Multiple object tracking

Prof. Laura Leal-Taixé

Technische Universität München

Dynamic Scene Understanding

Understand every pixel of a video

Semantic segmentation

Dynamic Scene Understanding

Understand every pixel of a video

Instance-based segmentation

Semantic segmentation

Dynamic Scene Understanding

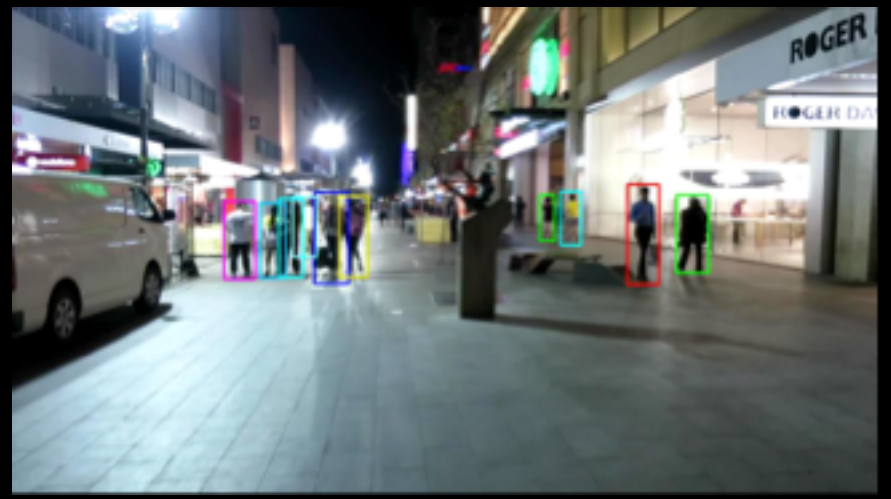
Understand every pixel of a video

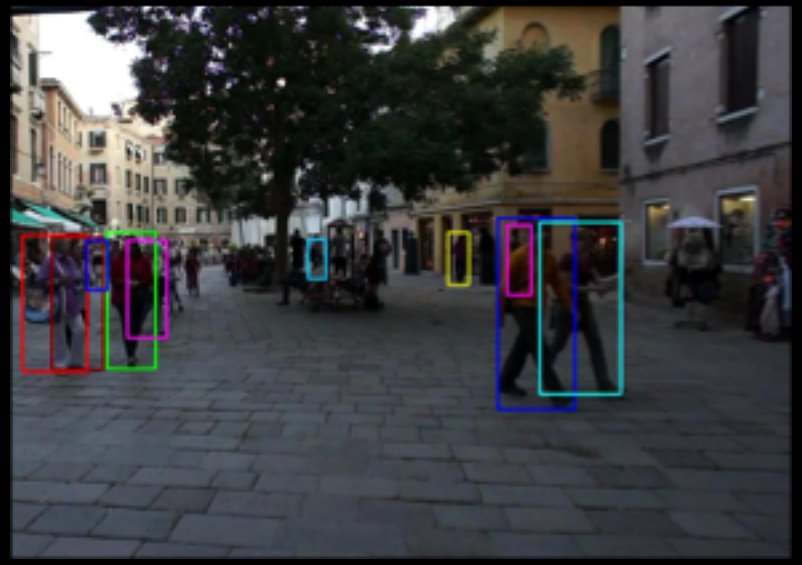
Multiple object tracking

Instance-based segmentation

Semantic segmentation

Multiple object tracking



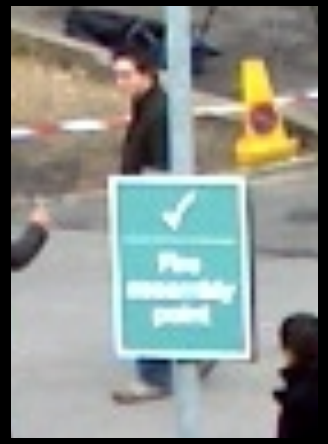


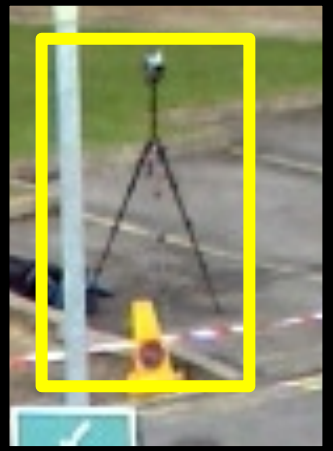
Goal: detect and track all objects in a scene

