Master in Computer Vision Barcelona

[http://pagines.uab.cat/mcv/]

Xavier Giro-i-Nieto

@DocXavi

xavier.giro@upc.edu

Associate Professor

Universitat Politècnica de Catalunya

Module 6 - Day 9 - Lecture 3

Video Object Tracking

5th April 2022

Acknowledgements

Andreu Girbau

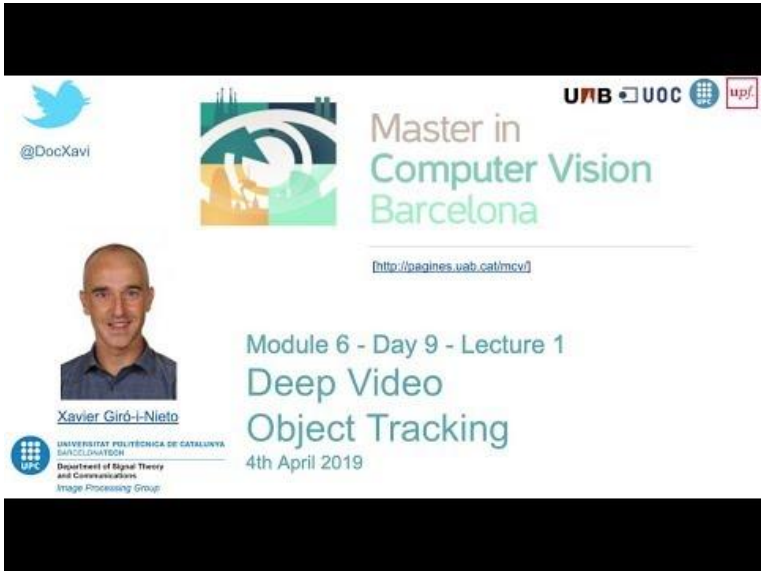
Andreu Girbau
andreu.girbau@upc.edu

PhD Candidate
Universitat Politècnica de Catalunya
AutomaticTV

Other Videos & Slides online

Laura Leal-Taixé
[Deep Learning for Computer Vision 2018]
UPC TelecomBCN, Barcelona

[slides]



Xavier Giro-i-Nieto
[Master in Computer Vision Barcelona 2019]
UPC TelecomBCN, Barcelona

[slides]

Outline

- **Motivation**
- First steps with NN/DNN
- Correlation filters
- Box regressors
- Tracking by detection
- Tracking with Language

Object Tracking

Usain Bolt

gif source

Challenge:

Assign a **unique ID** to objects across a video.

Set up:

The object initial location is known.

#DeepMOT Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixe, Xavier Alameda-Pineda, "How To Train Your Deep Multi-Object Tracker" CVPR 2020. [code]

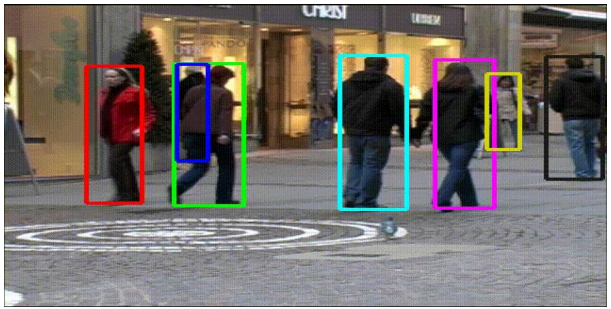
Single vs Multiple Object Tracking

Single object tracking (SOT)

Usain Bolt

gif source

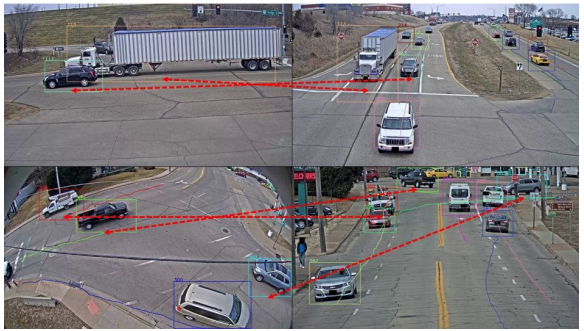
Multiple Object Tracking (MOT)



(P1, P2, P3, P4 ...)

video source

Multi-Target Multi-Camera Tracking (MTMC)

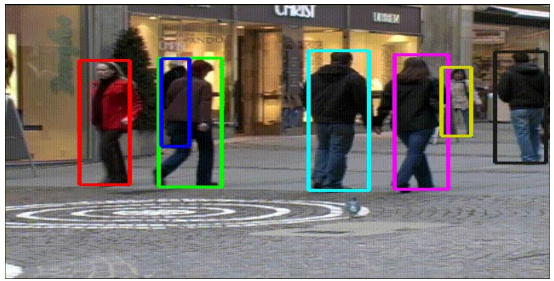


Source:
AI City Challenge 2019

Single object tracking vs multiple object tracking

1. SOT does not take into account the appearance of the rest of the objects in the scene, making it difficult for object disambiguation.

2. SOT is computationally expensive (1 pass through all the video per object to track).



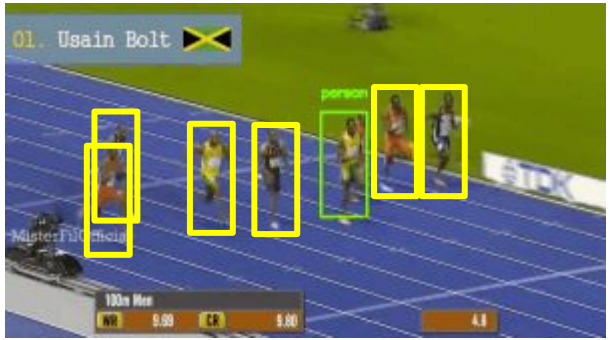
Some tracking concepts

Tracklet

Tracklet

A fragment of the track followed by a moving object.

