

# Basic video analysis

Vlad Shakhuro



Graphics & Media Lab  
Vision Group

15 November 2023

# Outline

I. Intro

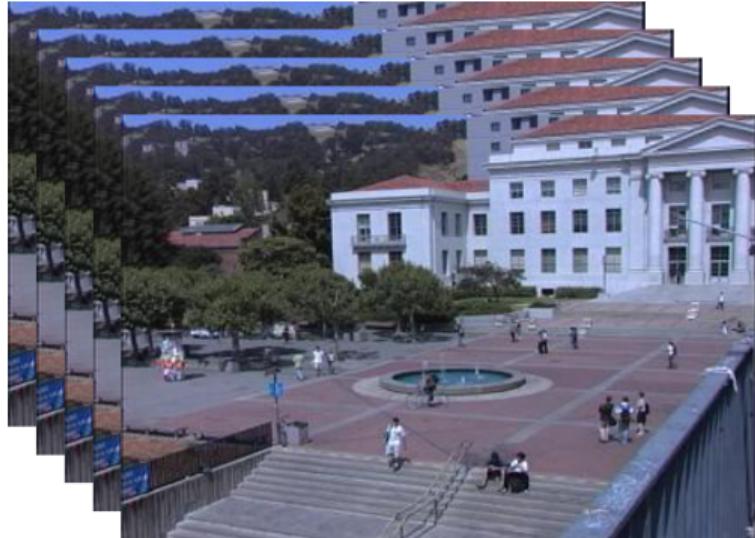
2. Optical flow

3. Visual Object Tracking

4. Multiple Object Tracking

5. Action recognition

# Types of video



Video — sequence of frames obtained from single camera in short periods of time

Videostream assumes online processing of frames, may be unbounded in time

Videosequence assumes offline processing, all frames are available

Raw data flow is greater than 1 Gb ethernet capacity:  
 $2\text{MB} \times 3 \times 30 = 180 \text{ MB/s}$

# Shooting scenarios

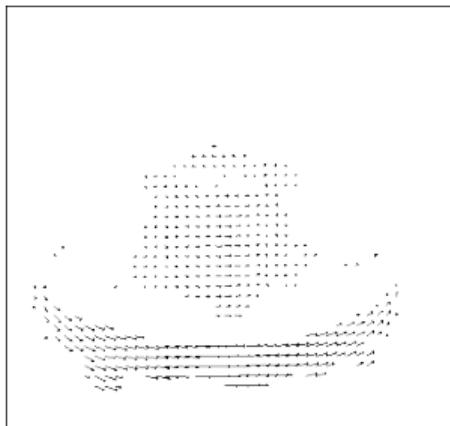


Camera view and appearance of objects may be very different. Production models are created for specific shooting scenarios

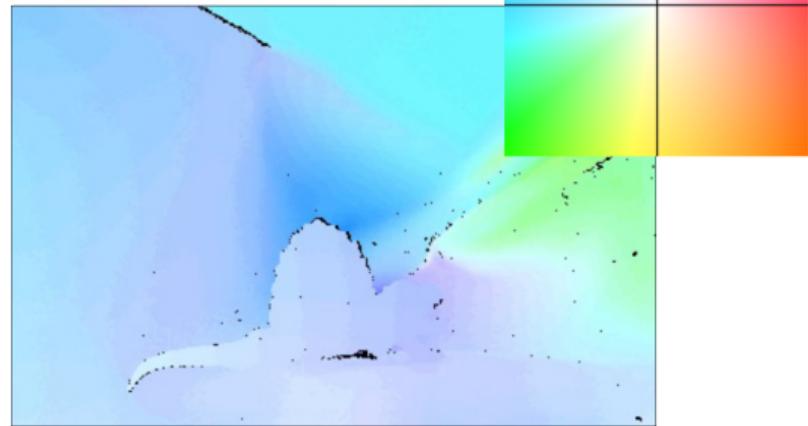
# Outline

- I. Intro
2. Optical flow
3. Visual Object Tracking
4. Multiple Object Tracking
5. Action recognition