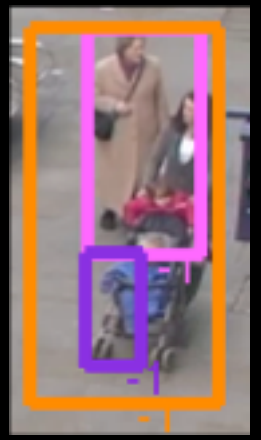
Tracking-by-detection

Video sequence

Person detection

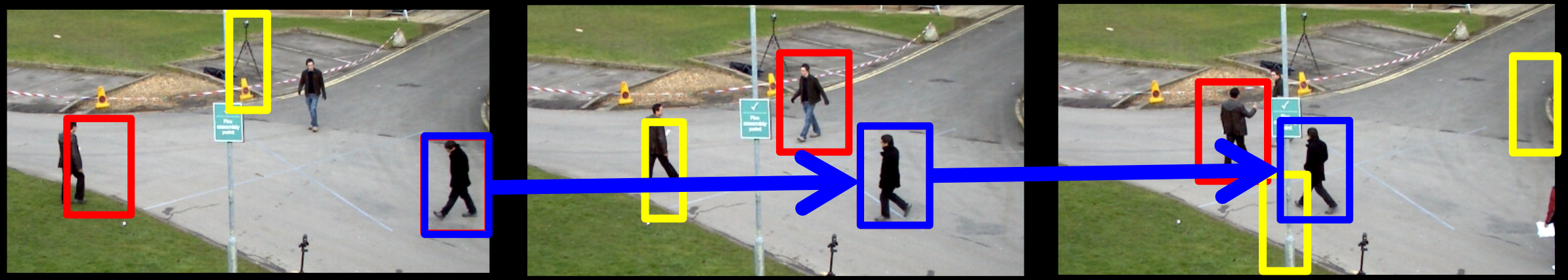




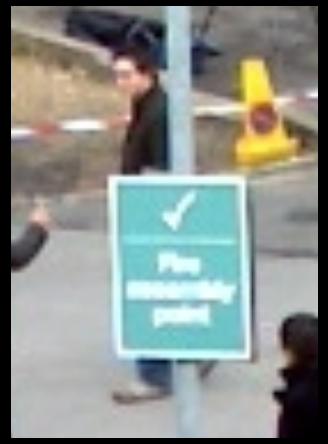


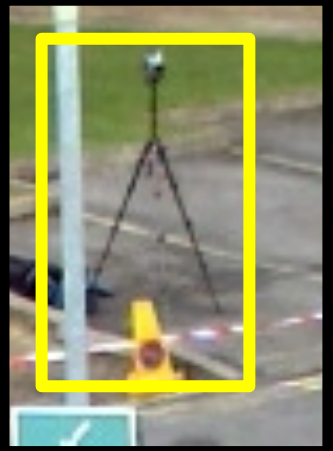
Tracking-by-detection

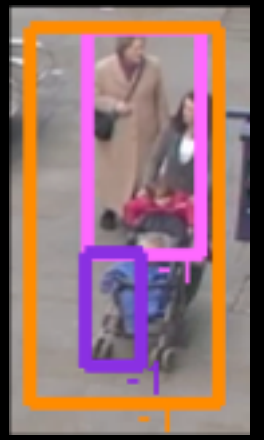
Video sequence



Person detection

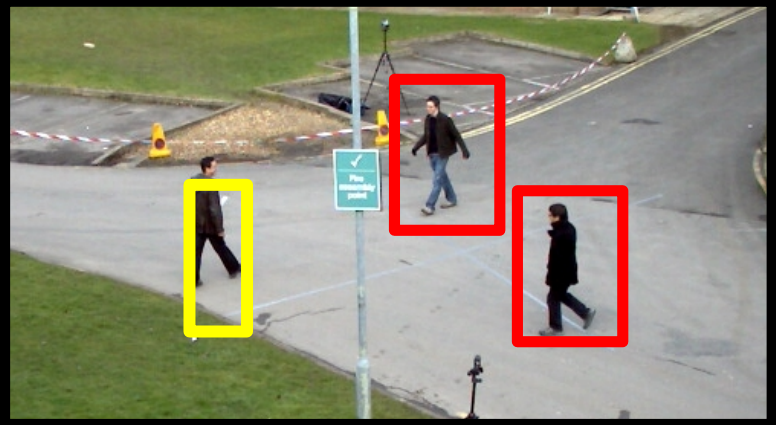




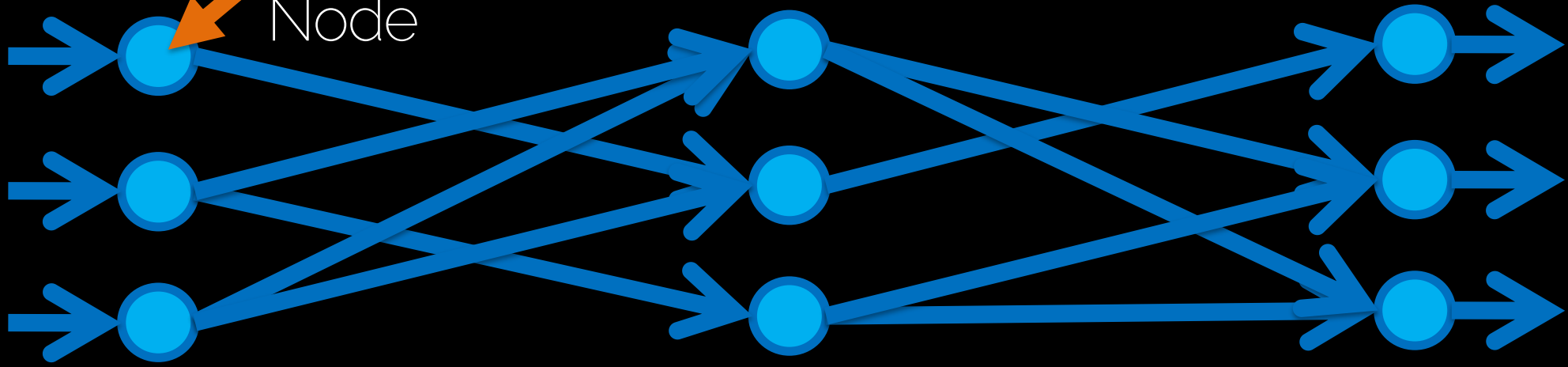


Tracking with network flows

Video sequence

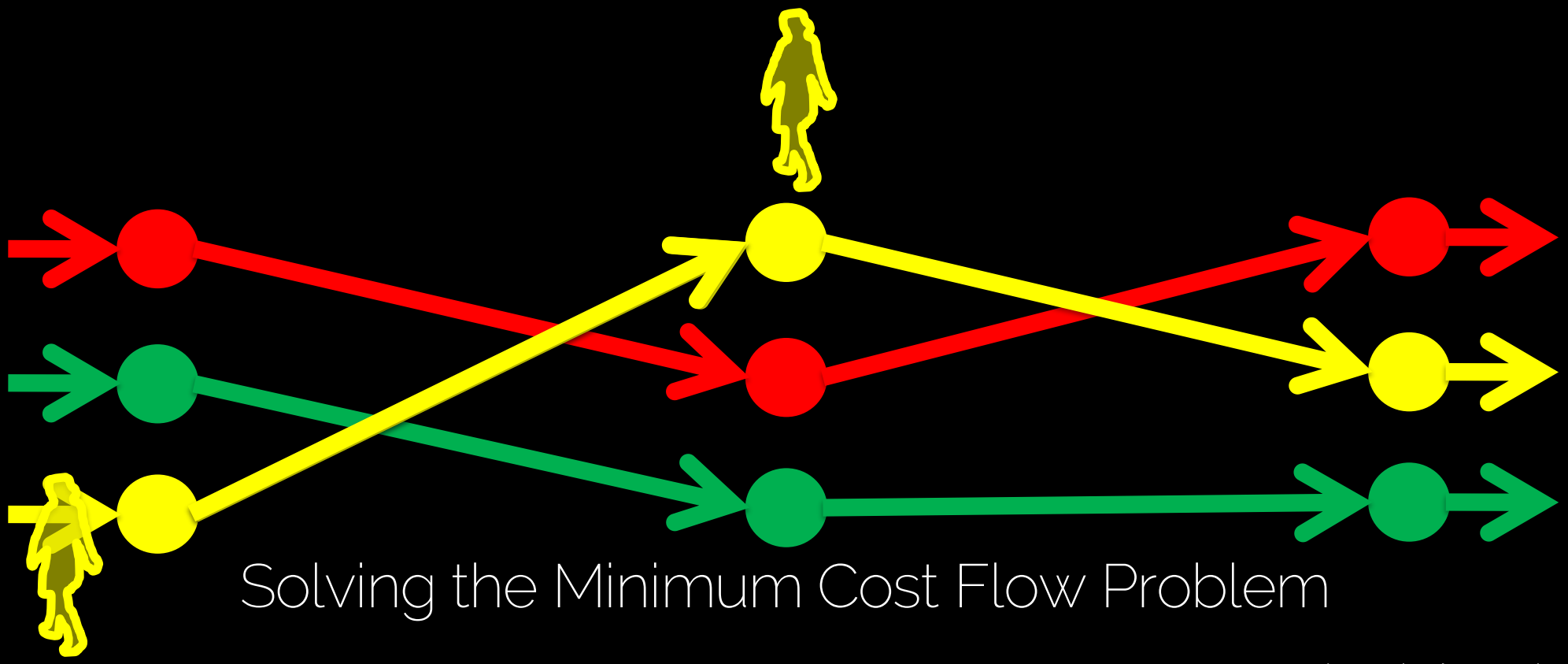


Node



Graphical model

Tracking with network flows



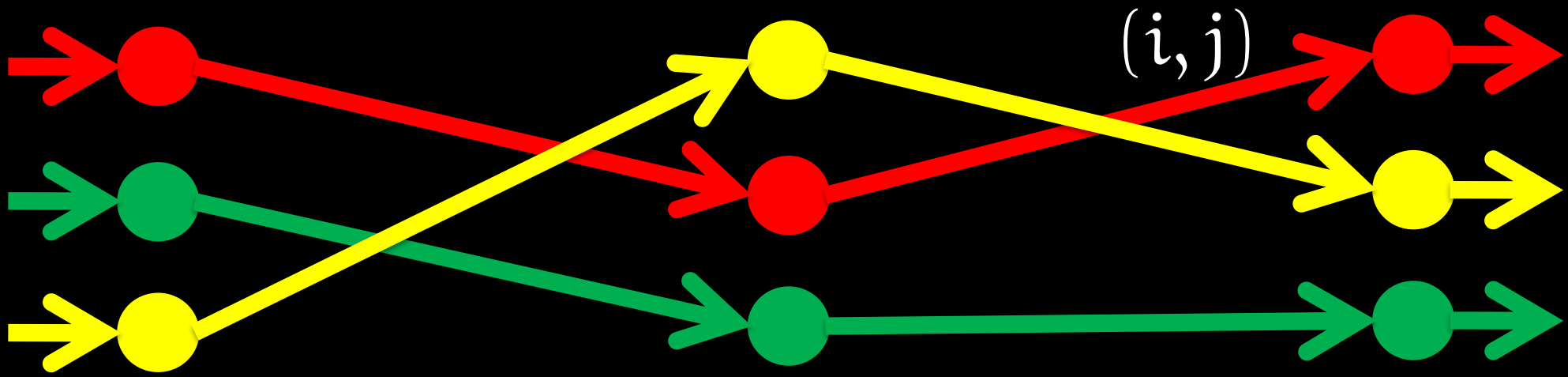
Solving the Minimum Cost Flow Problem

L. Leal-Taixé et al. ICCVW2011

Tracking with Linear Programming

- Objective function

