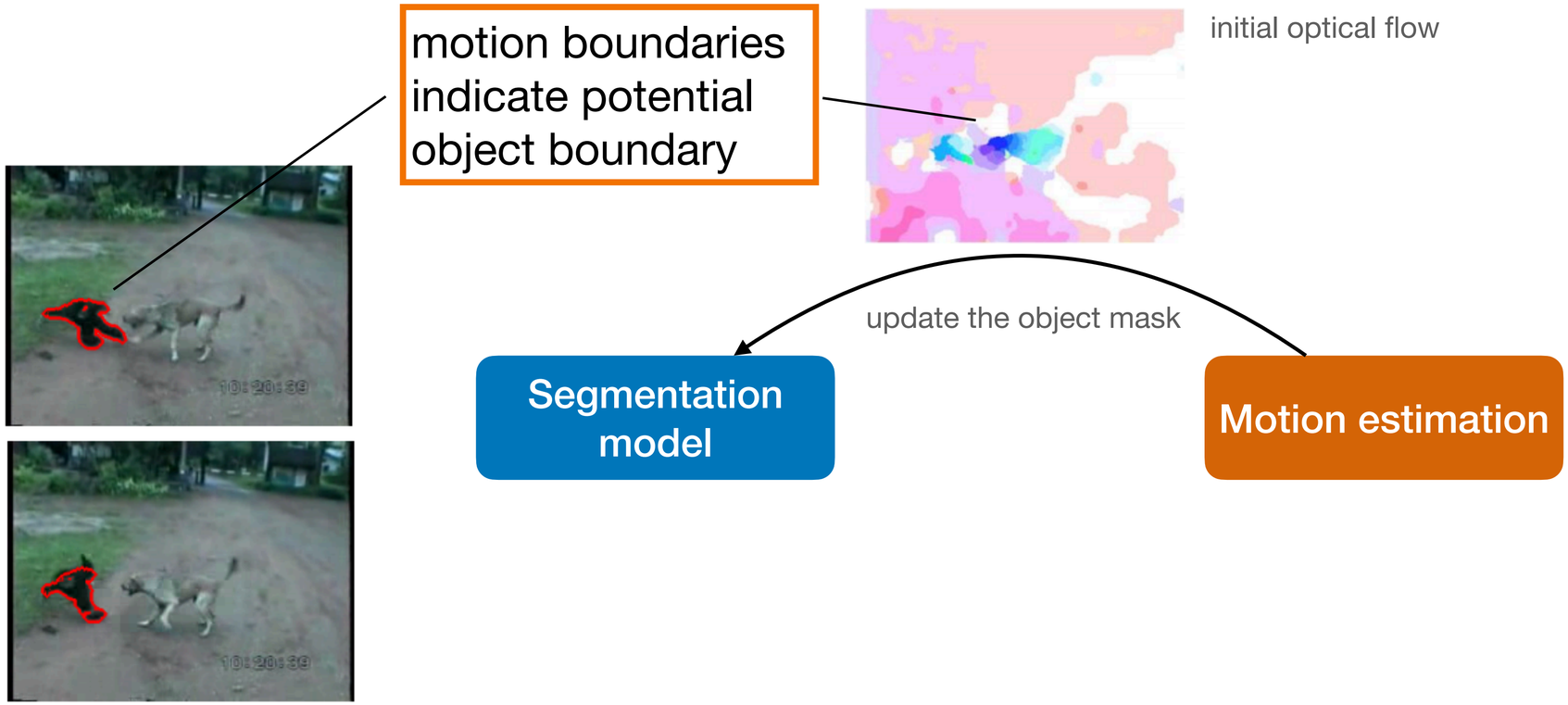
Segmentation model

Motion estimation

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

Can we do VOS with optical flow?

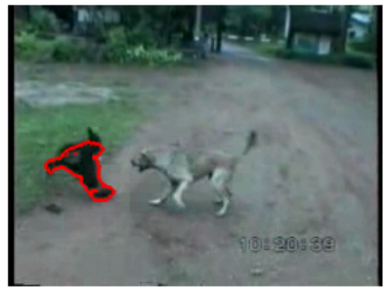
- Joint formulation of segmentation and optical flow estimation:



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

updated optical flow

Segmentation model

Motion estimation

update 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?