# Optical flow



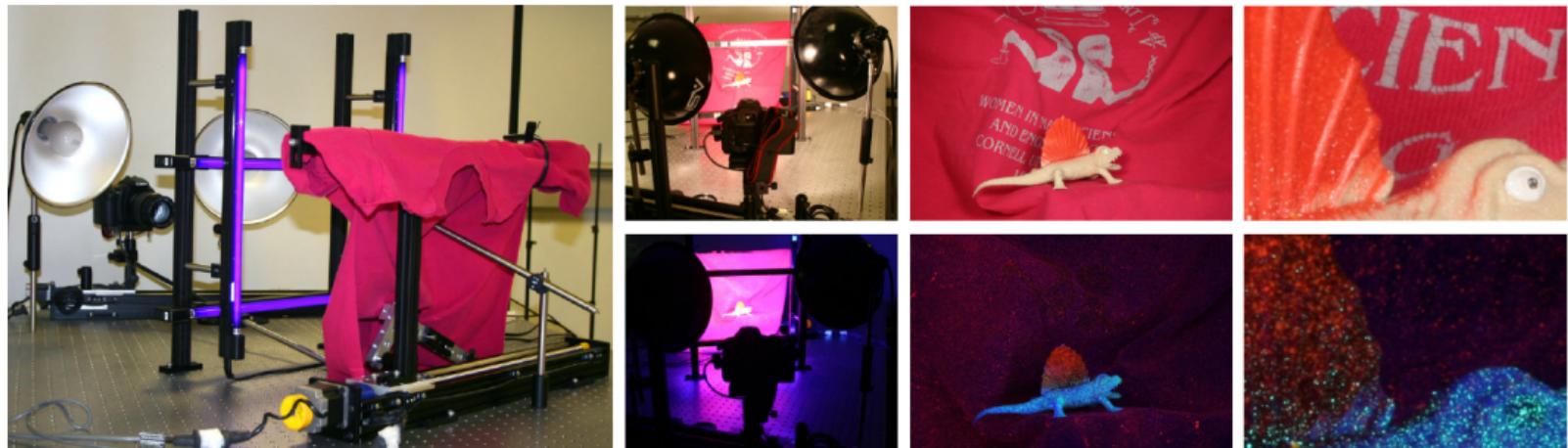
motion vectors for some points



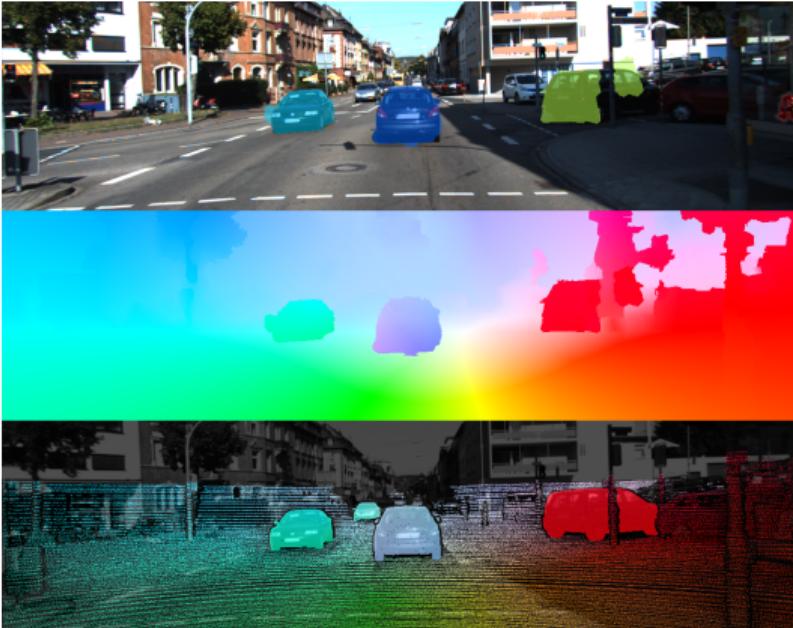
color coding of motion vectors

Optical flow — vector field of visible movement of pixels between frames

# Middlebury dataset



# KITTI

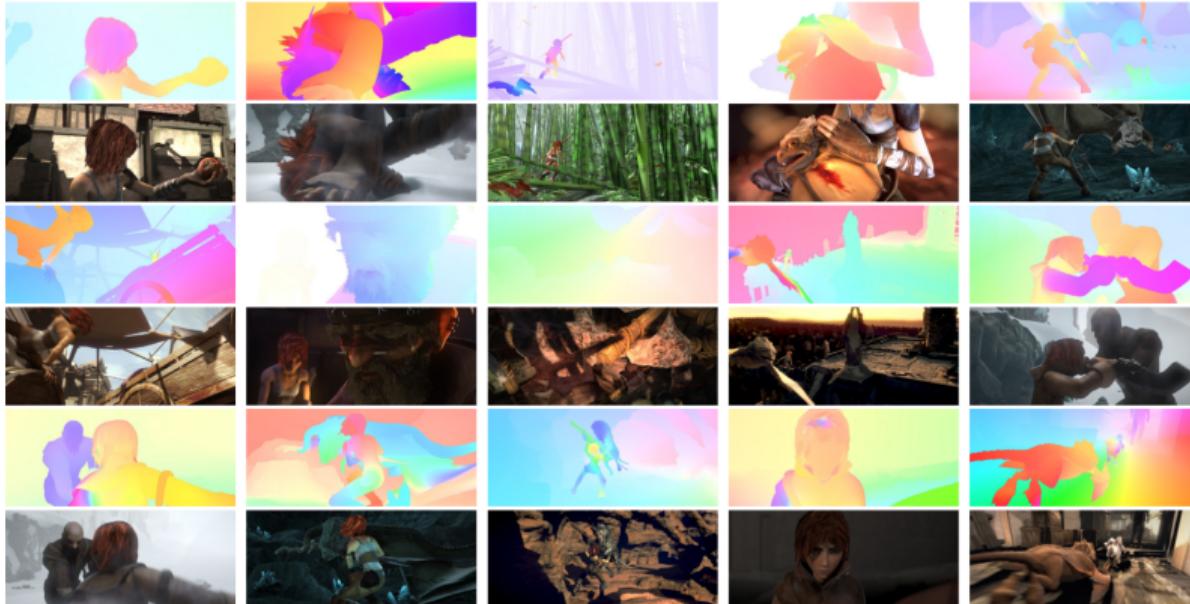


200 training and 200 testing frames

Optical flow was estimated:

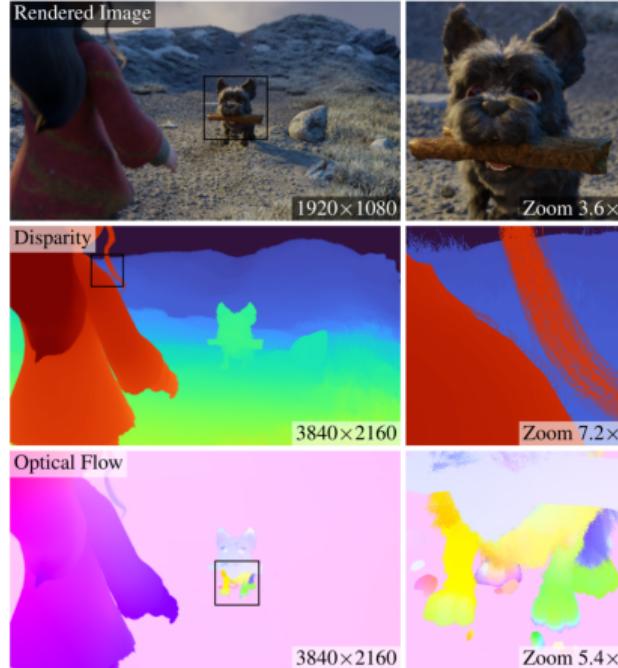
- via 3D reconstruction for background
- by fitting CAD models for moving objects (cars)

# Sintel



1064 training and 564 testing frames,  $1024 \times 436$  resolution

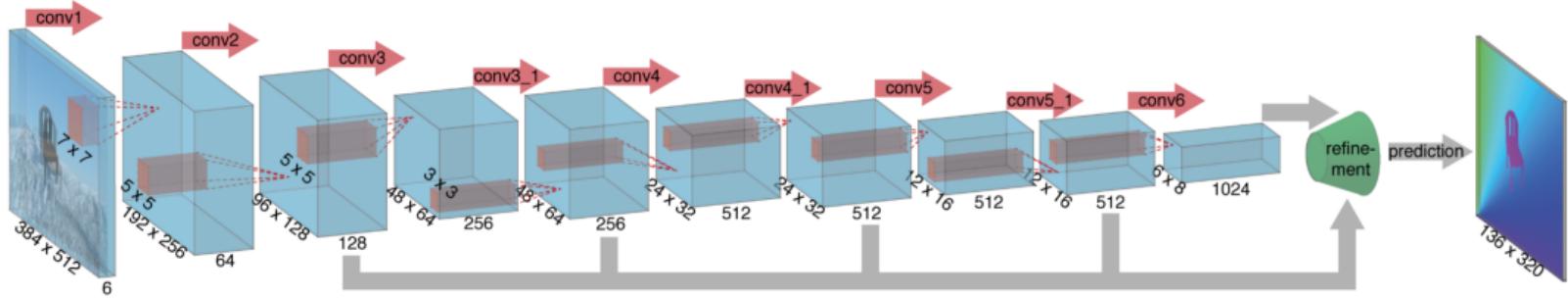
# Spring



6000 FullHD images with 4K ground truth

Mehl et al. Spring: A High-Resolution High-Detail Dataset and Benchmark for Scene Flow, Optical Flow and Stereo. CVPR 2023