Distractors



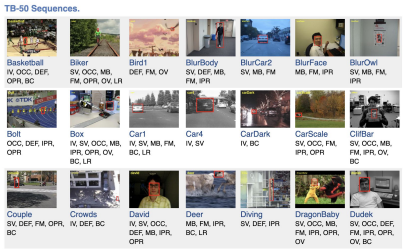
Salient objects other than the object of interest.

Set up

Tracking: Mostly, object detection as a first step and tracking as post-processing.

Objective: Infer a tracklet over multiple frames by simultaneous detection and tracking.

Single object Tracking: Datasets



Online Tracking Benchmark

100 videos

Visual Object tracking

100 videos (2018)
Every year holding a VOT challenge with new data

TrackingNet

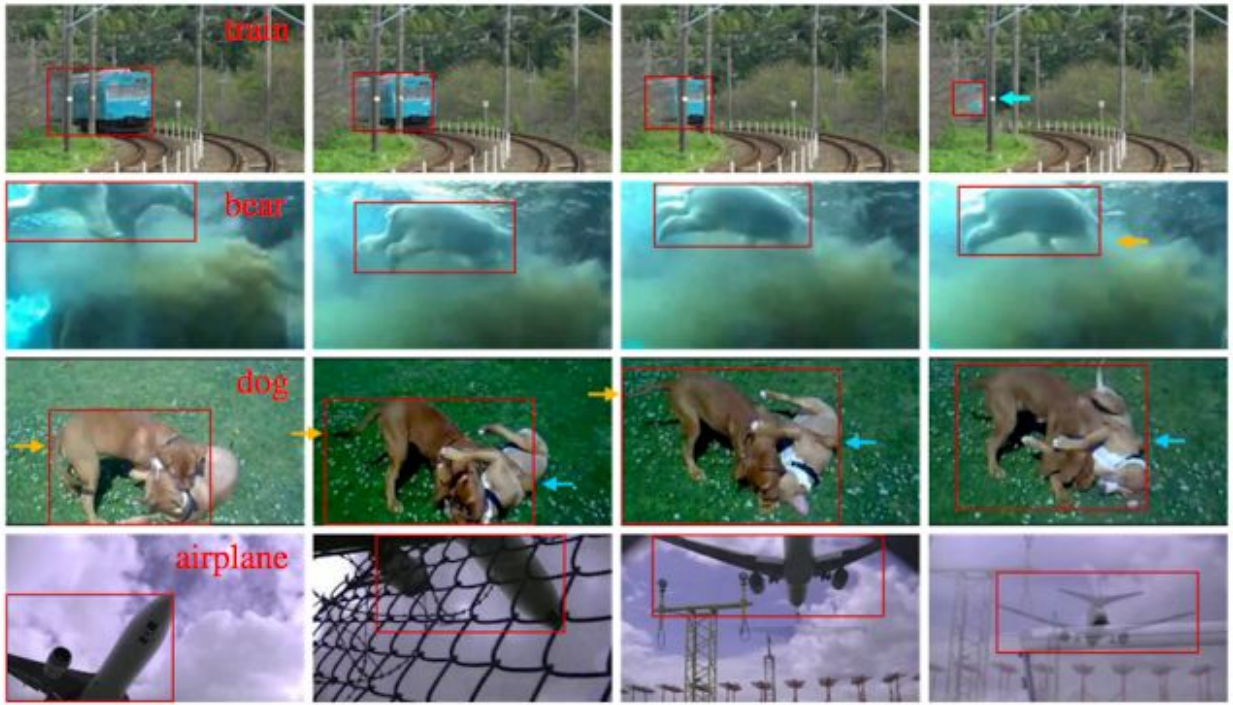
30k training videos
500 test videos
14M train bounding boxes
225k test bounding boxes

Datasets: ImageNet VID

Video Object Detection = Intra-frame Localization + Inter-frame tracking

Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015

Datasets: YouTube-BB



Real, Esteban, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. "Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video." CVPR 2017.

Datasets: MOT

Release	# Seq.	BBs	Persons	Length	Density	Tracks	HD
MOT15	22	101349	101349	16:36	8.98	1221	31.8%
MOT16	14	476532	292733	07:43	26.05	1276	85.7%

Leal-Taixé, Laura, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, and Stefan Roth. "Tracking the trackers: an analysis of the state of the art in multiple object tracking." arXiv (2017).

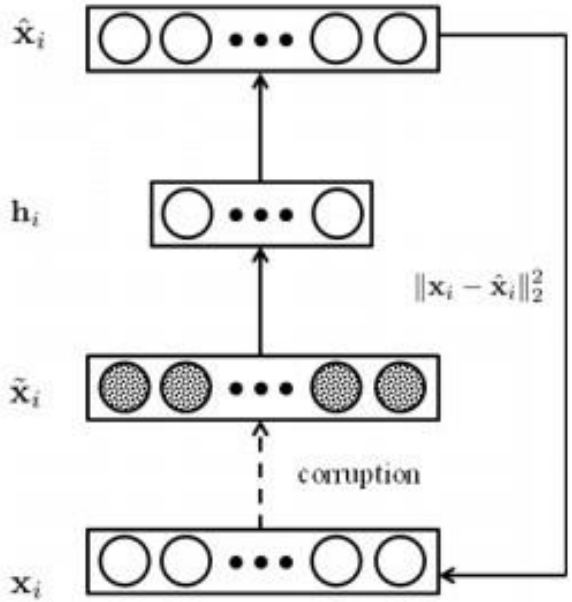
Milan, Anton, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. "MOT16: A benchmark for multi-object tracking." arXiv (2016).