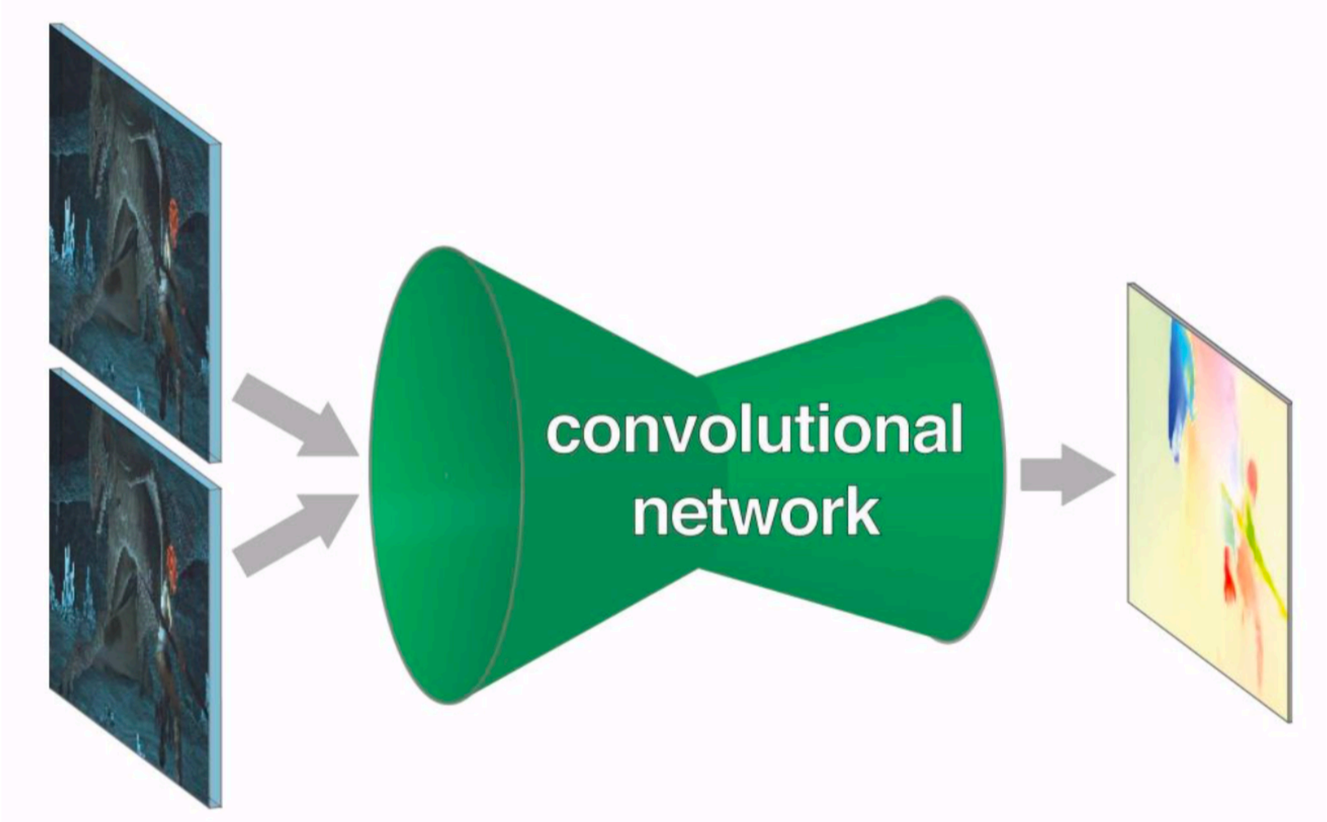
(a) frame $t-1$ (b) frame t

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

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

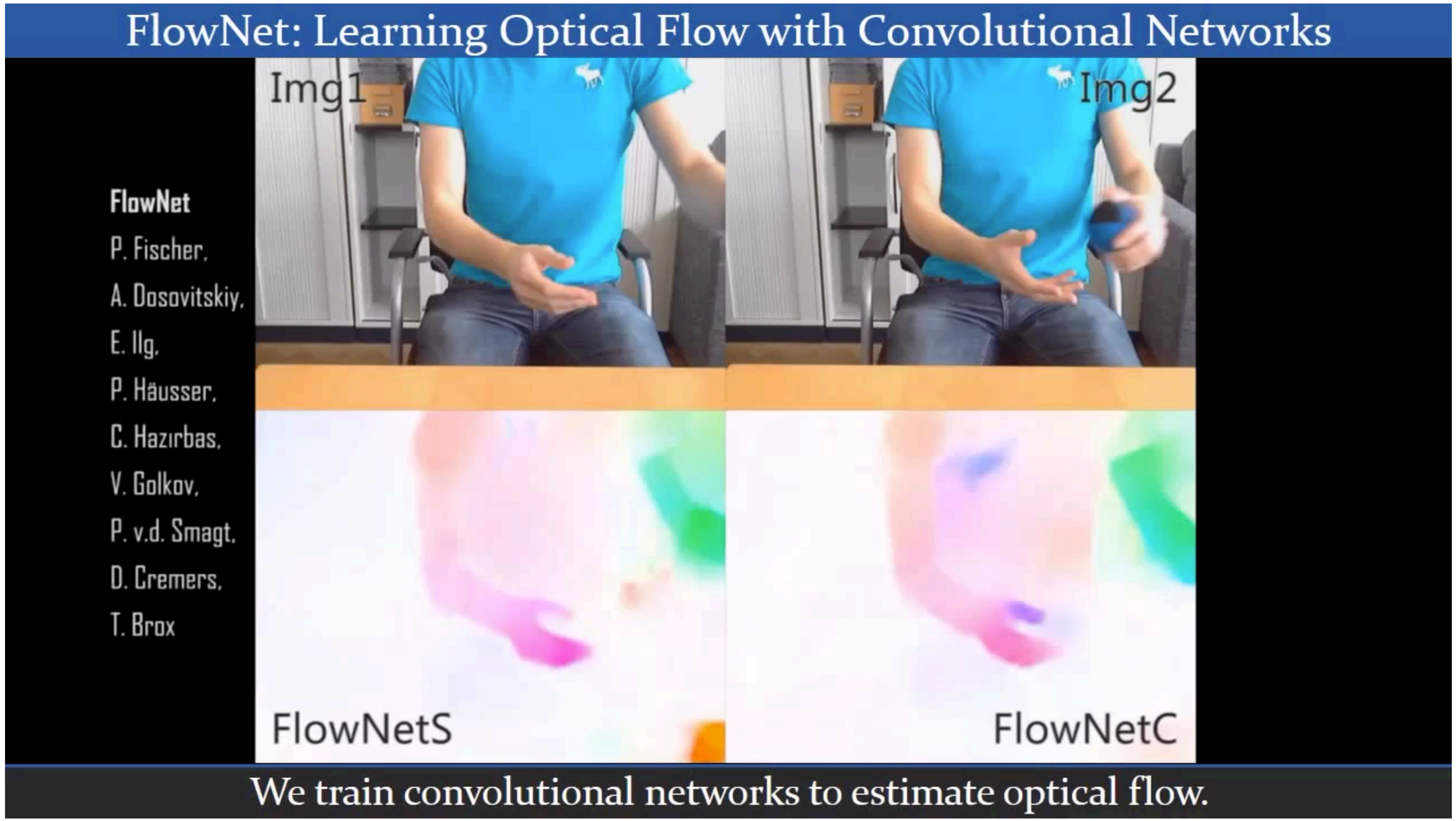
Optical flow with CNNs

- End-to-end supervised learning of optical flow



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

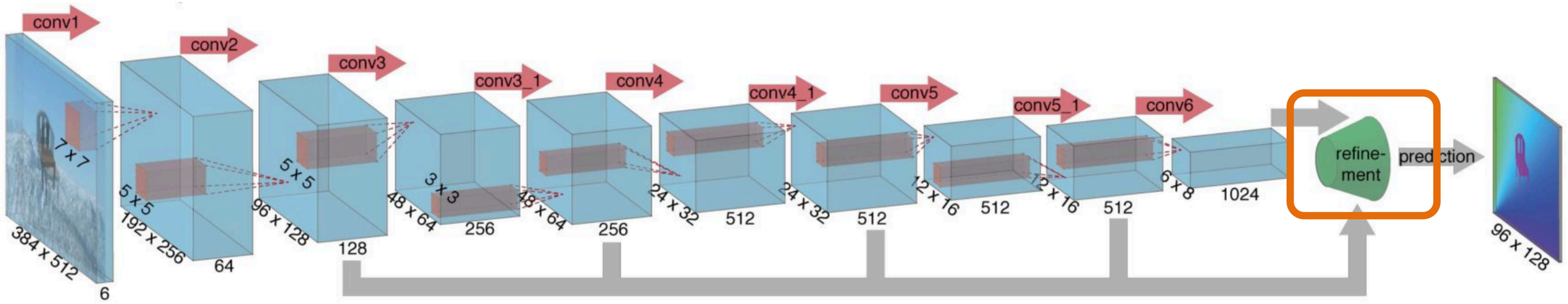
Optical flow with CNNs



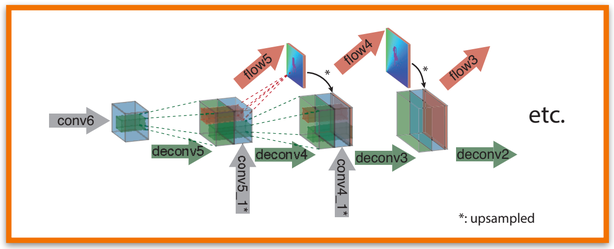
P. Fischer et al. „FlowNet: Learning Optical Flow With Convolutional Networks“. ICCV 2015

FlowNet: Architecture 1

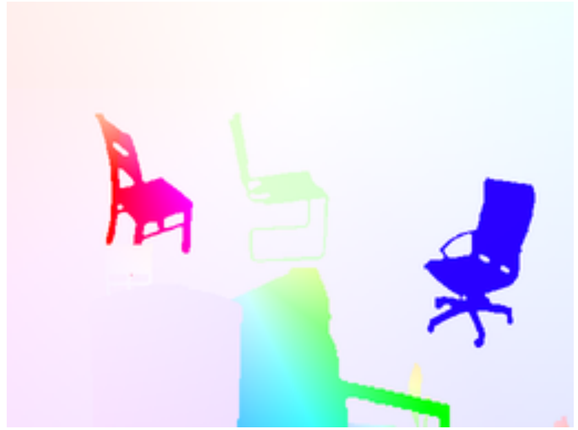
- Stack both images → input is now 2 x RGB = 6 channels

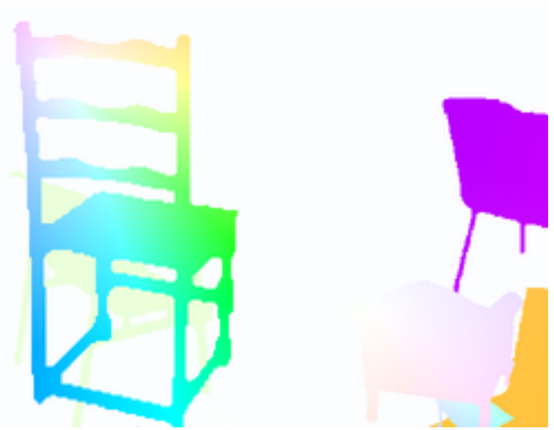


- Training with L2 loss from synthetic data



Optical flow from synthetic data

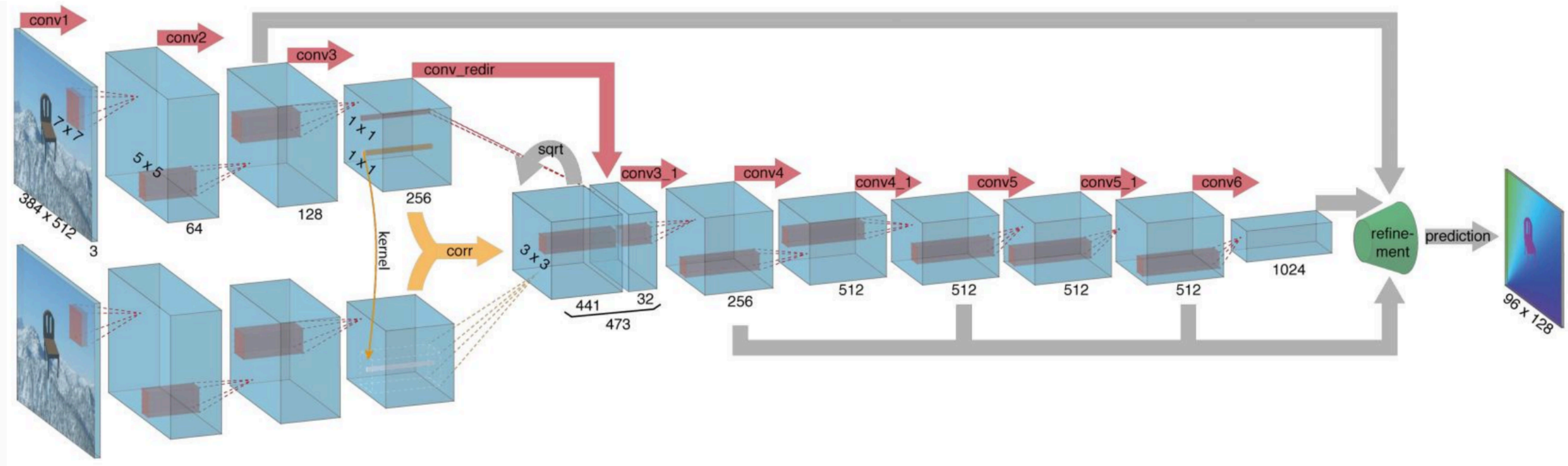




- Why chairs?
  - ...because we have large collections of 3D models.