# FlowNet



Concatenate a pair of images and pass it to dense prediction network

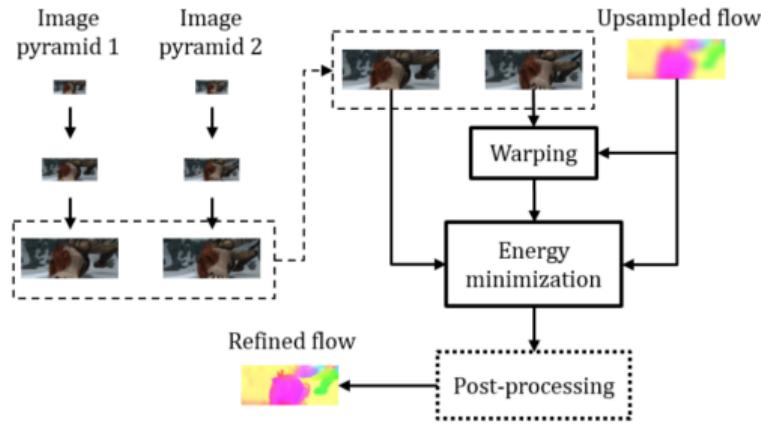
# FlowNet



	Frame pairs	Frames with ground truth	Ground truth density per frame
Middlebury	72	8	100%
KITTI	194	194	~50%
Sintel	1,041	1,041	100%
Flying Chairs	22,872	22,872	100%

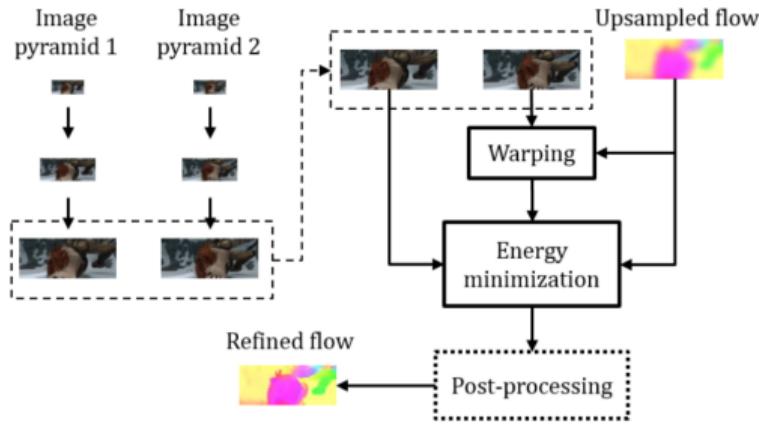
Use synthetic dataset Flying chairs for pretraining

# PWC-Net

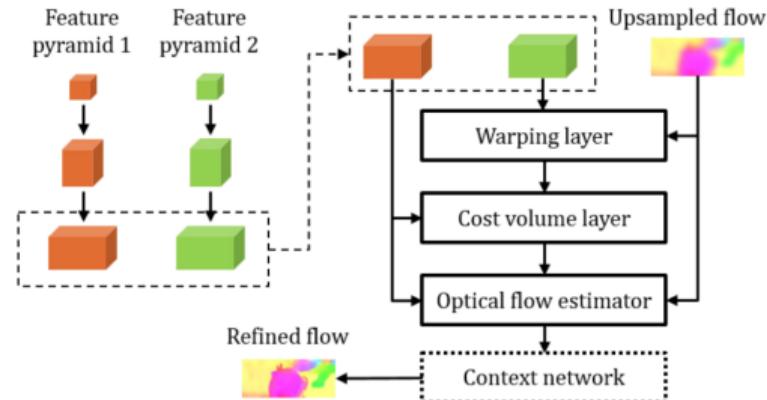


classical approach

# PWC-Net



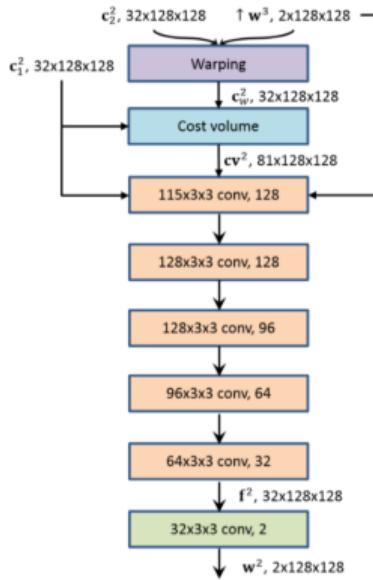
classical approach



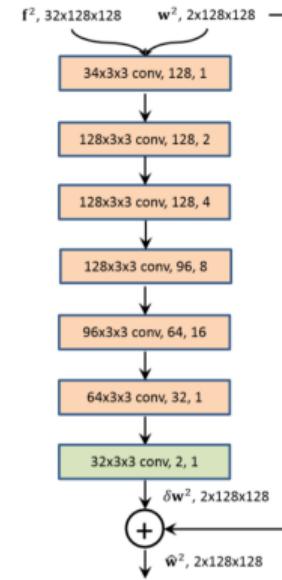
PWC-Net

Compute Cost Volume (CV) via correlation for pixels within distance  $d$ .  
Size of CV tensor will be  $d^2 \times H \times W$