$$\mathcal{T}* = \arg\min_{\mathcal{T}} \sum_{i,j} C(i,j) f(i,j)$$



Solving the Minimum Cost Flow Problem

L. Leal-Taixé et al. ICCVW2011

Tracking with Linear Programming

- Objective function

$$\mathcal{T}* = \arg\min_{\mathcal{T}} \sum_{i,j} C(i,j) f(i,j)$$

C

C

Costs – what will drive the tracking

Indicator {0,1}

Measuring bounding box similarity

Classic measures

- Bounding box distance
- Appearance similarity

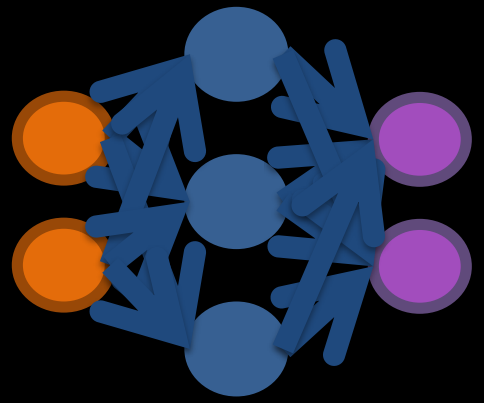
Proposed measures

Modeling pedestrian motion and interactions

Modeling pedestrian interaction

Using a physics-based model of crowd motion

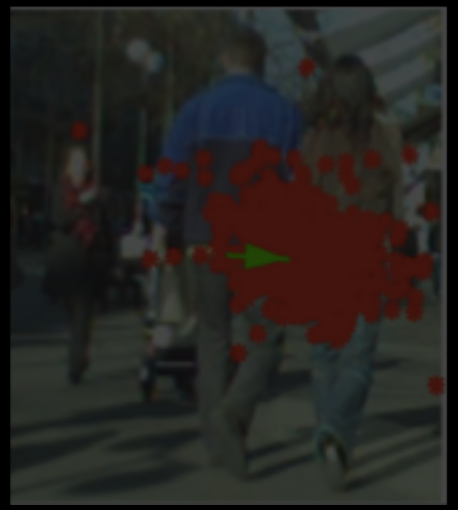
Learning an image-based motion context



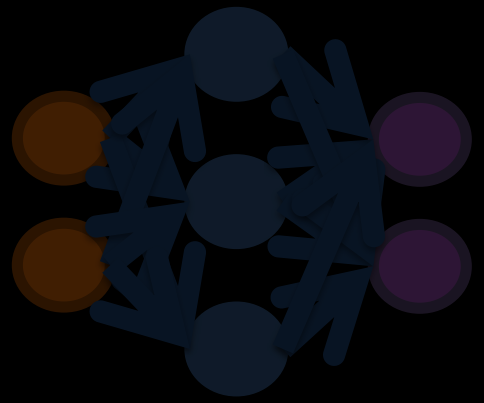
Learning appearance and interactions with Deep Learning