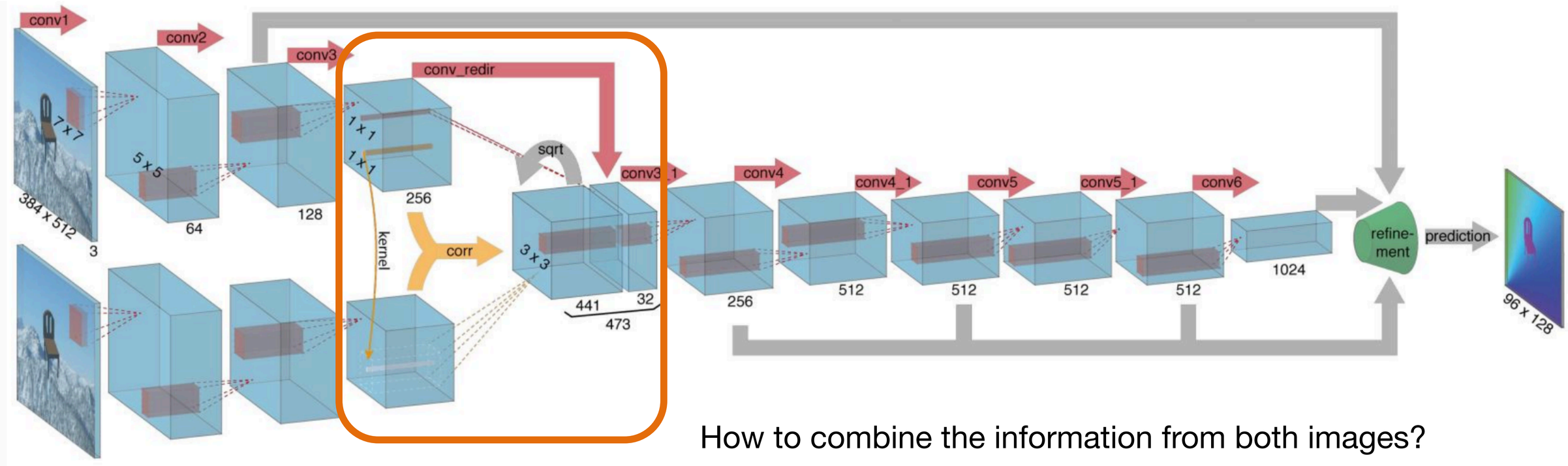
FlowNet: Architecture 2

- Siamese architecture



FlowNet: Architecture 2

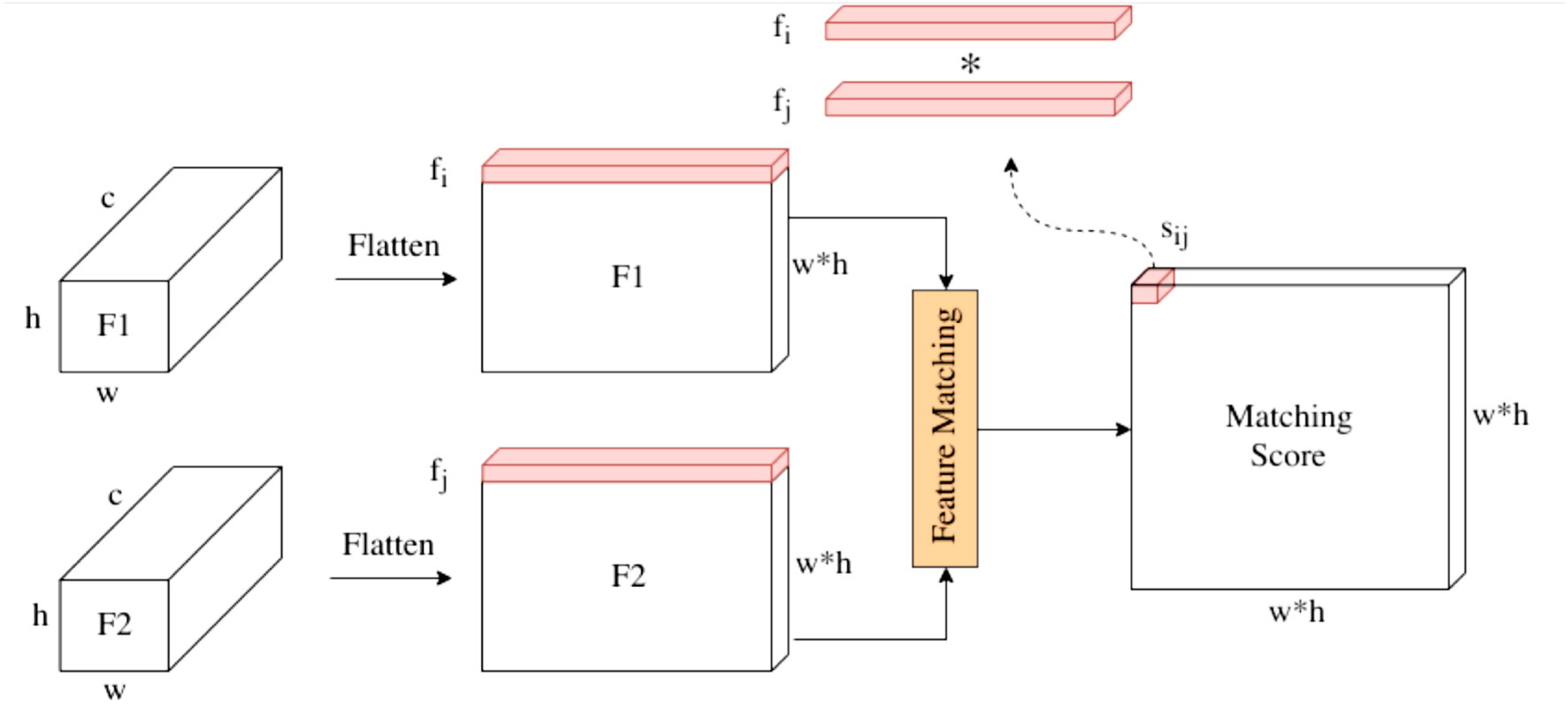
- Key design choice:



How to combine the information from both images?