# PWC-Net



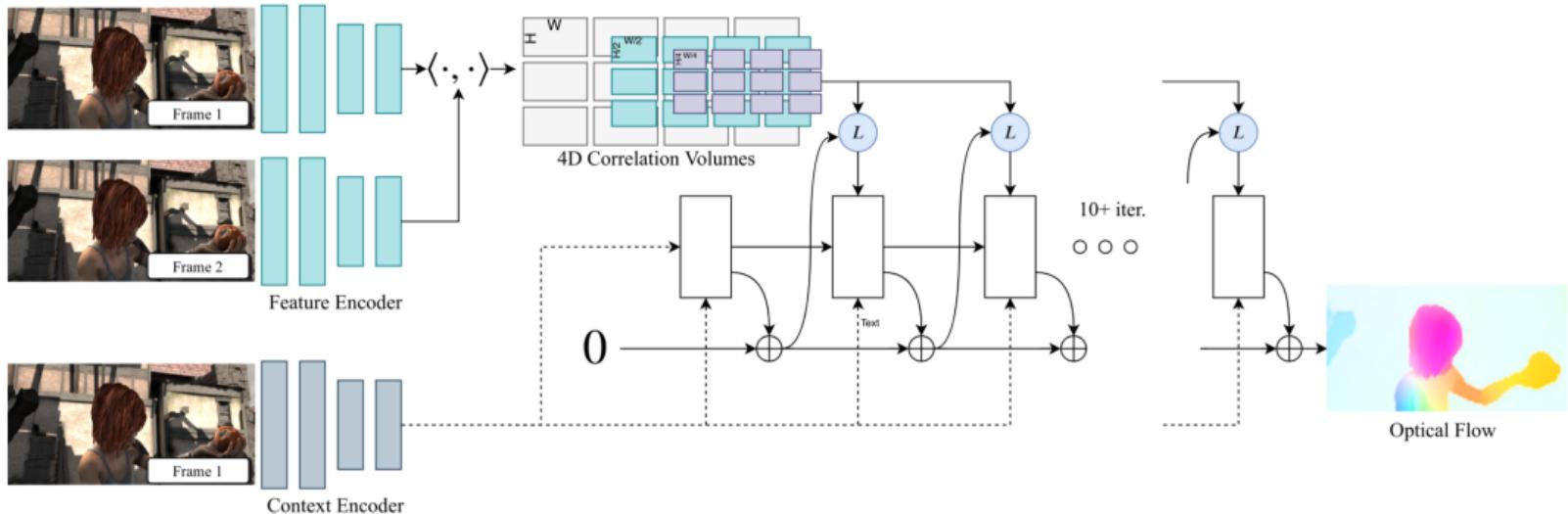
OF estimator



context network

Sun et al. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. CVPR 2018

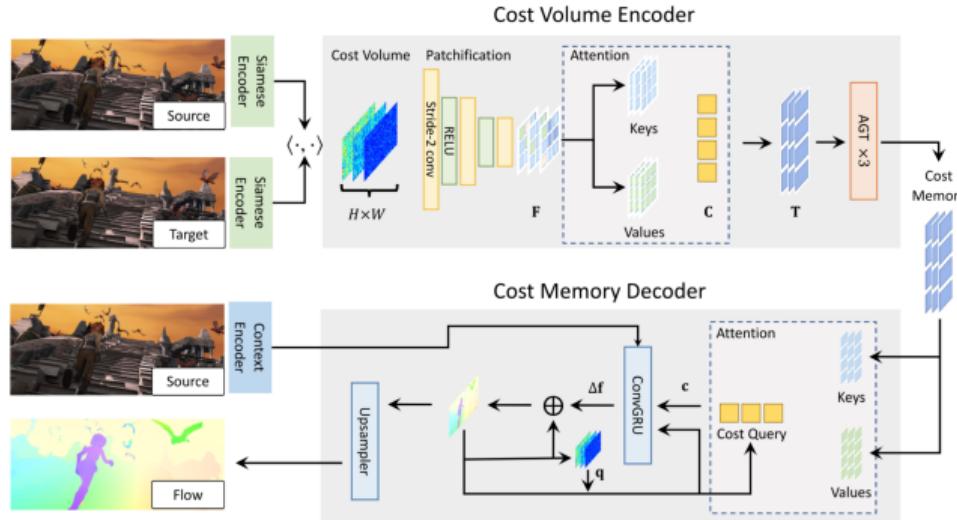
# RAFT



## Key ideas:

- maintain high-res estimation of optical flow without pyramid
- use recurrent unit for refinement of optical flow

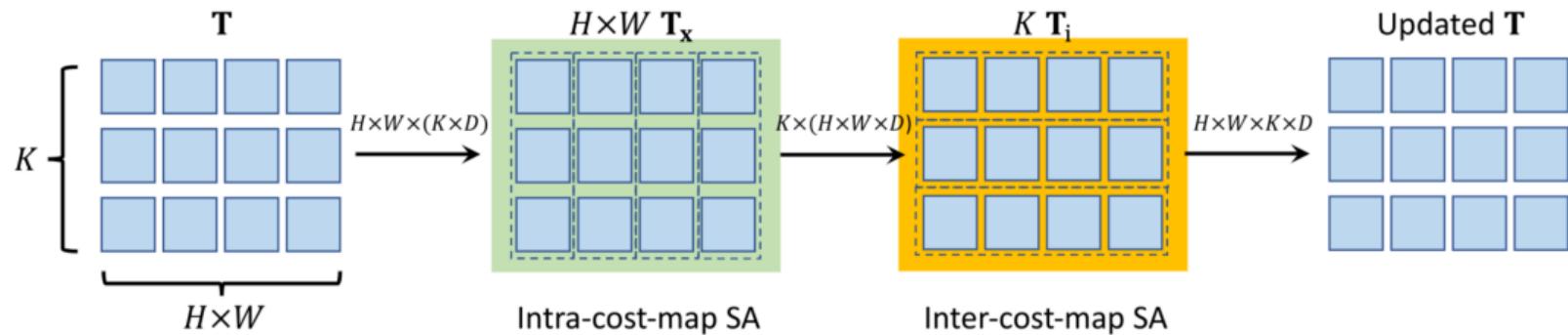
# FlowFormer



Key ideas:

- compress cost volume to tokens ( $H^2 \times W^2 \rightarrow H \times W \times K \times D$ )
- two-step attention: for tokens withing same cost map and across different cost maps
- decoder with cost queries

# FlowFormer



# Outline

- I. Intro
2. Optical flow
3. Visual Object Tracking
4. Multiple Object Tracking
5. Action recognition

# Visual Object Tracking



Single arbitrary object localized on the first frame

Object should be tracked on short time interval in online mode

# VOT model requirements



- model works in real time
- object appearance may change significantly
- similar and occluded objects make task more complex

# VOT Challenge

Small number (~50) of various videos

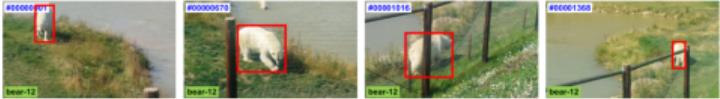
Most challenging:



Least challenging:



# LaSOT



Bear-12: "white bear walking on grass around the river bank"



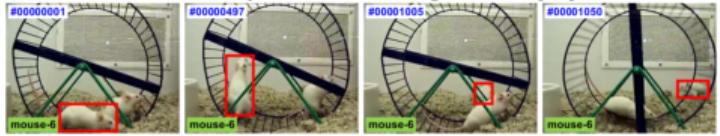
Bus-19: "red bus running on the highway"



Horse-1: "brown horse running on the ground"