Modeling pedestrian interaction

Using a physics-based model of crowd motion



Learning an image-based motion context

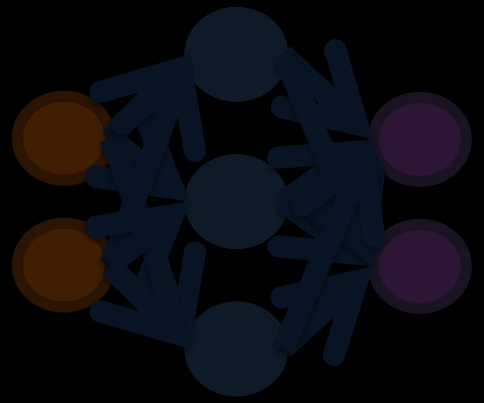


Learning appearance and interactions with Deep Learning

Modeling pedestrian interaction

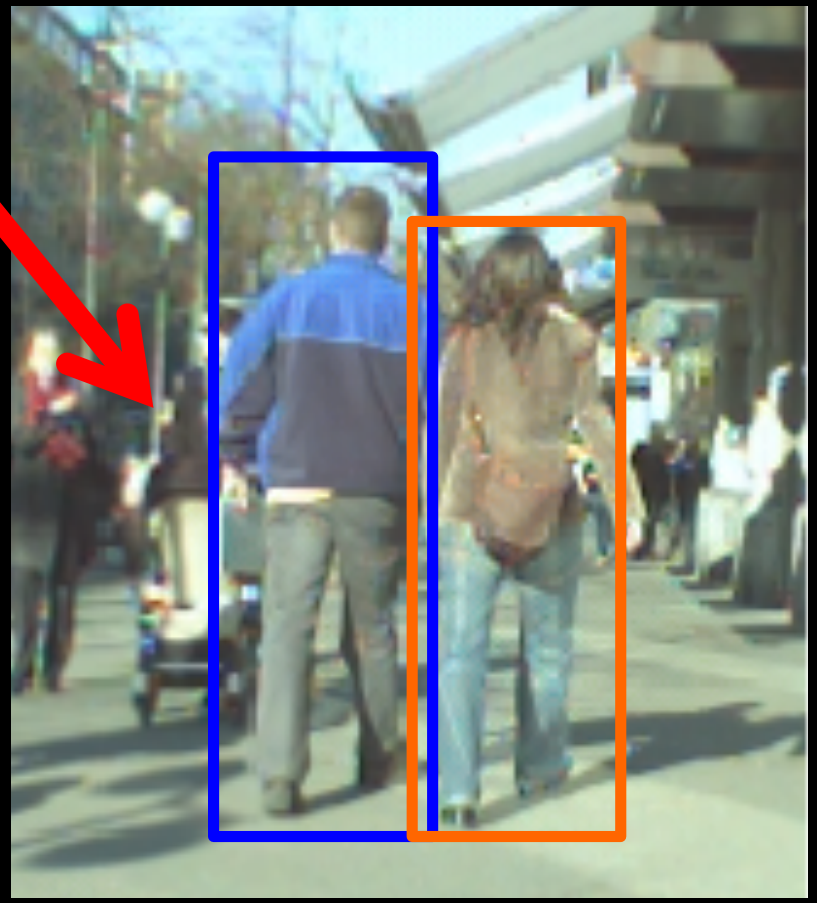
Using a physics-based model of crowd motion

Learning an image-based motion context



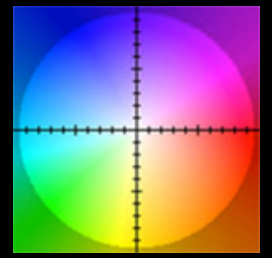
Learning appearance and interactions with Deep Learning