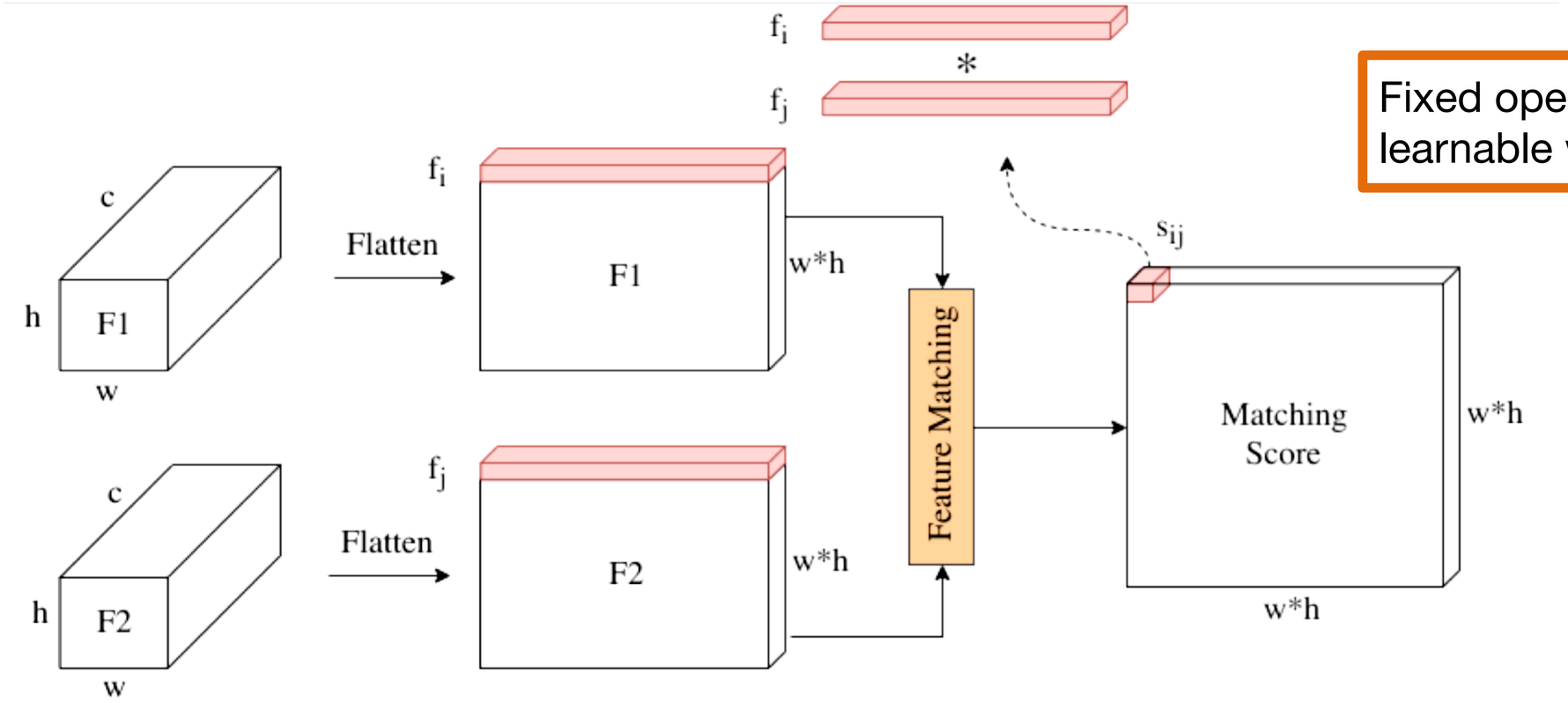
Correlation layer

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



Correlation layer

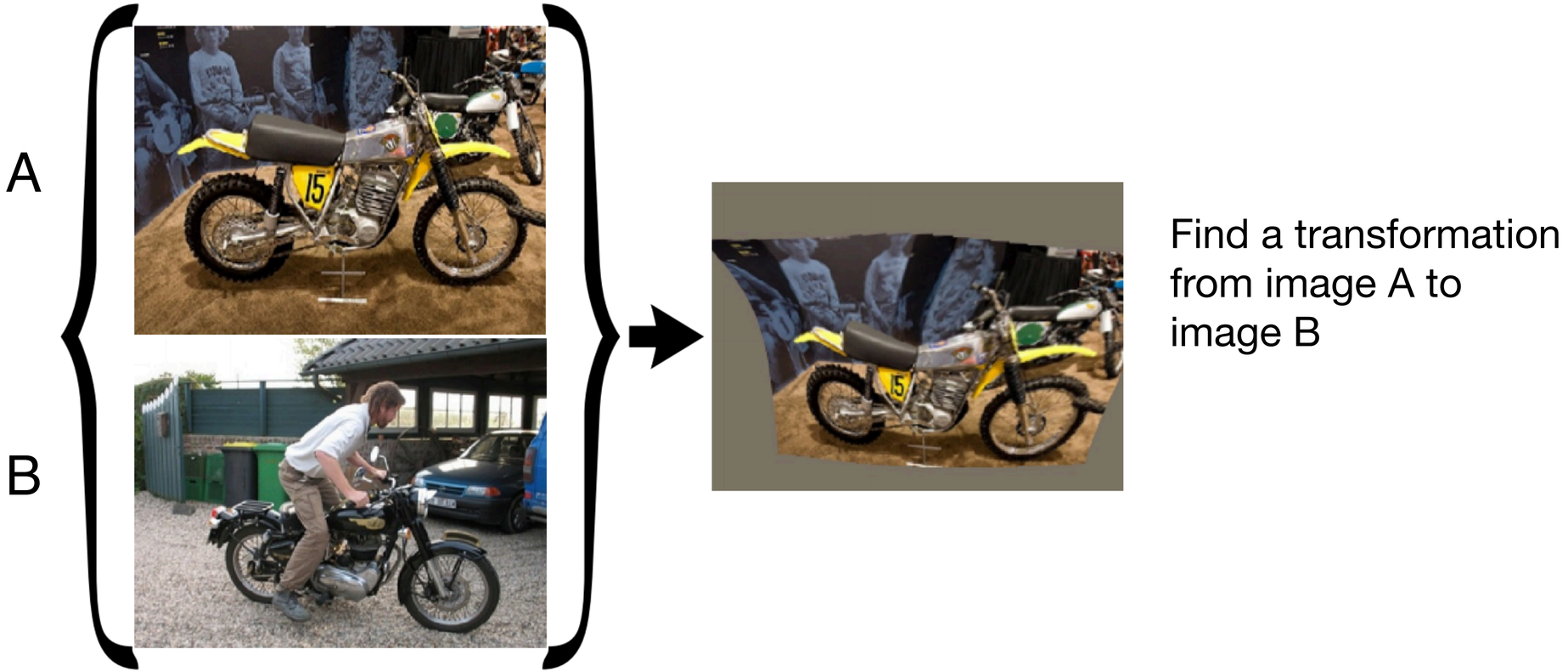
- The dot product measures similarity of two features



Fixed operation. No learnable weights!

Correlation layer

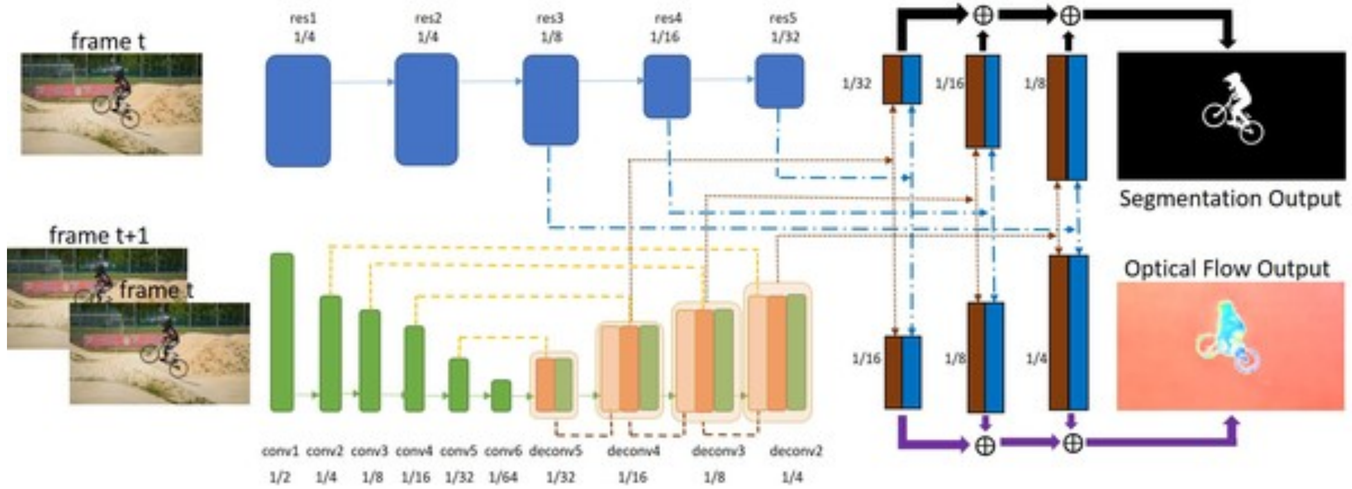
- Correlation layer is useful for finding image correspondences