Outline

- Motivation
- **First steps with NN/DNN**
- Correlation filters
- Box regressors
- Tracking by detection
- Tracking with Language

Feature Learning: MLP + Particle Filter

Appearance features can be learned with a denoising MLP autoencoder...

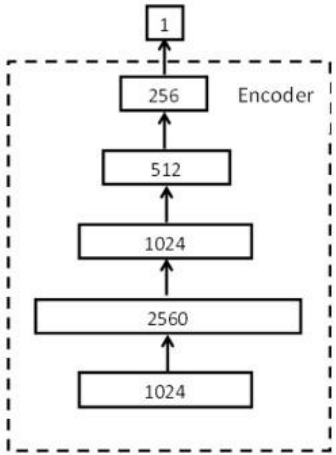


Artificial noise (“corruption”) is added to force the autoencoder to be robust to it.

Wang, Naiyan, and Dit-Yan Yeung. "Learning a deep compact image representation for visual tracking." NIPS 2013
[Project page with code]

Feature Learning: MLP + Particle Filter

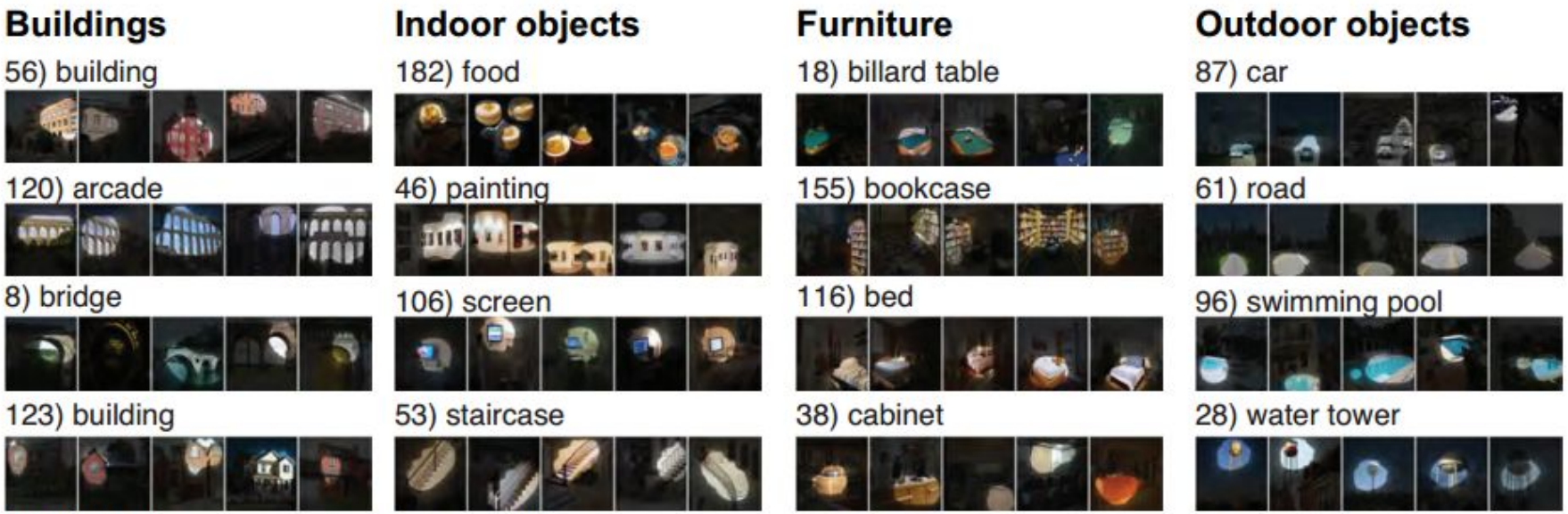
...the encoder part is used as feature extractor to train a binary classifier used in a particles filter.



Wang, Naiyan, and Dit-Yan Yeung. "Learning a deep compact image representation for visual tracking." NIPS 2013
[Project page with code]

Off-the-shelf convs as weak trackers

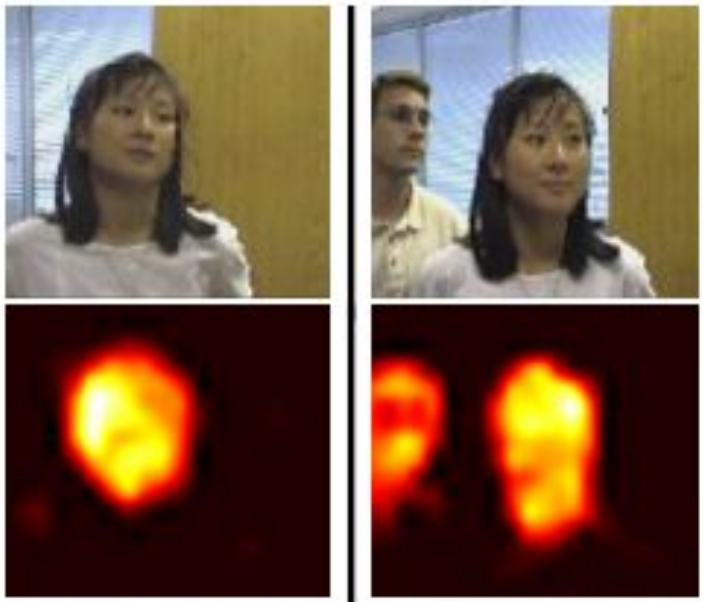
Previous work: Off-the-shelff CNN filters as weak object detectors.



Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Object detectors emerge in deep scene CNNs." ICLR 2015.

Off-the-shelf convs as weak trackers

Similarly to Zhou (ICLR 2016), in this work they identified feature maps in **conv5-3** as weak object detectors… but were not discriminative enough to discriminate between instances of the same class.

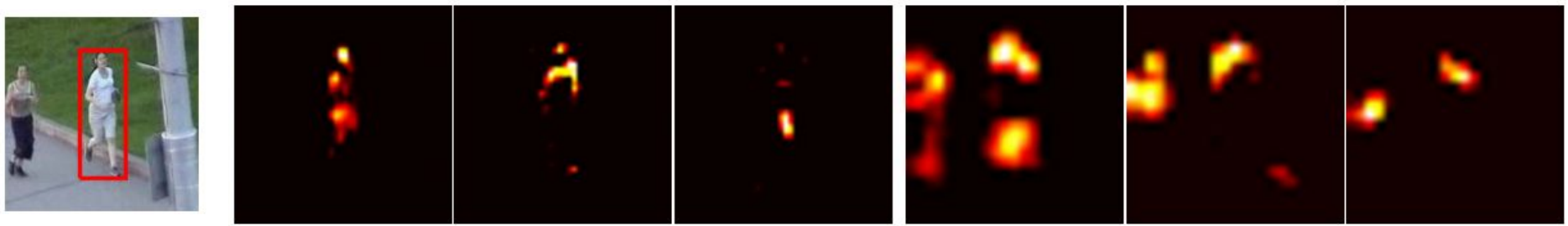


#FCNT Wang, Lijun, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. "Visual Tracking with Fully Convolutional Networks." ICCV 2015. [code]

Off-the-shelf convs as weak trackers

On the other hand, feature maps from **conv4-3** are more sensitive to intra-class appearance variation…

conv4-3 (specific) conv5-3 (general)



#FCNT Wang, Lijun, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. "Visual Tracking with Fully Convolutional Networks." ICCV 2015. [code]

Off-the-shelf convs as weak trackers + Online learning

SNet=Specific Network (**online learning**)

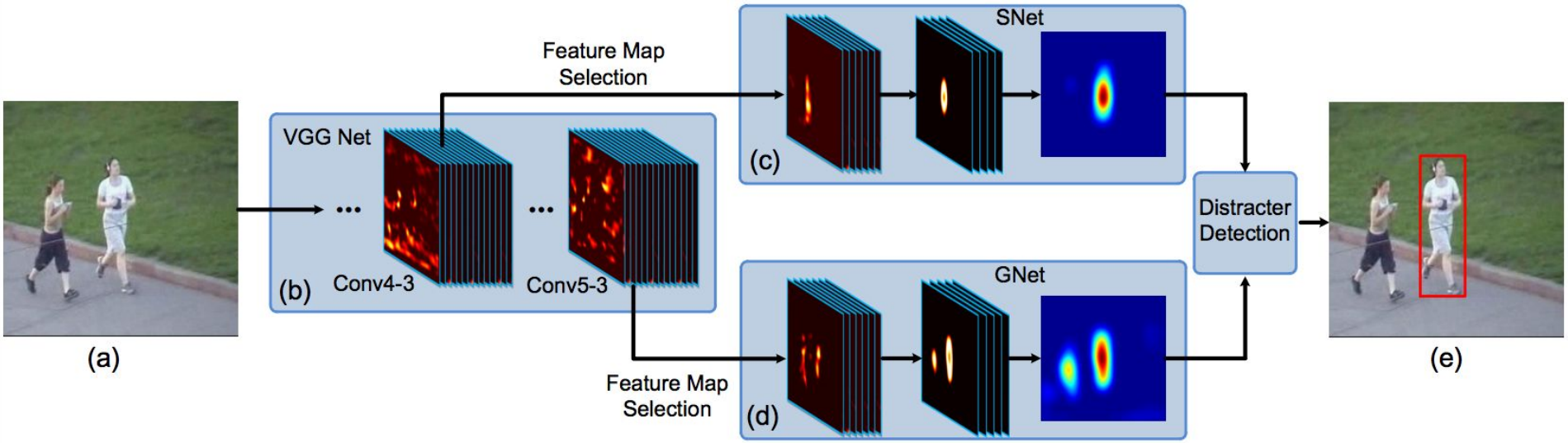


Figure 5. Pipeline of our algorithm. (a) Input ROI region. (b) VGG network. (c) SNet. (d) GNet. (e) Tracking results.