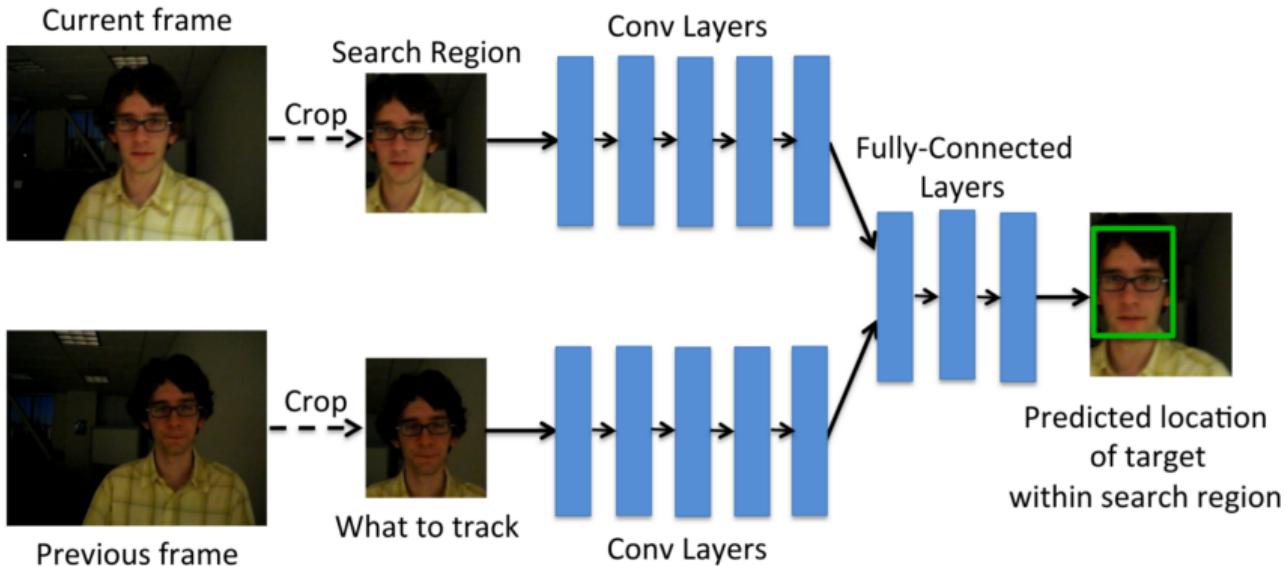
Person-14: "boy in black suit dancing in front of people"



Mouse-6: "white mouse moving on the ground around another white mouse"

- 1400 videos (YouTube CC license)
- 84s duration on average
- 3.5M frames in total
- 70 classes chosen for popular applications
- labelled by 10 volunteers and PhD students

# GOTURN



# Training GOTURN



Previous frame centered on object



Generate next synthetic frame from current using random crops and resizes

# STARK

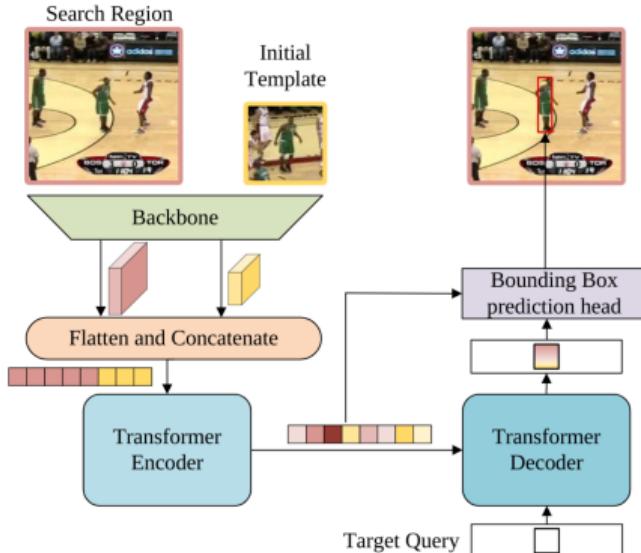


Figure 2: Framework for spatial-only tracking.

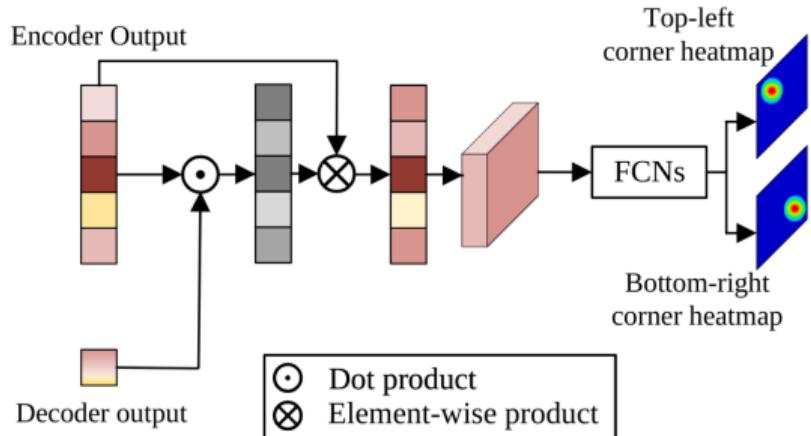
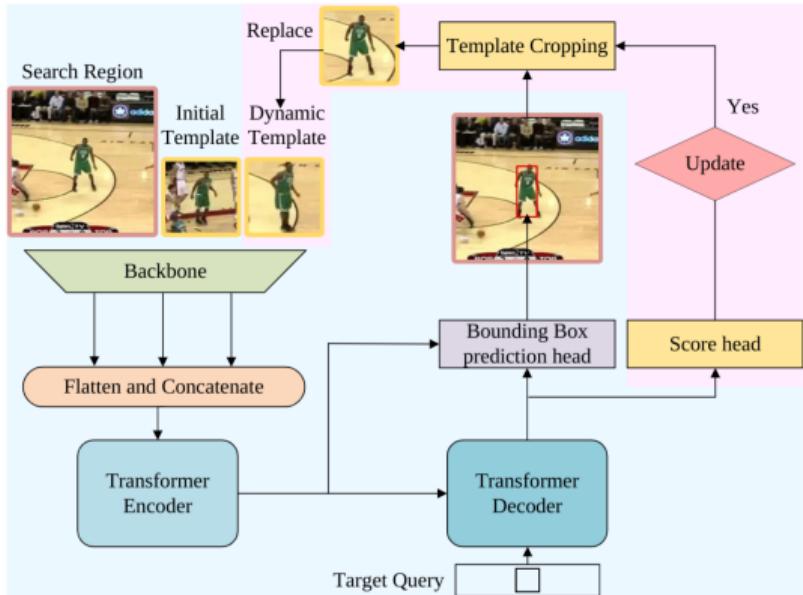


Figure 3: Architecture of the box prediction head.

# STARK



Two stage-training: first localization, then classification

Figure 4: Framework for spatio-temporal tracking. The differences with the spatial-only architecture are highlighted in pink.