The effect of undetected pedestrians



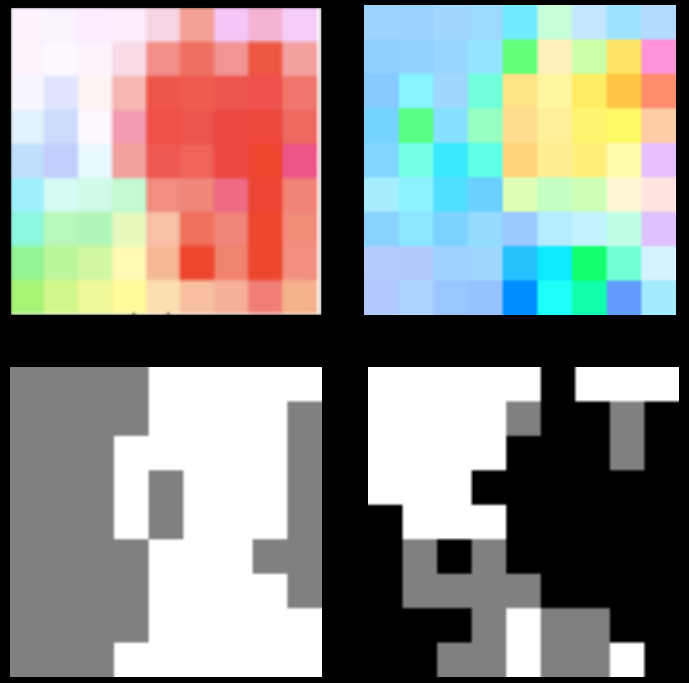
Image

Optical flow



Learning an image-based motion context

Image



Features

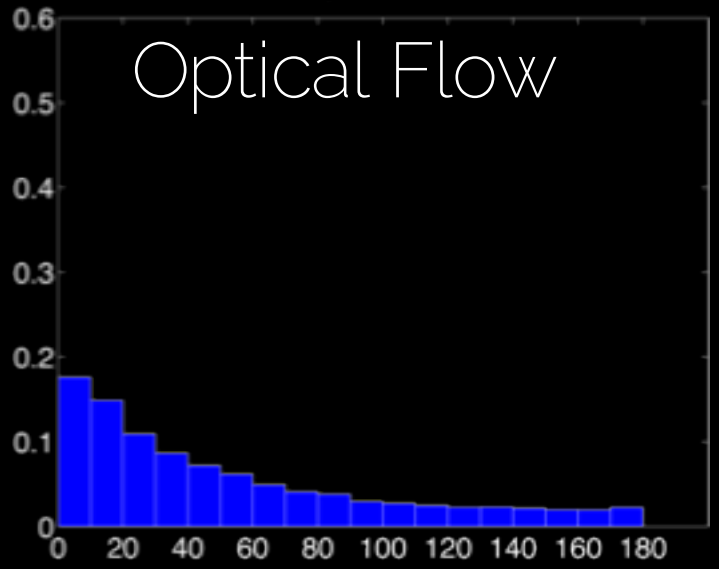
Hough forests

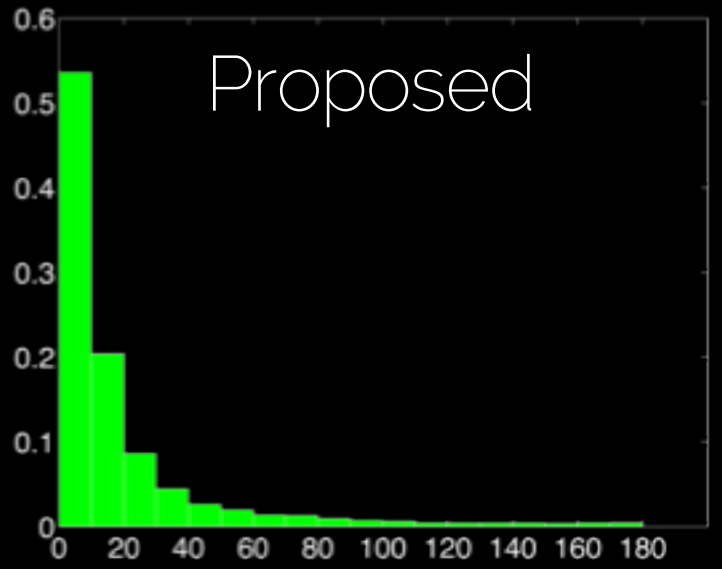
Estimation of the velocity

L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, S. Savarese. CVPR 2014.