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

SegFlow

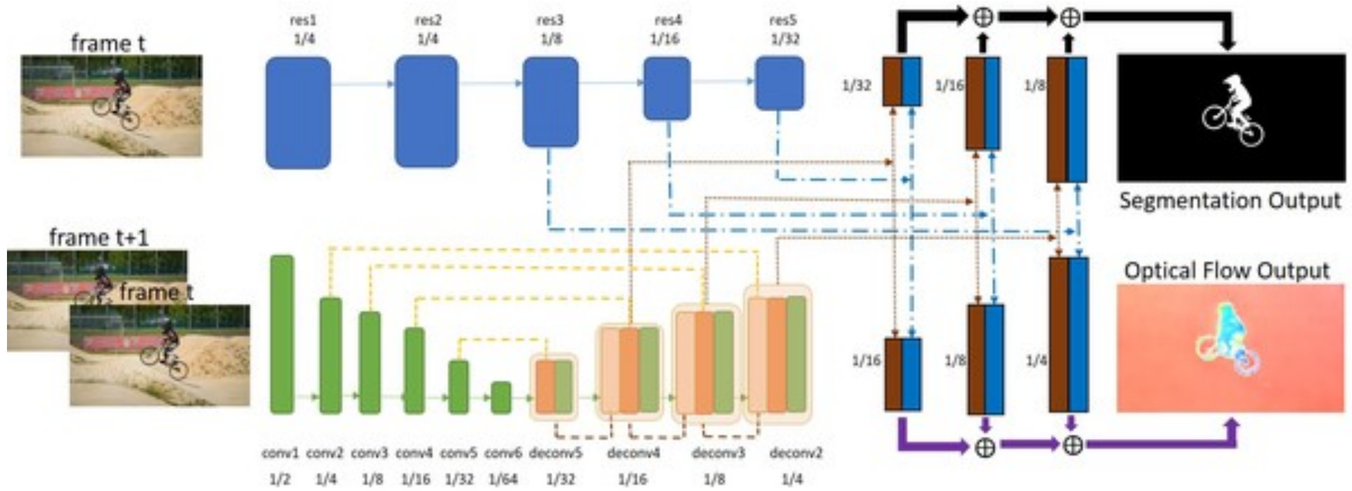
- Joint estimation of optical flow and object segment:



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

SegFlow

- Joint estimation of optical flow and object segment:

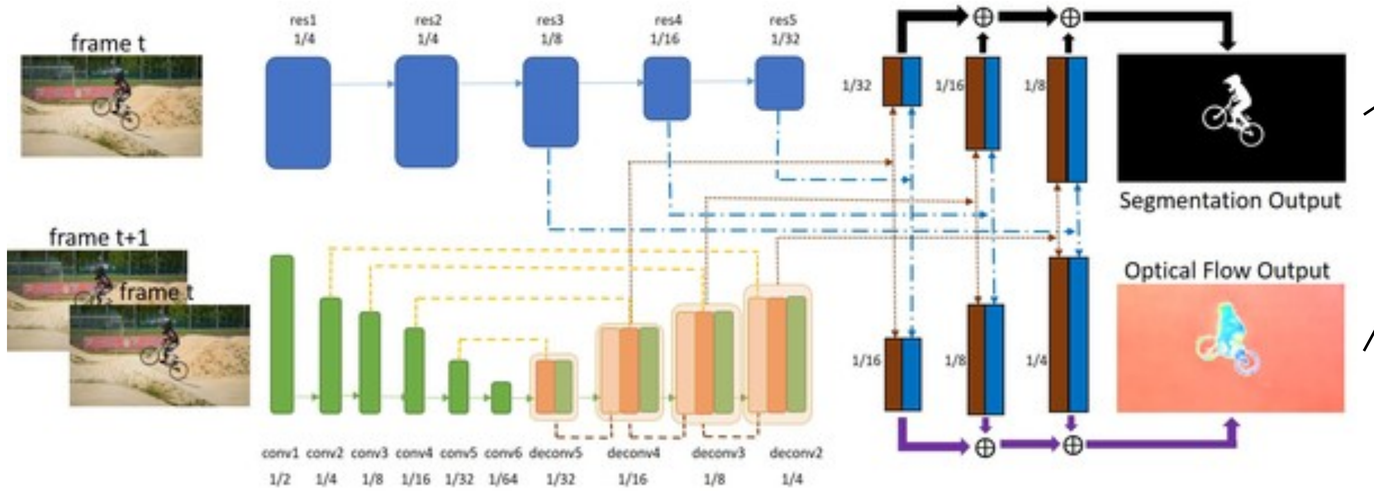


Joint feature representation at multiple scales

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

SegFlow

- Joint estimation of optical flow and object segment:



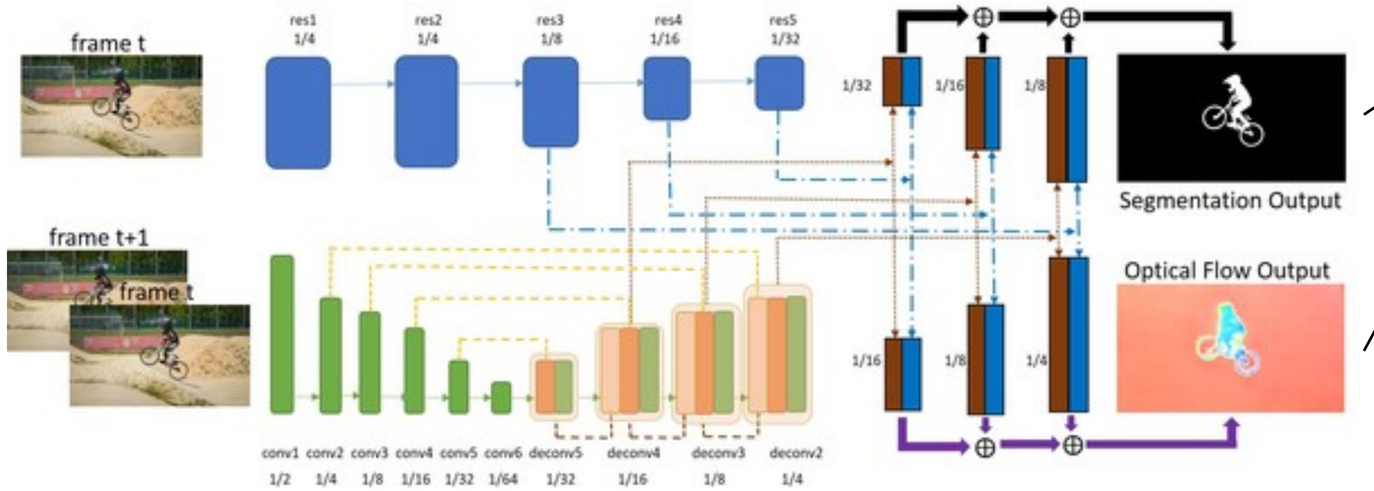
Supervision need not be synchronised

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

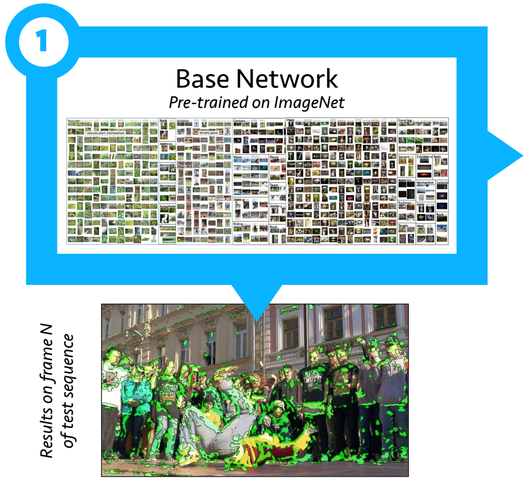
Appearance-only models

- 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.

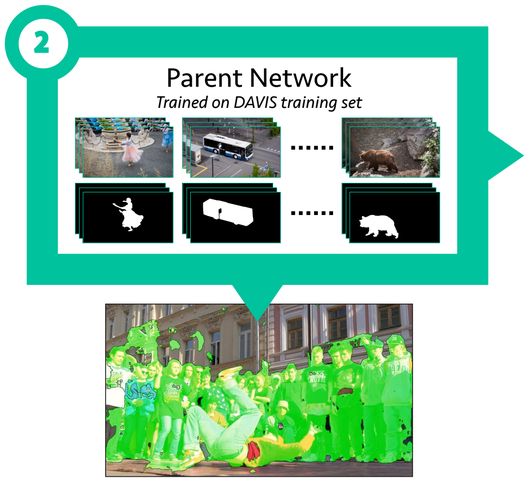
One-shot VOS

Pre-training



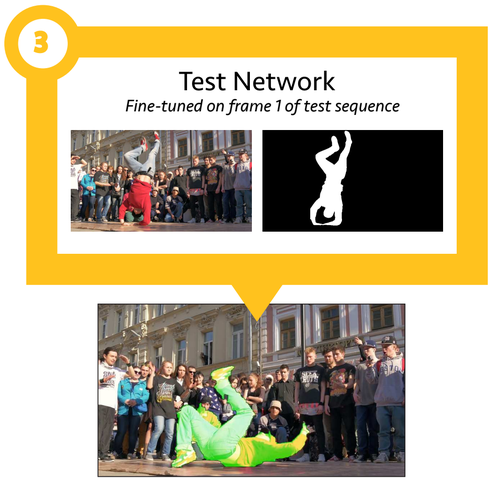
Edges and basic image features

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.