# Outline

- I. Intro
2. Optical flow
3. Visual Object Tracking
4. Multiple Object Tracking
5. Action recognition

# Multiple Object Tracking

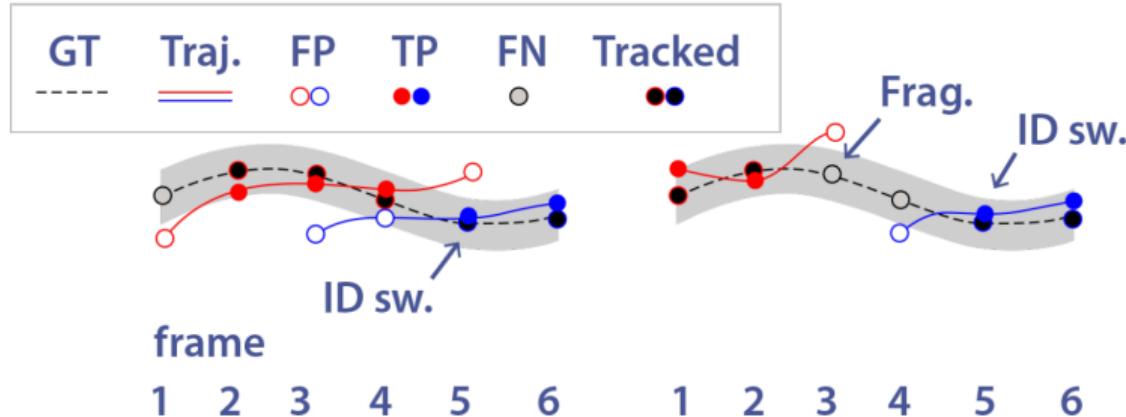


Track multiple objects on a long period of time. Variants:

- Detection Based Tracking
- Detection Free Tracking

Output a set of trajectories for all visible objects in video. Object location is usually described with bounding box

# MOT errors



- ID switches — two or more trajectories are predicted for single trajectory
- Fragmentations — two trajectories are predicted for a single trajectory with gaps

# MOT metrics

Multiple Object Tracking Accuracy:

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t}$$

Multiple Object Tracking Precision  
(object localization accuracy, average overlap):

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{c_t}$$

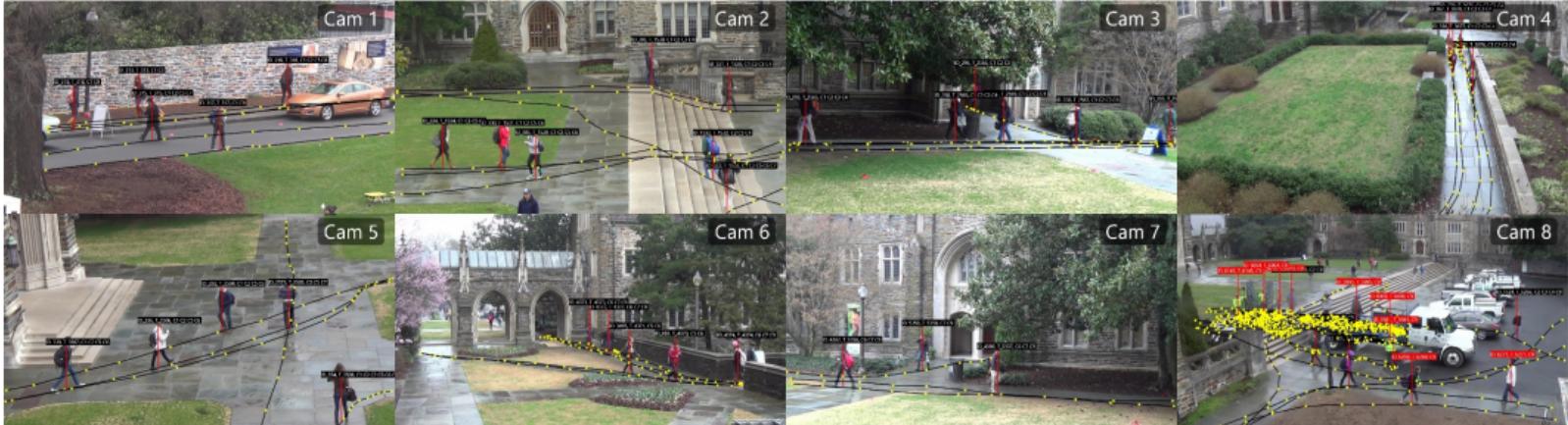
Mostly Tracked (MT), Mostly Lost (ML), Partially Tracked (PT):  
#objects tracked for > 80%, < 20% of the trajectory and in between  
these thresholds

# MOT20 Challenge



- 4 train and 4 test videos with challenging crowded scenes
- 14k frames, 9 minutes
- 1.5M and 0.7M bboxes for training and testing pedestrian detector

# Duke MTMC



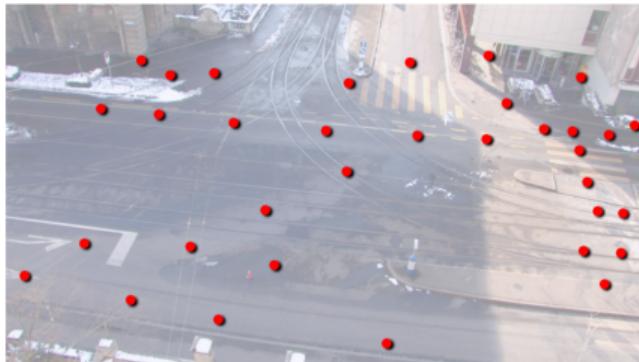
- 8 static cameras  $\times$  85 minutes of 1080p 60 fps video
- >2M manually annotated frames
- >2k identities
- manual annotation by 5 people over 1 year
- more identities than all existing MTMC datasets combined
- unconstrained paths, diverse appearance

# UA-DETRAC

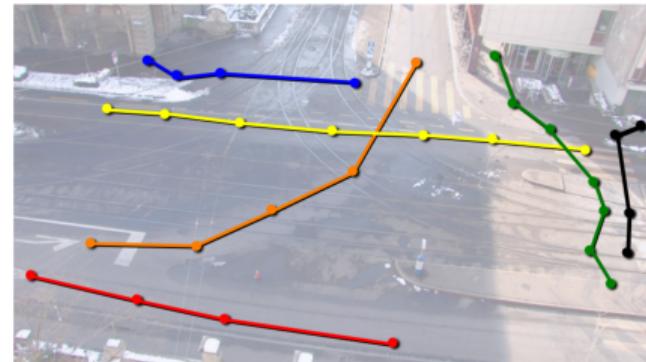


- 10 hours of videos, 25 fps, resolution  $960 \times 540$  pixels
- 24 different locations at Beijing and Tianjin in China
- > 140k frames
- 8250 vehicles
- 1.2M labeled boxes