Pedestrian velocity estimation

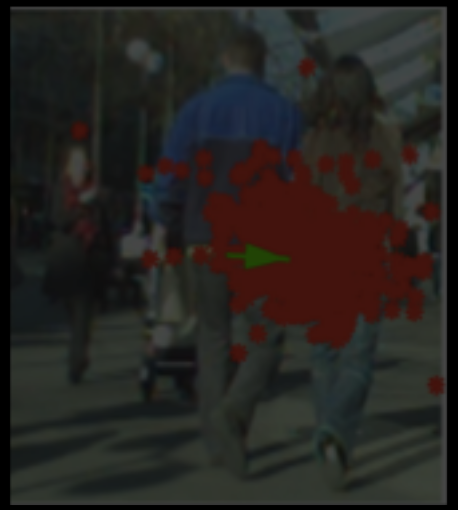
Histogram
of the errors



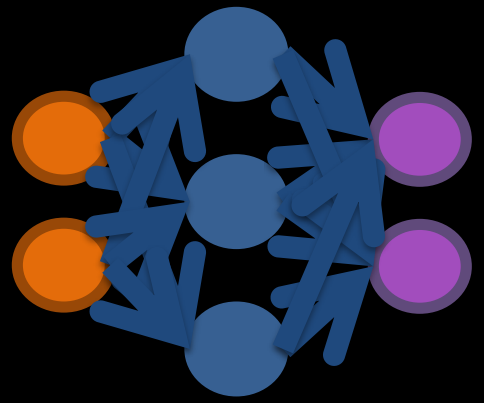


Modeling pedestrian interaction

Using a physics-based model of crowd motion

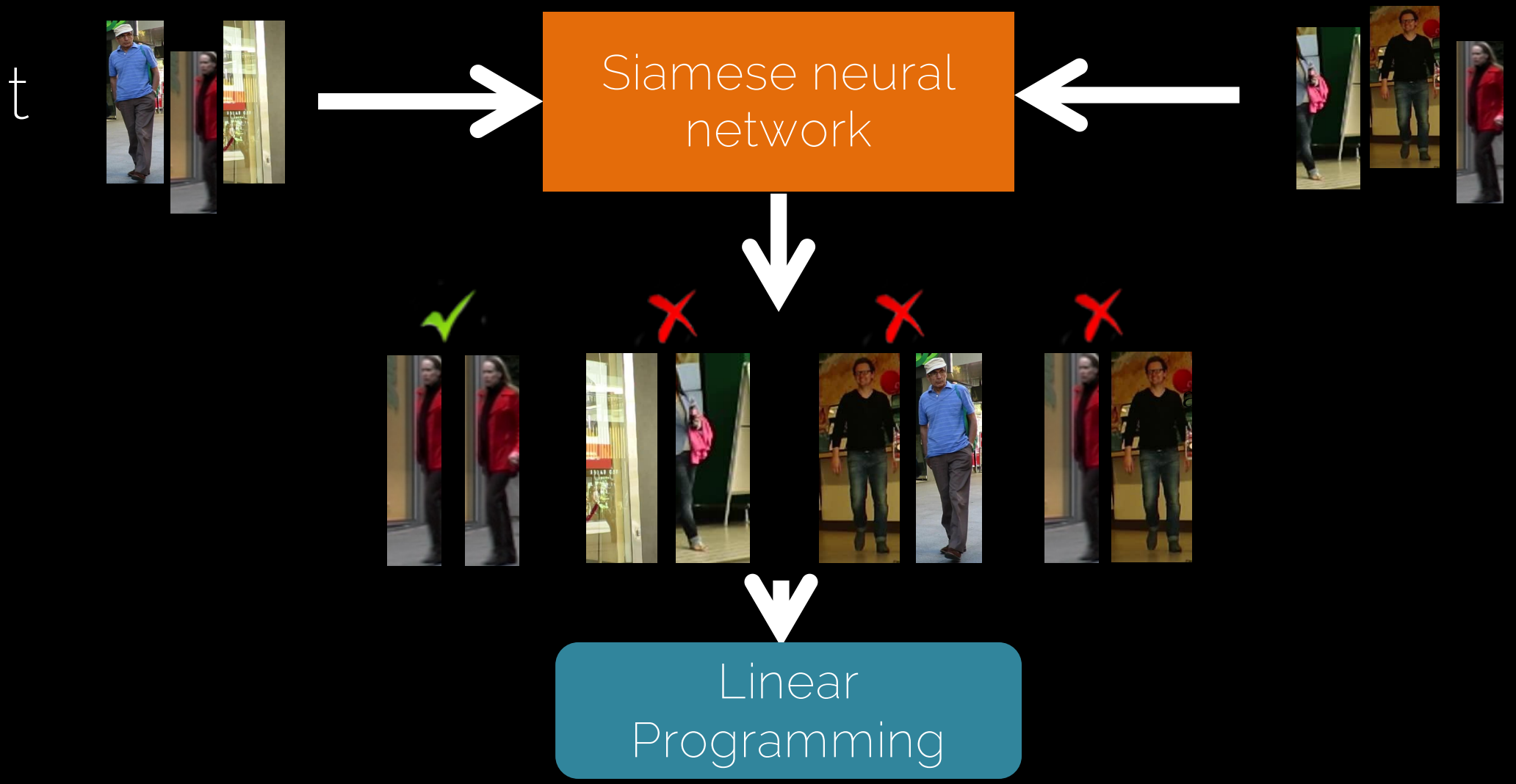


Learning an image-based motion context



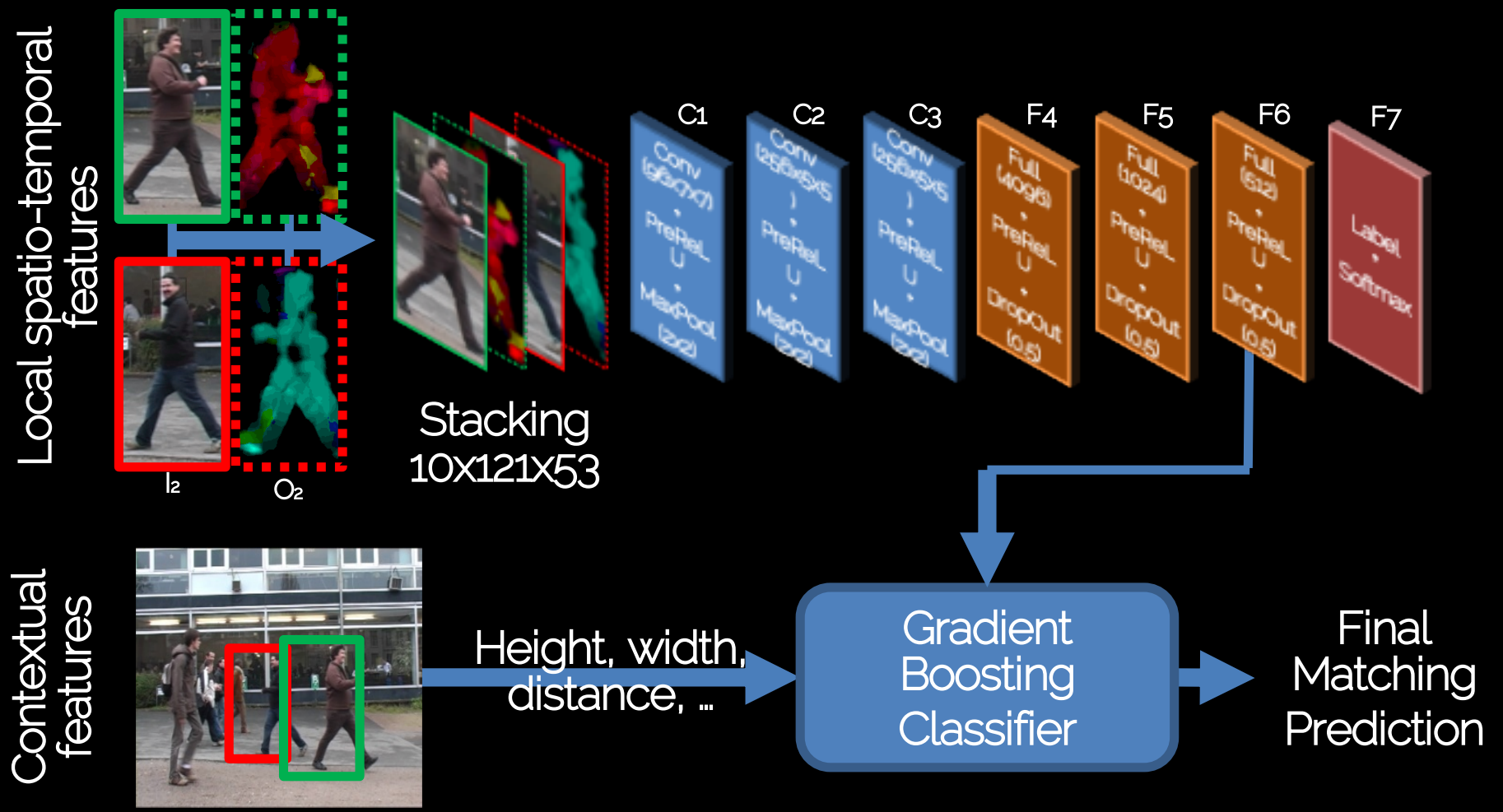
Learning appearance and interactions with Deep Learning

Learning appearance and interactions



L. Leal-Taixé, C. Canton, K. Schindler. CVPRW 2016

Design


Local spatio-temporal features
I2
O2
Stacking
10x121x53
C1
C2
C3
F4
F5
F6
F7
Label
Softmax
Contextual features
Height, width, distance, ...
Gradient Boosting Classifier
Final Matching Prediction

Improvement directions

- Better appearance models
  - Ristani & Tomasi. Features for Multi-Target Multi-Camera Tracking and Re-Identification. CVPR 2018

- Multi-detector fusion
  - Henschel et al. Fusion of Head and Full-Body Detectors for Multi-Object Tracking. CVPRW 2018.