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

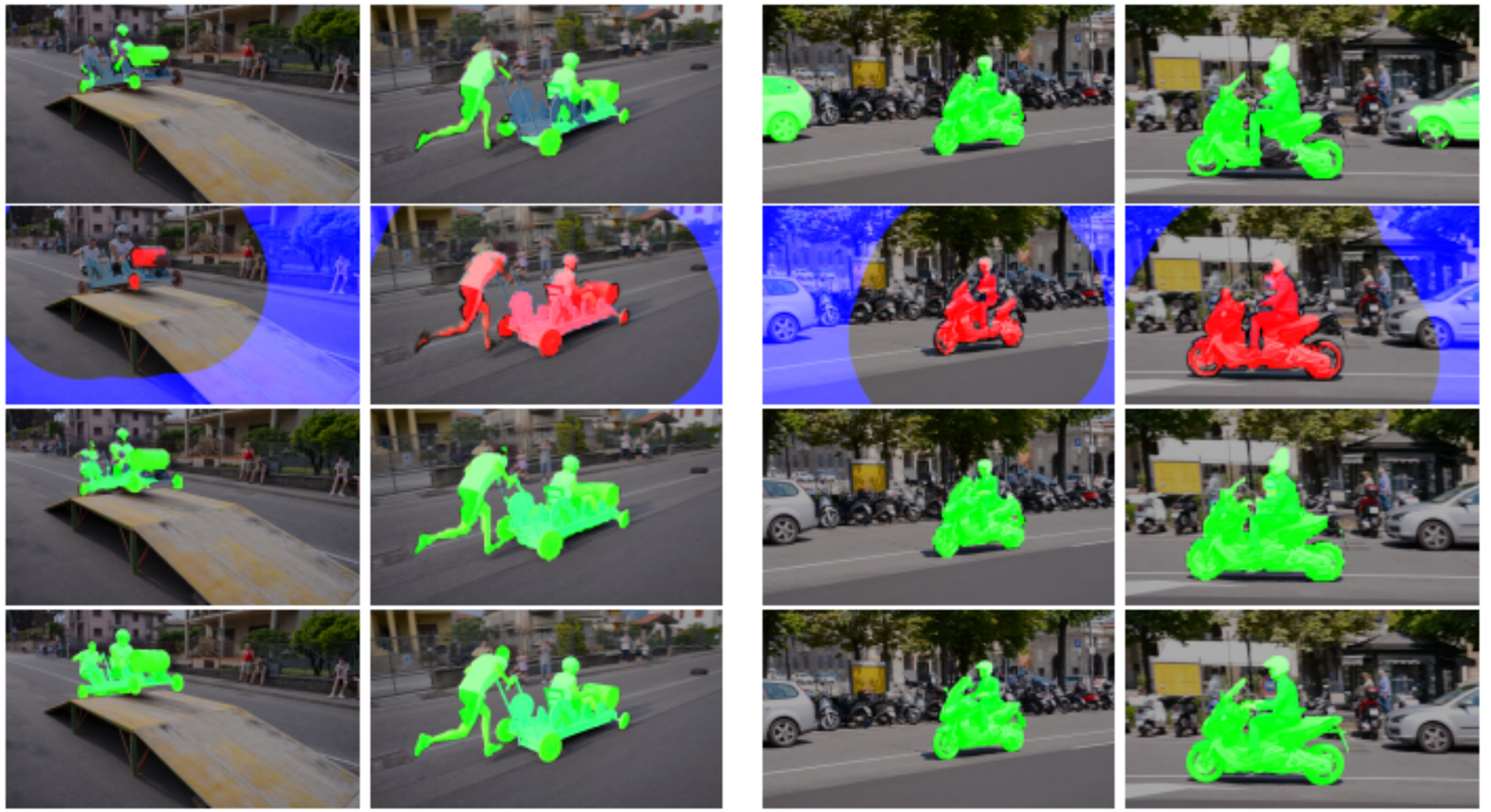
OnAVOS: Online Adaptation

Before adaptation

Pseudo labels

After adaptation

Ground truth

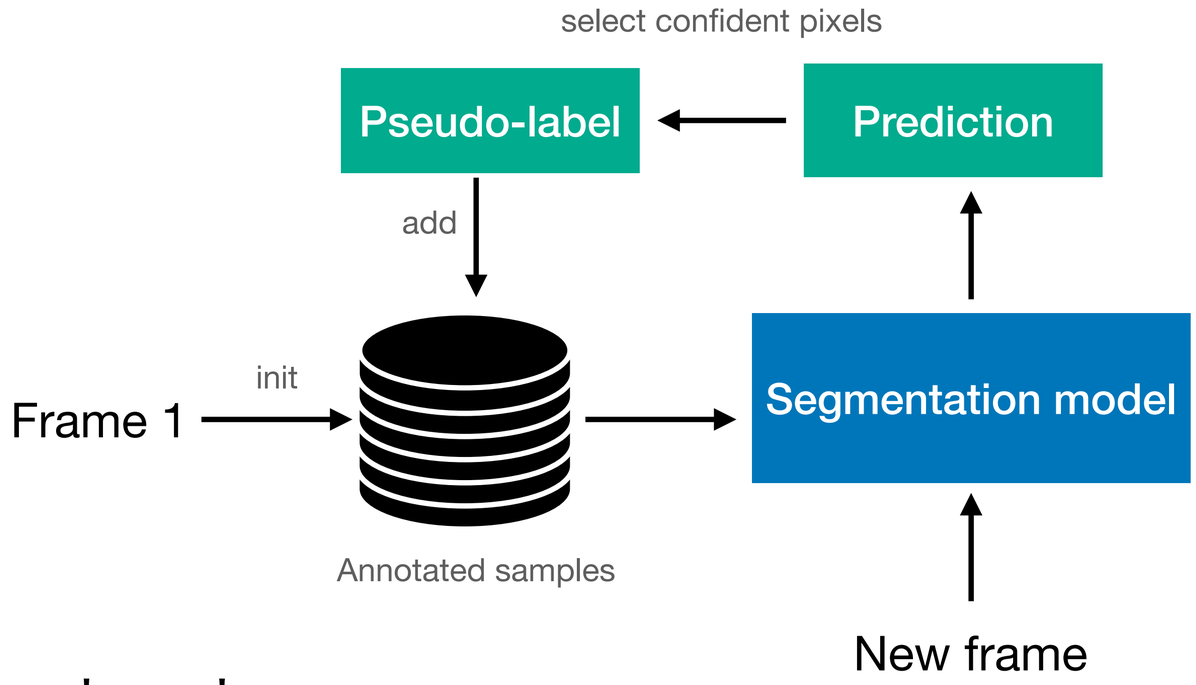


Blue = background samples
Red = foreground samples

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

OnAVOS: Online Adaptation

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



- Drawback: can be slow.

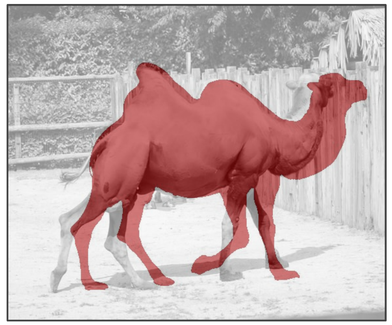
Issues

- 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.

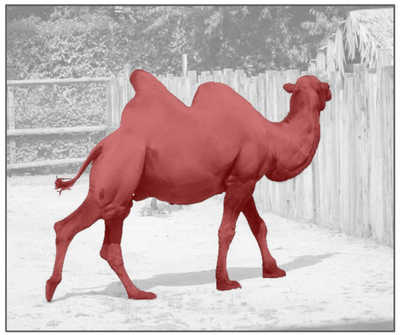
- Next: Can ensure we fine-tune the model with a correct signal?