# Tracking by detection



object detections



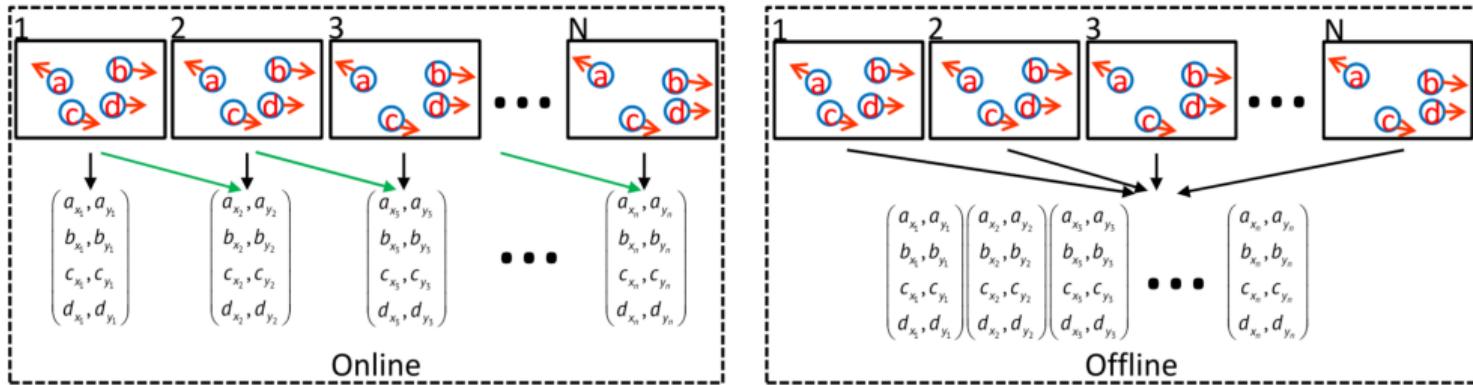
association of detections

# MOT and detection errors

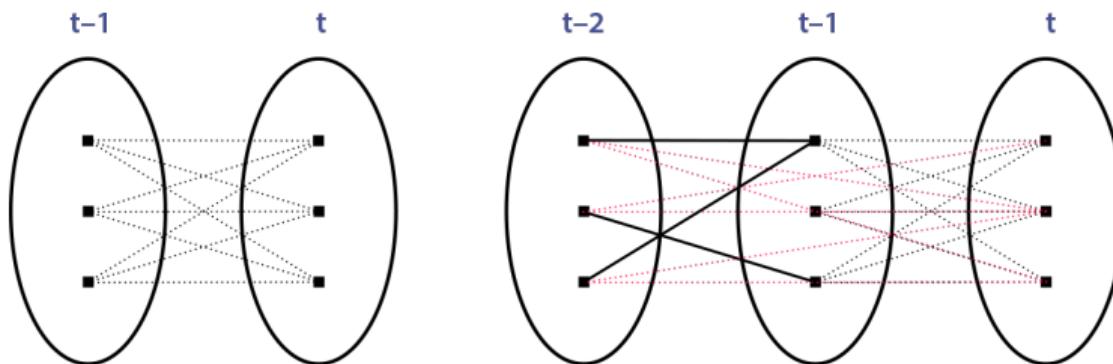


Trackers usually reduce FP and raise FN. Therefore, good detector is a key for a good MOT method