GNet=General Network (fixed)

#FCNT Wang, Lijun, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. "Visual Tracking with Fully Convolutional Networks." ICCV 2015. [code]

Outline

- Motivation
- First steps with NN/DNN
- **Correlation filters**
- Box regressors
- Tracking by detection
- Tracking with Language

Correlation Filters

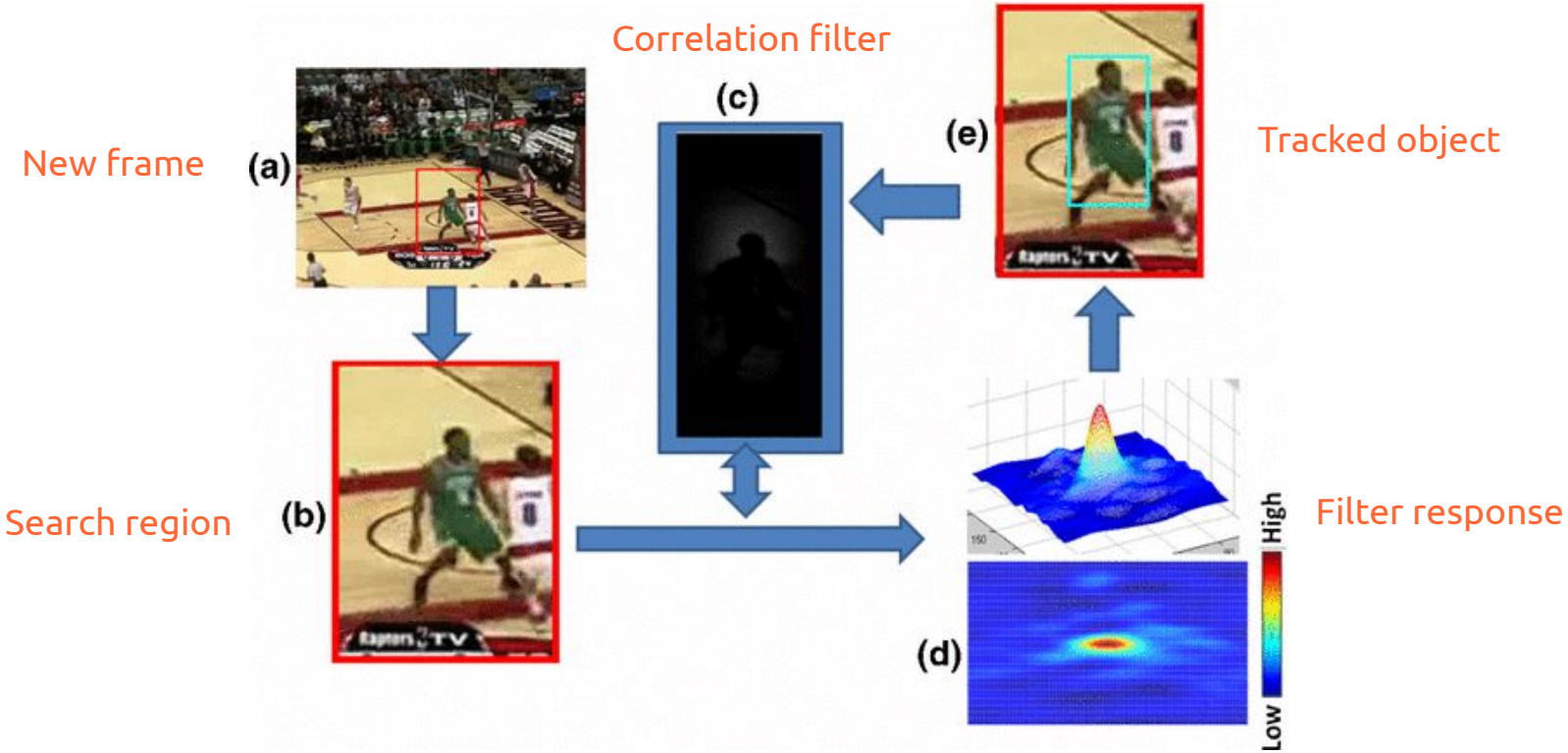
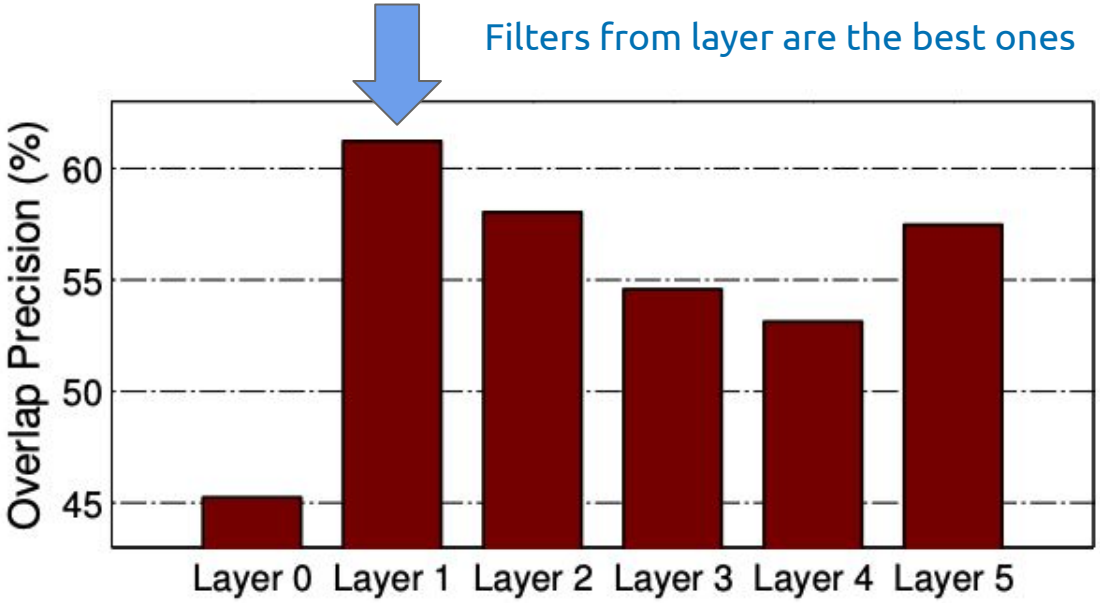


Figure: Han, Ke. "Image object tracking based on temporal context and MOSSE." Cluster Computing 20, no. 2 (2017): 1259-1269.