MaskTrack

Input frame *t*



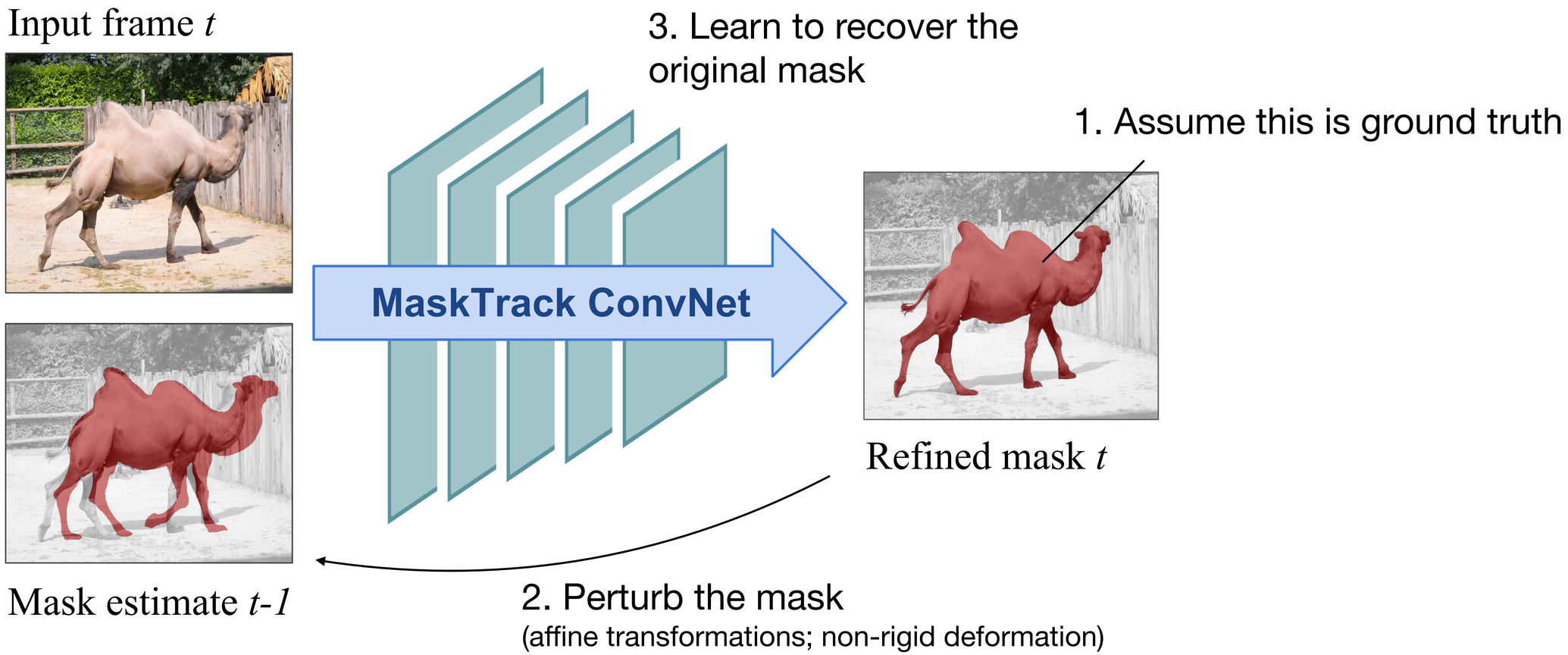
Mask estimate *t-1*



Refined mask *t*

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

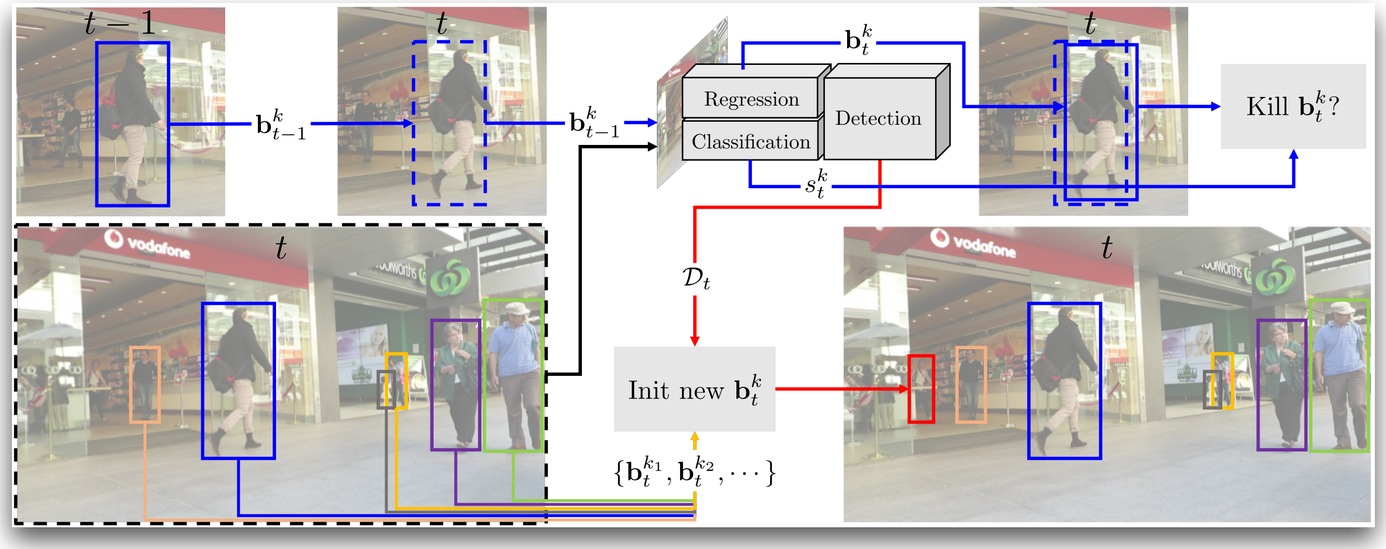
MaskTrack



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!

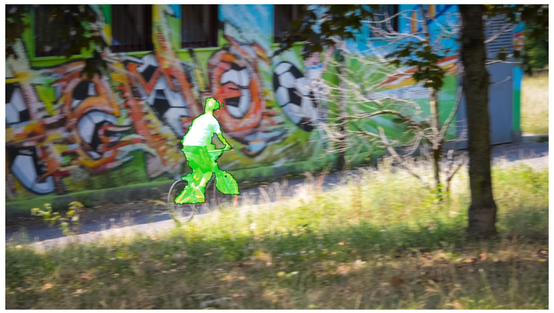
(a) Annotated image

(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);