# Online vs offline tracking

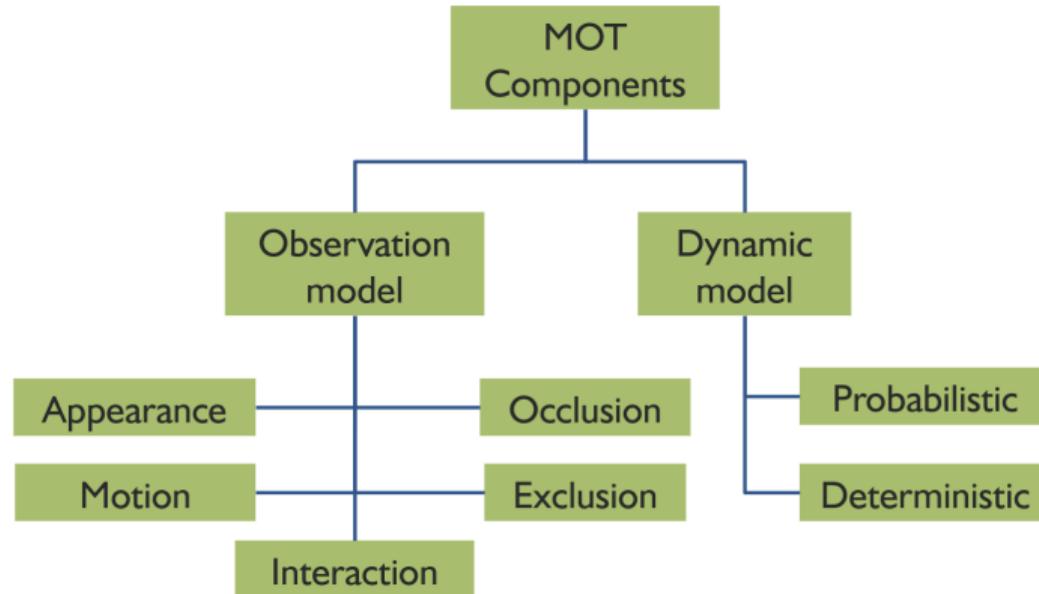


# Data association

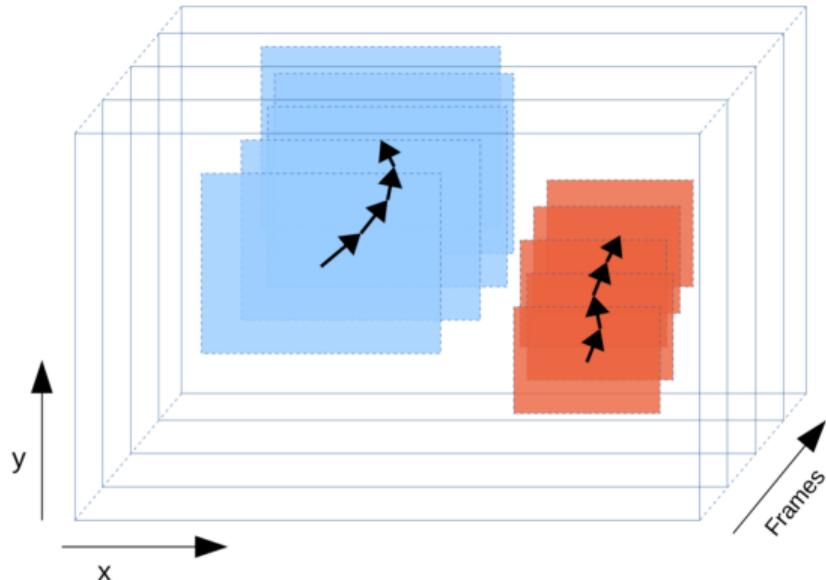


two-frame vs multi-frame methods

# Affinity function

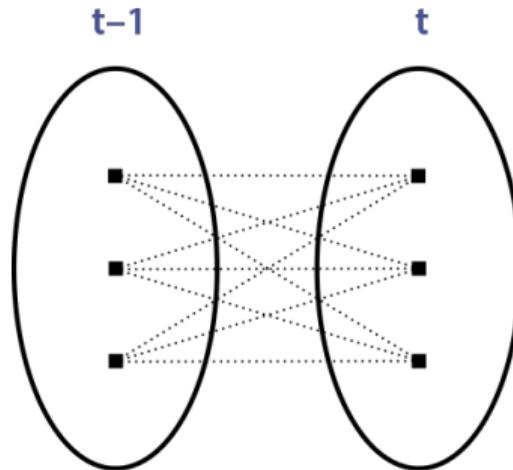


# Association with IoU



# Simple Online and Realtime Tracking (SORT)

- CNN-based object detector
- Kalman filter for predicting object position in current frame based on positions in previous frames
- Hungarian algorithm for matching object detections in current frames with predicted positions
- IoU of detected and predicted bounding boxes as affinity measure for matching detection and track



# SORT and IoU tracker comparison

	MOTA	MOTP	MT	ML	FP	FN	ID sw	Frag	Hz
SORT	<b>59.8</b>	<b>79.6</b>	<b>25.4%</b>	<b>22.7%</b>	8698	<b>63245</b>	<b>1423</b>	<b>1835</b>	59.5
IOU	57.1	77.1	23.6%	32.9%	<b>5702</b>	70278	2167	3028	<b>3004</b>

# Re-identification



Detections in video



Probe



Gallery



Matches

# DeepSORT and SORT comparison

	MOTA	MOTP	MT	ML	FP	FN	ID sw	Frag	Hz
SORT	59.8	79.6	25.4%	22.7%	<b>8698</b>	63245	1423	<b>1835</b>	<b>59.5</b>
Deep SORT	<b>61.4</b>	79.1	<b>32.8%</b>	<b>18.2%</b>	12852	<b>56668</b>	<b>781</b>	2008	40

Addition of re-identification to affinity function

# Outline

- I. Intro
2. Optical flow
3. Visual Object Tracking
4. Multiple Object Tracking
5. Action recognition

# Action recognition



Human actions are the main content of movies, TV news and shows, home video and video surveillance. Applications of action recognition:

- surveillance, abnormal situation detection
- video archive indexing and retrieval
- content navigation (automatical video timestamps)

# Actions



walking



jogging



running



boxing



waving



clapping

# Actions

Short meaningful movements



answer phone



handshake

# Actions → Events

A set of small actions with a specific common goal can still be called an “action” but we can also call them “events”



make sandwich



doing homework

# Events

An event can include a lot of different actions of different people



birthday party



parade

# UCF101



- 13320 videos from YouTube in 101 classes
- 5 groups: Human-Object Interaction; Body-Motion Only; Human-Human Interaction; Playing Musical Instruments; Sports
- quality is saturated (99% accuracy for SOTA methods)

# Kinetics



riding a bike



playing violin



braiding hair



dribbling basketball



riding unicycle



playing trumpet



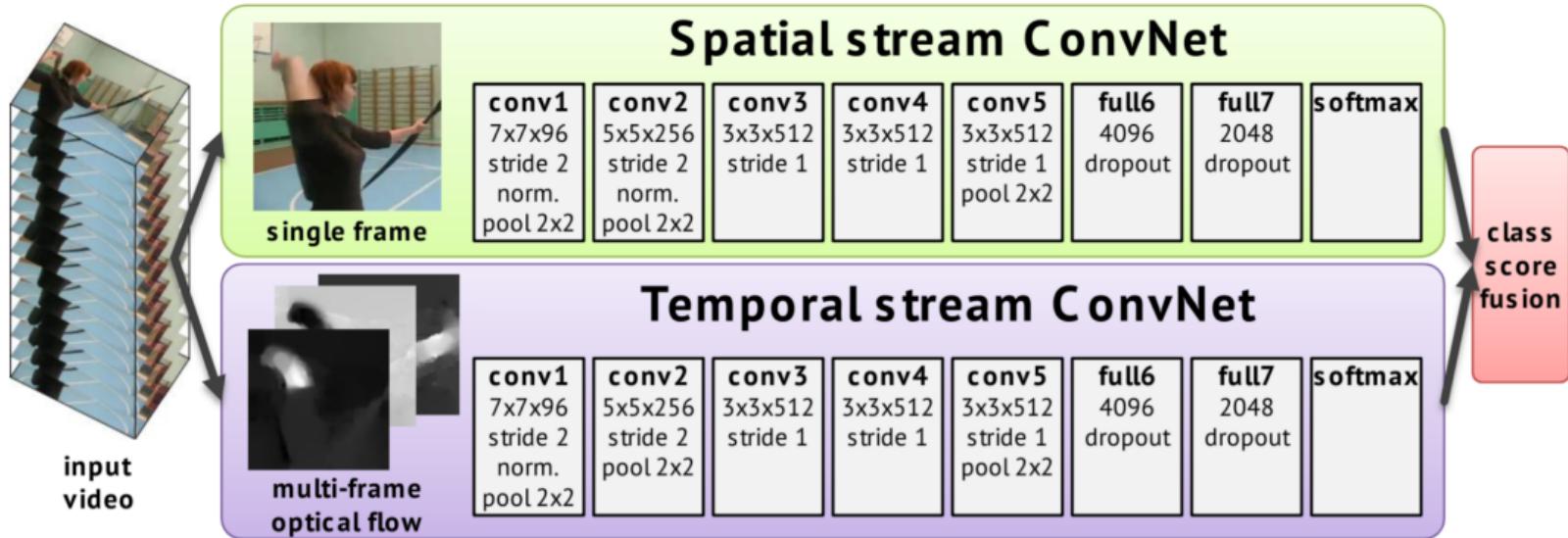
brushing hair



dunking basketball

- large video dataset scraped from YouTube
- 400, 600, 700 classes
- 300k, 650k, 700k videos

# Two-stream ConvNet



Simonyan, Zisserman. Two-stream convolutional networks for action recognition in videos. NIPS 2014

# Conclusion

We reviewed following topics:

- optical flow computation — an important low-level task for further video analysis
- visual tracking — general tracking method for a single object in video
- multiple object tracking that is used mostly for surveillance tasks
- action recognition — classification task for videos