Bolme, David S., J. Ross Beveridge, Bruce A. Draper, and Yui Man Lui. "Visual object tracking using adaptive correlation filters." CVPR 2010.
Henriques, João F., Rui Caseiro, Pedro Martins, and Jorge Batista. "High-speed tracking with kernelized correlation filters." TPAMI 2014.

Correlation Filters: Off-the-shelf

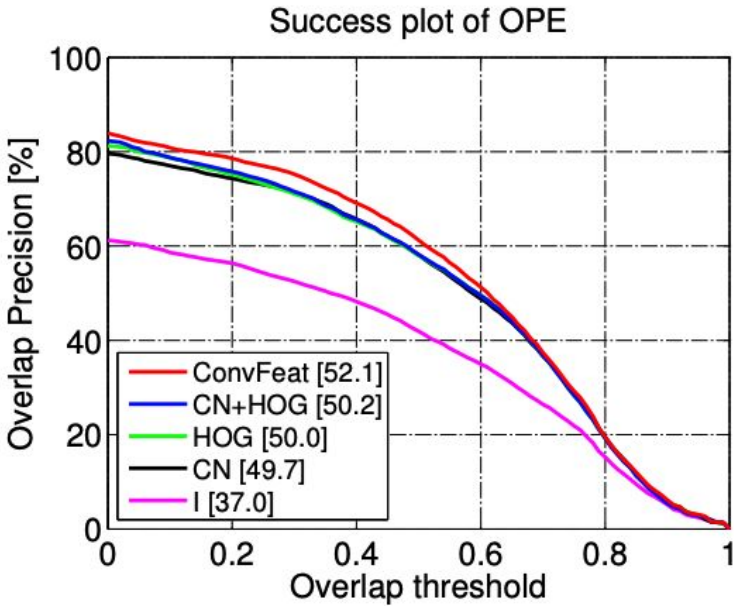
The activations from the convolutional layers of an off-the-shelf CNN (VGG-M) are used as discriminative correlation filters for tracking.



Danelljan, Martin, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. "Convolutional features for correlation filter based visual tracking." ICCCVW 2015.

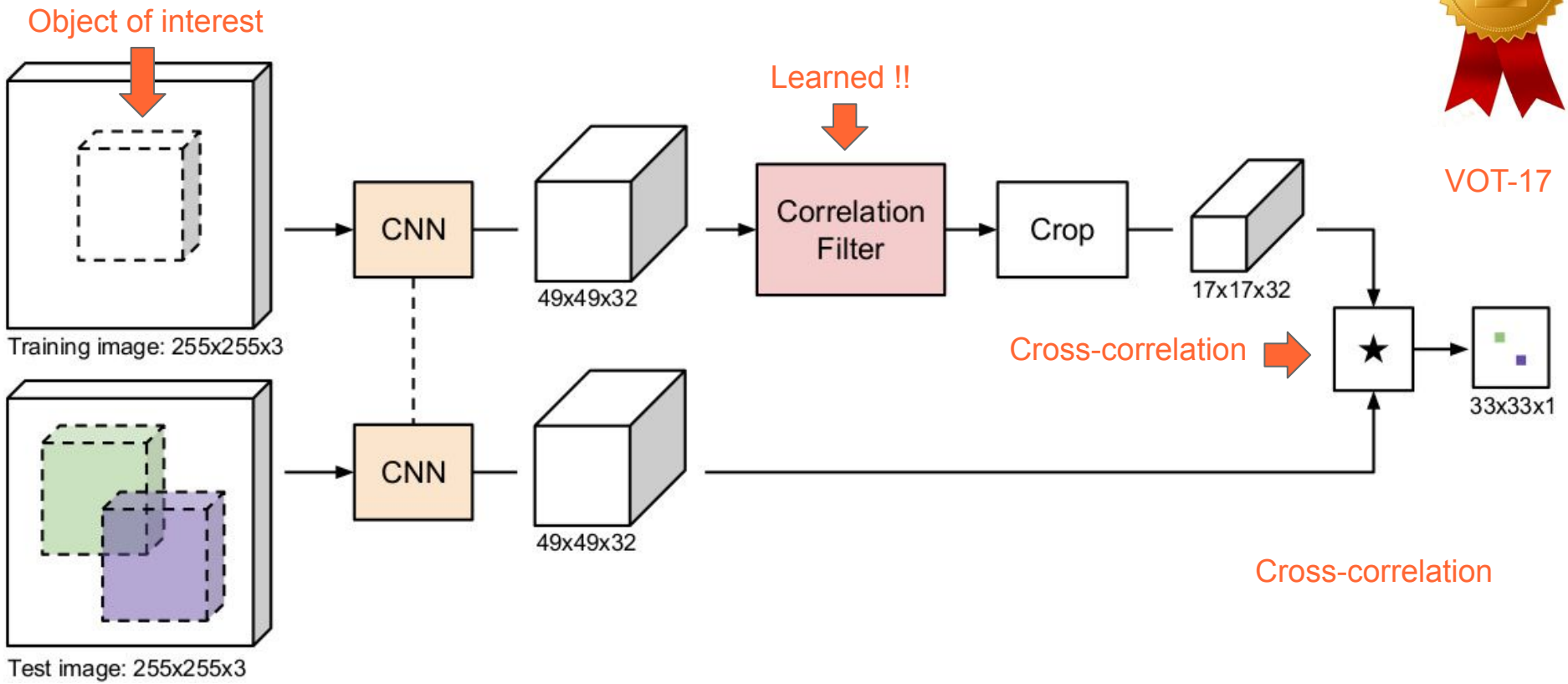
Correlation Filters: Off-the-shelf

Off-the-shelf filters from the first VGG-M layer outperform handcrafted correlation filters.