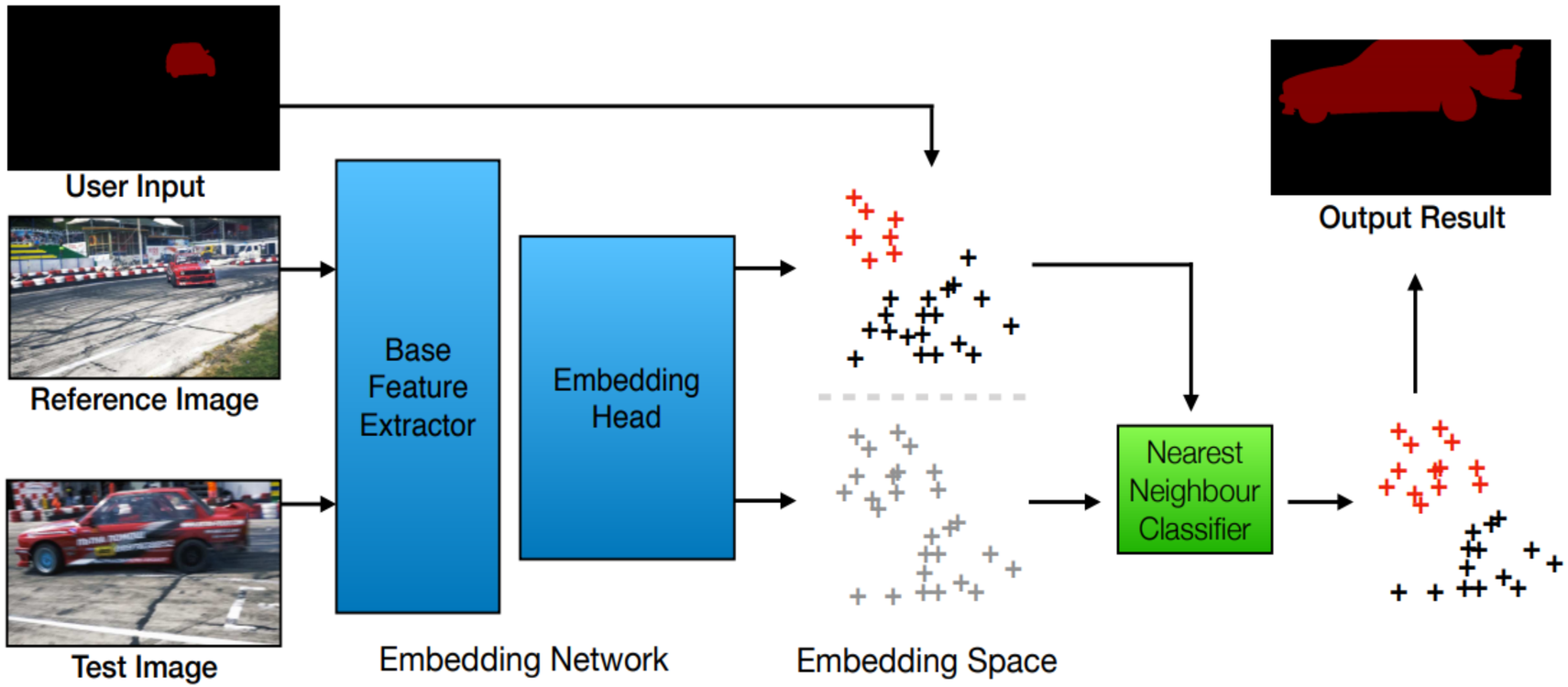
Metric-based approaches

Pixel-wise retrieval

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

Pixel-wise retrieval

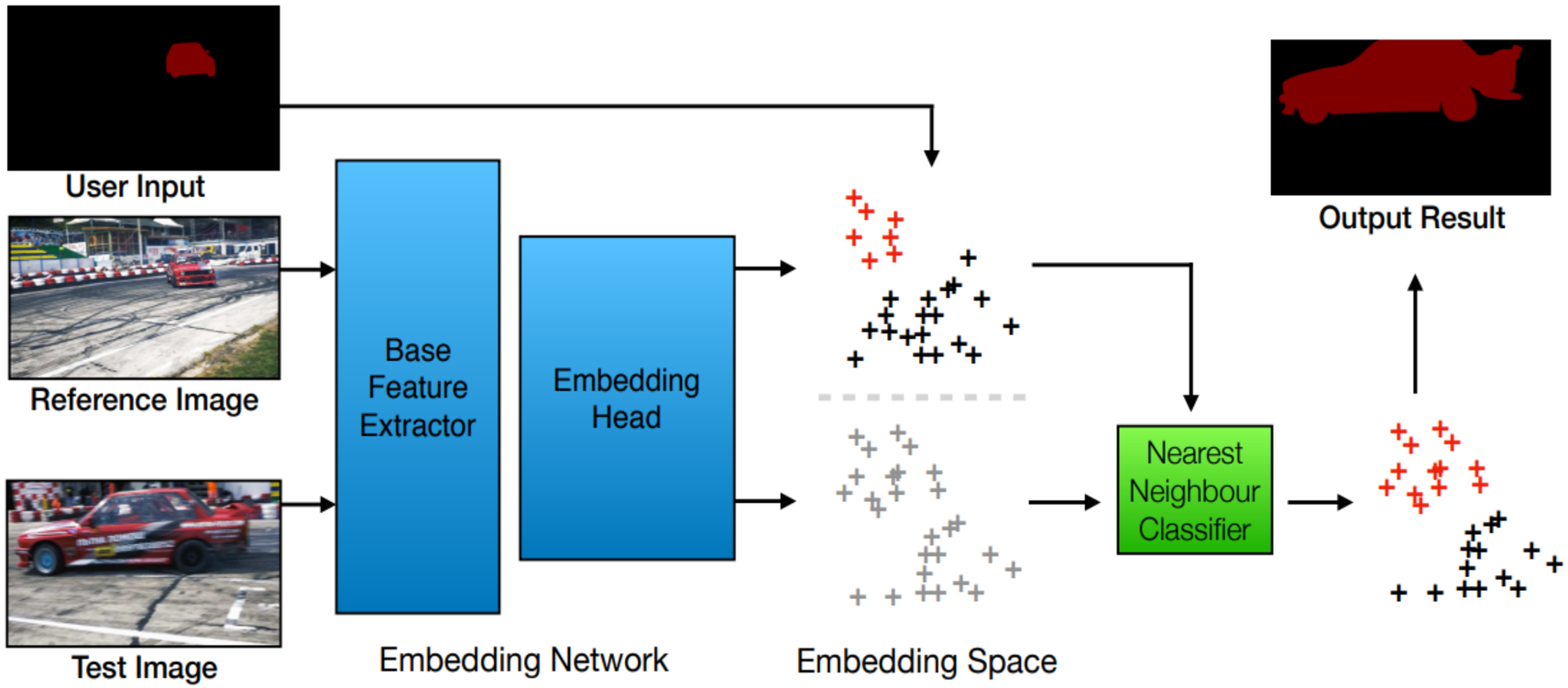
- The user input can be in any form, first-frame ground-truth mask, scribble...



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

Pixel-wise retrieval

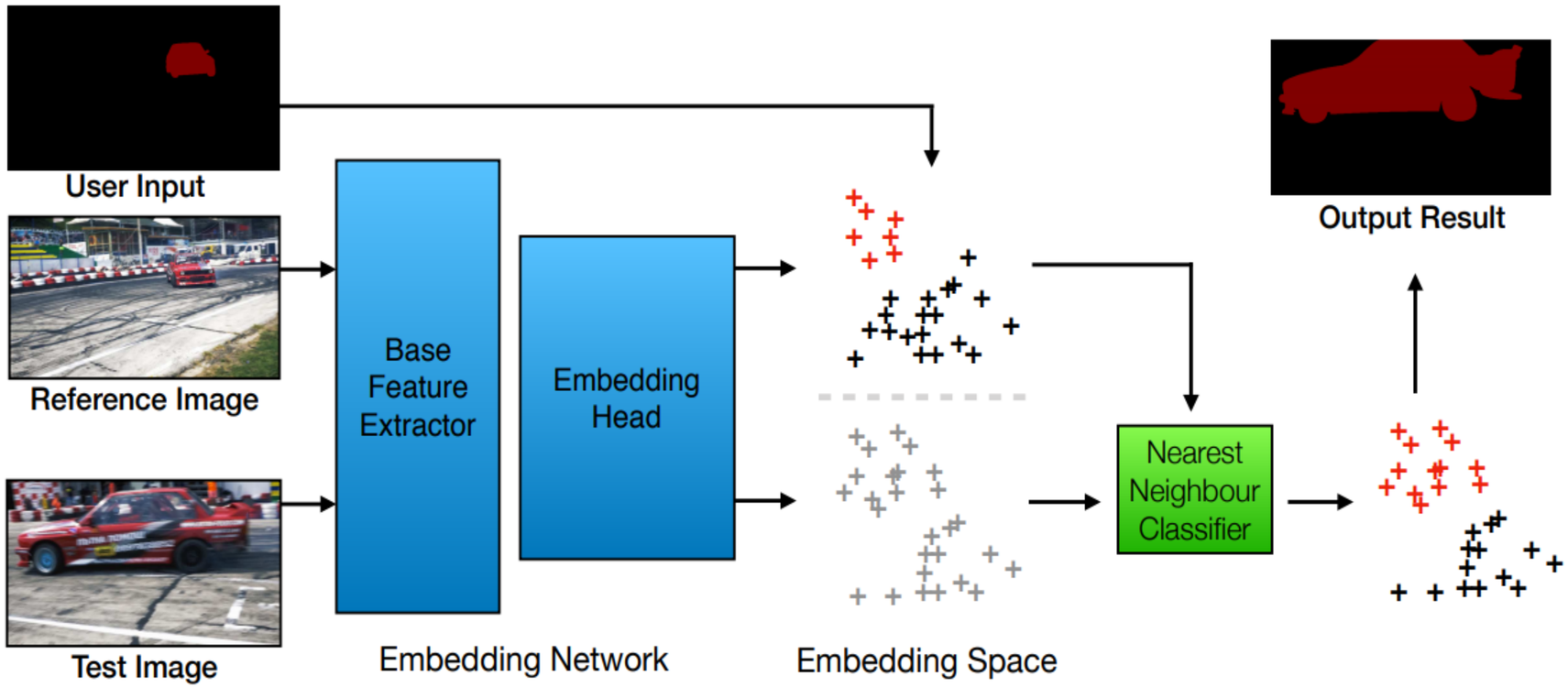
- Training: use the triplet loss to bring foreground pixels together and separate them from background pixels



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 (e.g. contrastive learning – later in the course).

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

Summary for today

- Motion-based models:
  - Optical flow (Lucas-Kanade);
  - FlowNet; SegFlow;
- Appearance-based models:
  - OSVOS: First-frame fine-tuning;
  - OnAVOS: Online Adaptation;
  - MaskTrack: Mask Refinement;
  - Pixel-wise retrieval

Computer Vision III:
Video object segmentation

Nikita Araslanov
14.12.2023

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