Can we learn it all?

- Problem 1: dimensionality of the output
  - Neural networks can handle fixed sized outputs (tensors)
  - Tracking:
    - Unknown number of detections per frame
    - Unknown length of each trajectory

- Solution 1: see the Set Learning lecture

- Solution 2: recursive

Can we learn it all?

- Solution 2: recursive
  - Using a Recurrent Neural Network to predict trajectory properties
  - Birth/death, transition → motion model

- Does not work well….

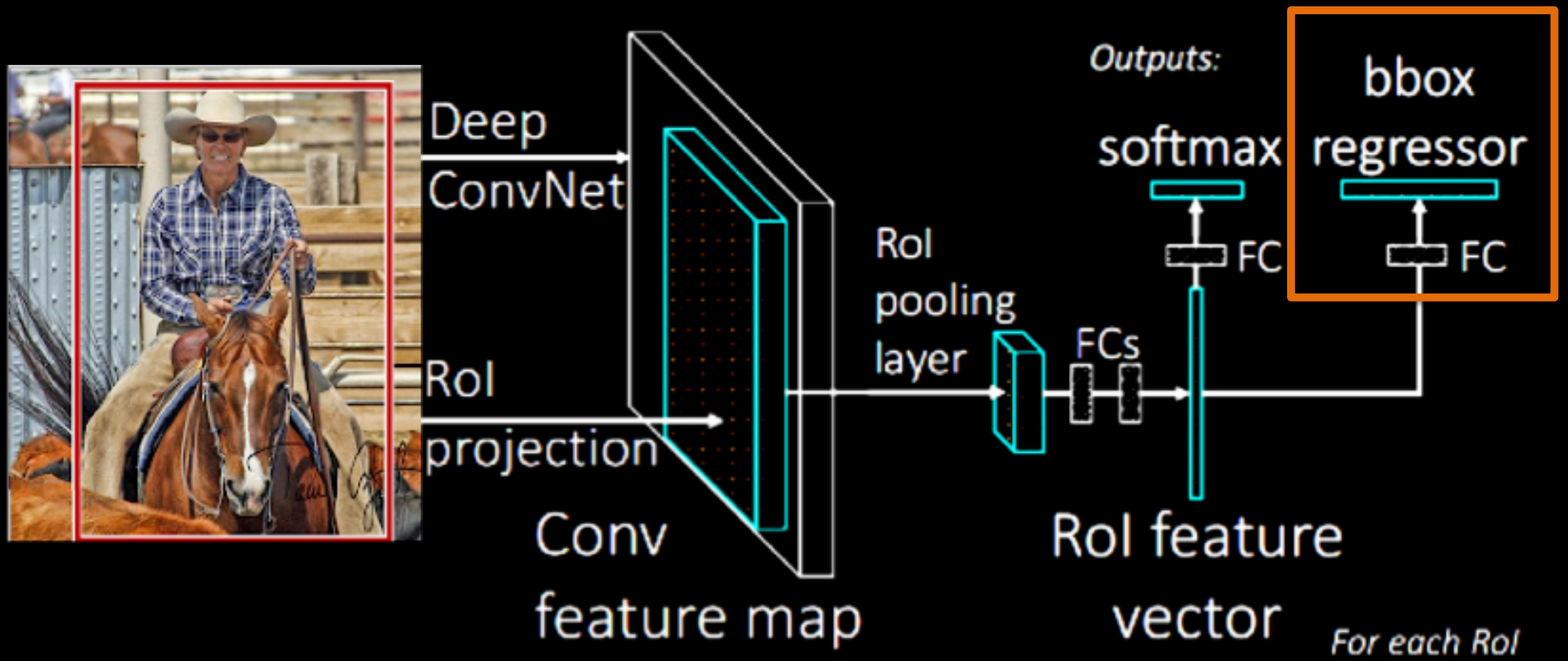
- Is it likely that more data is needed to train such a motion model?

Milan et al. Online Multi-Target Tracking Using Recurrent Neural Networks. AAAI 2017

Can we learn it all?

- Our Solution 2: recursive
  - One trajectory at a time, one frame at a time

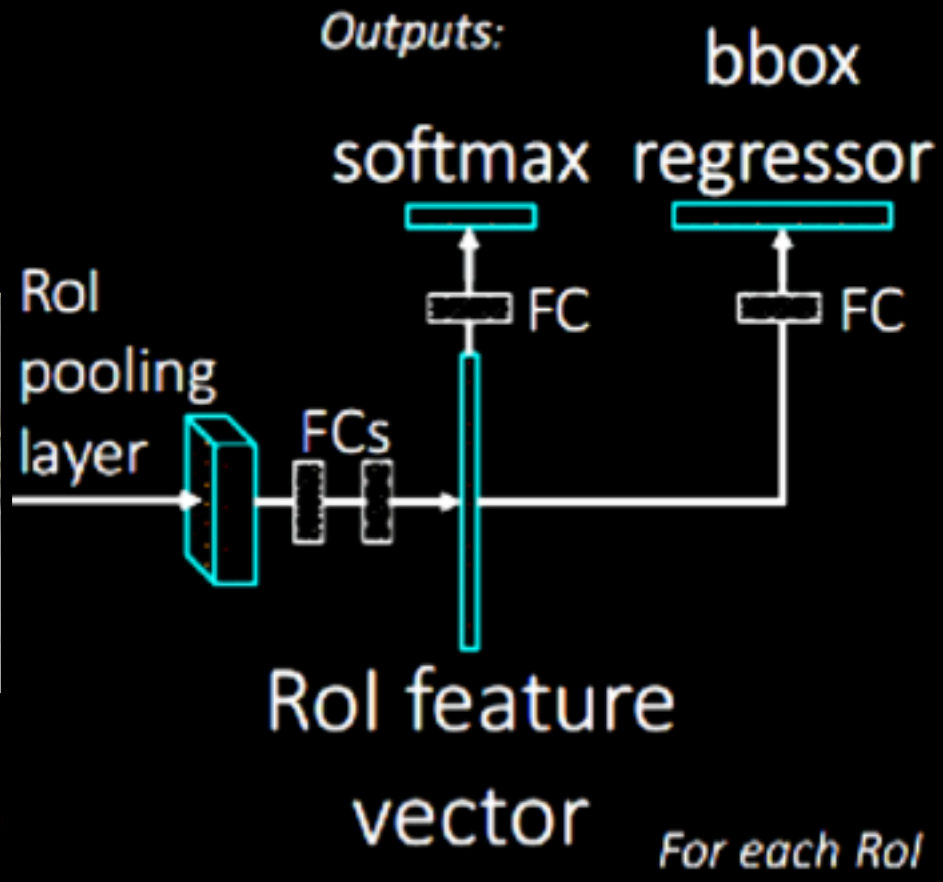
Can we learn it all?