Danelljan, Martin, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. "Convolutional features for correlation filter based visual tracking." ICCCVW 2015.

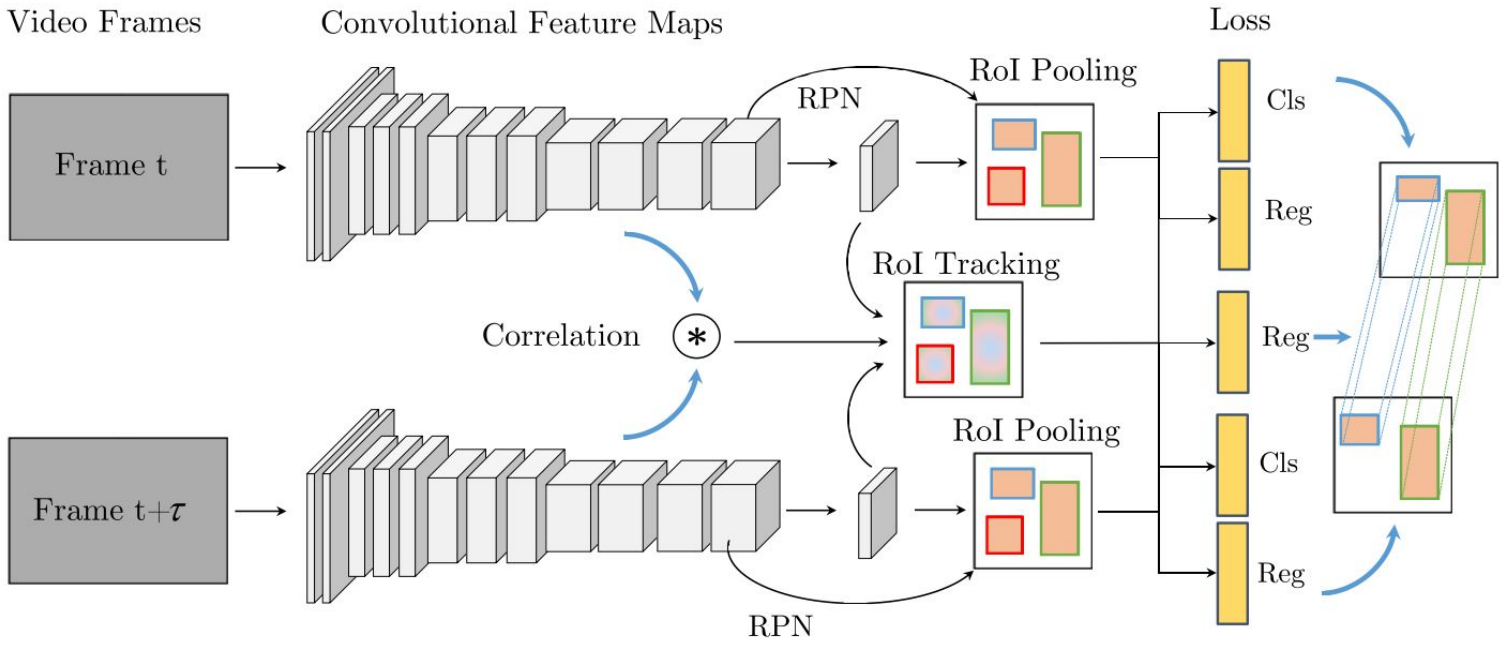
Correlation Filters: Learned



#CFNET Valmadre, Jack, Luca Bertinetto, João F. Henriques, Andrea Vedaldi, and Philip HS Torr. "End-to-end representation learning for Correlation Filter based tracking." CVPR 2017

Correlation Filters: Learned + Detection

Simultaneous detection and tracking in a single architecture.