Girshick Fast R-CNN. ICCV 2015

Can we learn it all?

- We use the detections from the last frame as proposals



- Where did the detection with ID 1 go in the next frame?

P. Bergmann. Master Thesis. TUM 2018

Pros and cons

- PRO We can reuse an extremely well-trained regressor
  - We get well-positioned bounding boxes

- CON The regressor only shifts the box by a small quantity
  - We need to compensate for large camera motions → optical flow

P. Bergmann. Master Thesis. TUM 2018

Surprisingly good results

- Extremely good MOTA improvement

Tracker	Avg Rank	↑MOTA	IDF1	MT	ML	FP	FN	ID Sw.	Frag	Hz	Detector
REG_FRCNN	18.5	53.6 ±13.6	52.8	19.0%	36.6%	5,217	78,471	909 (16.0)	1,742 (30.6)	3.0	Public
1.											Anonymous submission
TPM	21.4	49.1 ±9.1	46.9	20.0%	38.9%	9,038	83,031	679 (12.5)	850 (15.6)	0.8	Public
2.											Anonymous submission
AFN	18.7	49.0 ±10.2	48.2	19.1%	35.7%	9,508	82,506	899 (16.4)	1,383 (25.3)	0.6	Public
3.											Anonymous submission
KCF16	23.1	48.8 ±9.6	47.2	15.8%	38.1%	5,875	86,567	906 (17.3)	1,116 (21.2)	0.1	Public
4.											Paper ID 2988
LMP	15.2	48.8 ±9.8	51.3	18.2%	40.1%	6,654	86,245	481 (9.1)	595 (11.3)	0.5	Public
5.											S. Tang, M. Andriluka, B. Andres, B. Schiele. Multiple People Tracking with Lifted Multicut and Person Re-identification. In CVPR, 2017.
GCRA	19.9	48.2 ±8.3	48.6	12.9%	41.1%	5,104	88,586	821 (16.0)	1,117 (21.7)	2.8	Public
7.											C.Ma, C.Yang, F.Yang, Y.Zhuang, Z.Zhang, H.Jia, D.Xie. Trajectory Factory: Tracklet Cleaving and Re-connection by Deep Siamese Bi-GRU for Multiple Object Tracking. In ICME 2018.
FWT	22.3	47.8 ±9.4	44.3	19.1%	38.2%	8,886	85,487	852 (16.0)	1,534 (28.9)	0.6	Public
8.											R. Henschel, L. Leal-Taixé, D. Cremers, B. Rosenhahn. Fusion of Head and Full-Body Detectors for Multi-Object Tracking. In Trajnet CVPRW, 2018.

Surprisingly good results

- Quite some ID switches but still good identification overall

Tracker	Avg Rank	↑MOTA		IDF1	MT	ML	FP	FN	ID Sw.	Frag	Hz	Detector
REG_FRCNN	18.5	**53.6**	±13.6	52.8	19.0%	36.6%	5,217	78,471	909 (16.0)	1,742 (30.6)	3.0	Public
1.												Anonymous submission
TPM	21.4	49.1	±9.1	46.9	20.0%	38.9%	9,038	83,031	679 (12.5)	850 (15.6)	0.8	Public
2.												Anonymous submission
AFN	18.7	49.0	±10.2	48.2	19.1%	35.7%	9,508	82,506	899 (16.4)	1,383 (25.3)	0.6	Public
3.												Anonymous submission
KCF16	23.1	48.8	±9.6	47.2	15.8%	38.1%	5,875	86,567	906 (17.3)	1,116 (21.2)	0.1	Public
4.												Paper ID 2988
LMP	15.2	48.8	±9.8	51.3	18.2%	40.1%	6,654	86,245	481 (9.1)	595 (11.3)	0.5	Public
5.												S. Tang, M. Andriluka, B. Andres, B. Schiele. Multiple People Tracking with Lifted Multicut and Person Re-identification. In CVPR, 2017.
GCRA	19.9	48.2	±8.3	48.6	12.9%	41.1%	5,104	88,586	821 (16.0)	1,117 (21.7)	2.8	Public
7.												C.Ma, C.Yang, F.Yan, Y.Zhuang, Z.Zhang, H.Jia, D.Xie. Trajectory Factory: Tracklet Cleaving and Re-connection by Deep Siamese Bi-GRU for Multiple Object Tracking. In ICME 2018.
FWT	22.3	47.8	±9.4	44.3	19.1%	38.2%	8,886	85,487	852 (16.0)	1,534 (28.9)	0.6	Public
8.												R. Henschel, L. Leal-Taixé, D. Cremers, B. Rosenhahn. Fusion of Head and Full-Body Detectors for Multi-Object Tracking. In Trajnet CVPRW, 2018.

Conclusions

- The old assumption that objects have small displacements still holds and can be used to reach SOA performance

- What is next? Hard cases.
  - Long occlusions
  - Crowded scenes

Questions?

Prof. Laura Leal-Taixé

Technische Universität München