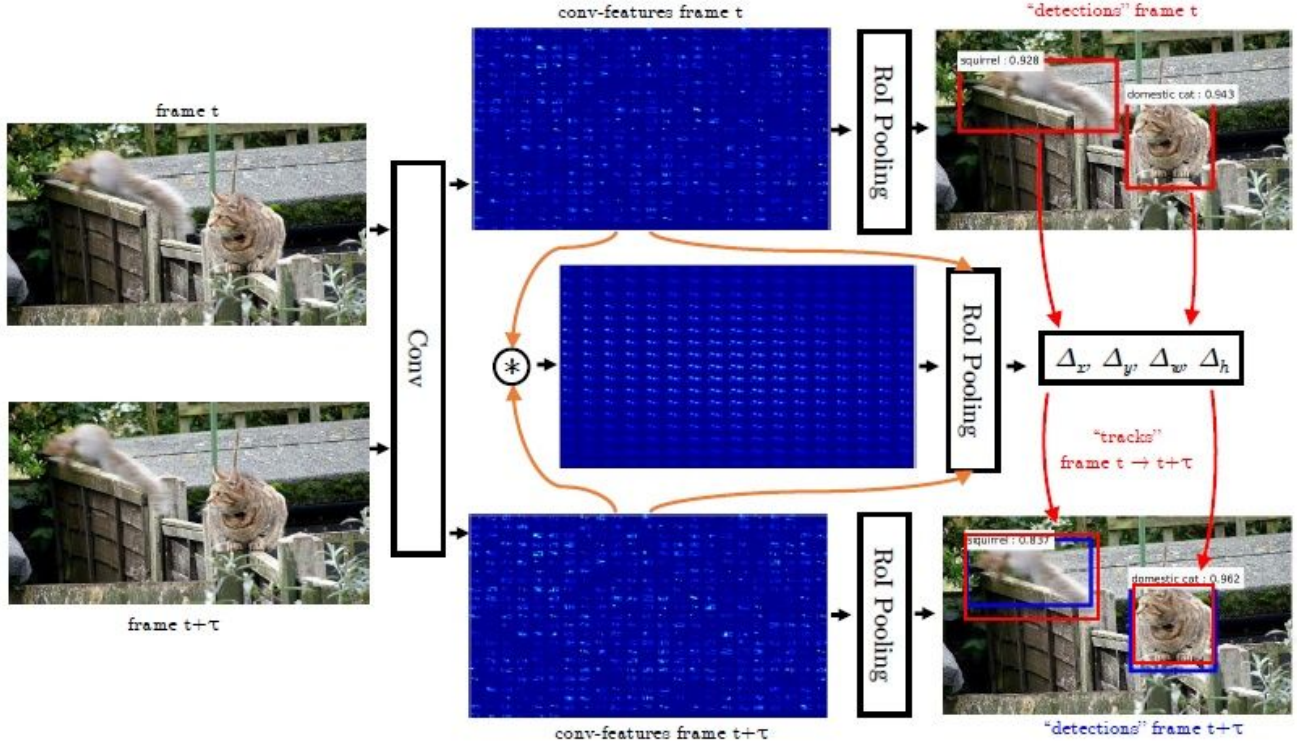


Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Detect to track and track to detect." ICCV 2017

[Slides by Andreu Girbau]

Correlation Filters Learned + Detection

Simultaneous detection and tracking in a single archietcture.



Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Detect to track and track to detect." ICCV 2017

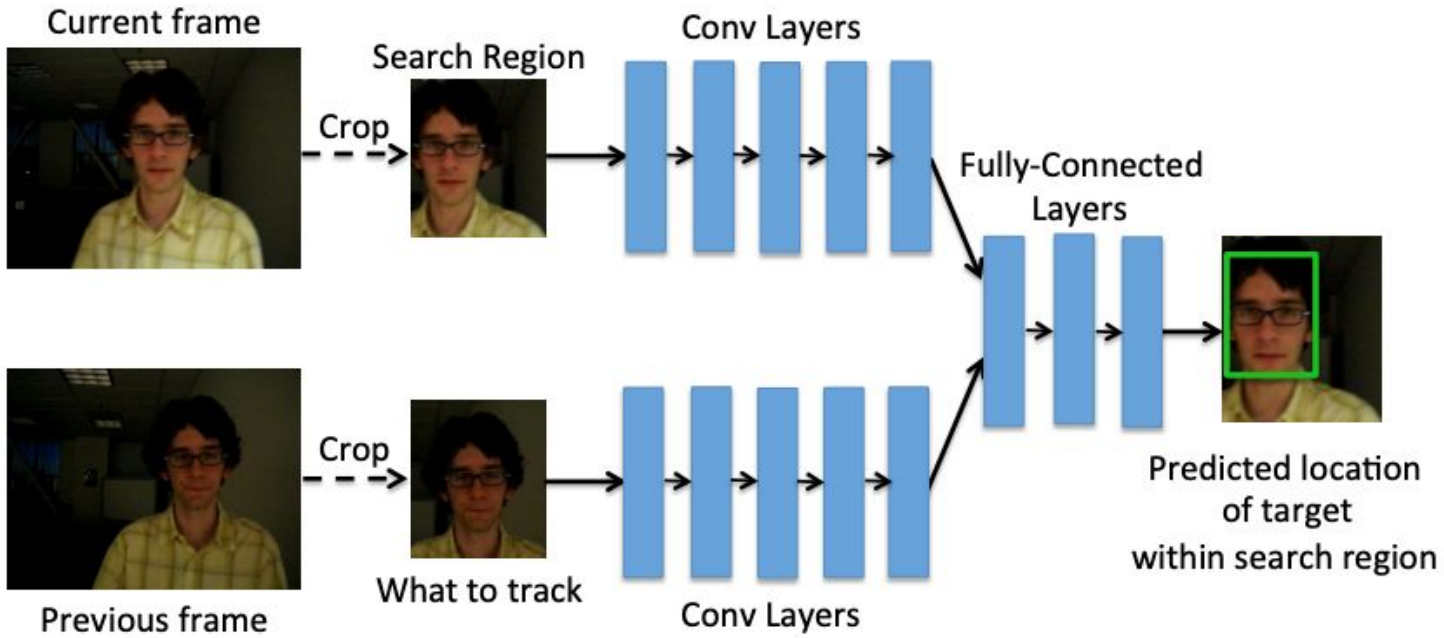
[Slides by Andreu Girbau]

Outline

- Motivation
- First steps with NN/DNN
- Correlation filters
- **Box regressors**
- Tracking by detection
- Tracking with Language

Regression Network

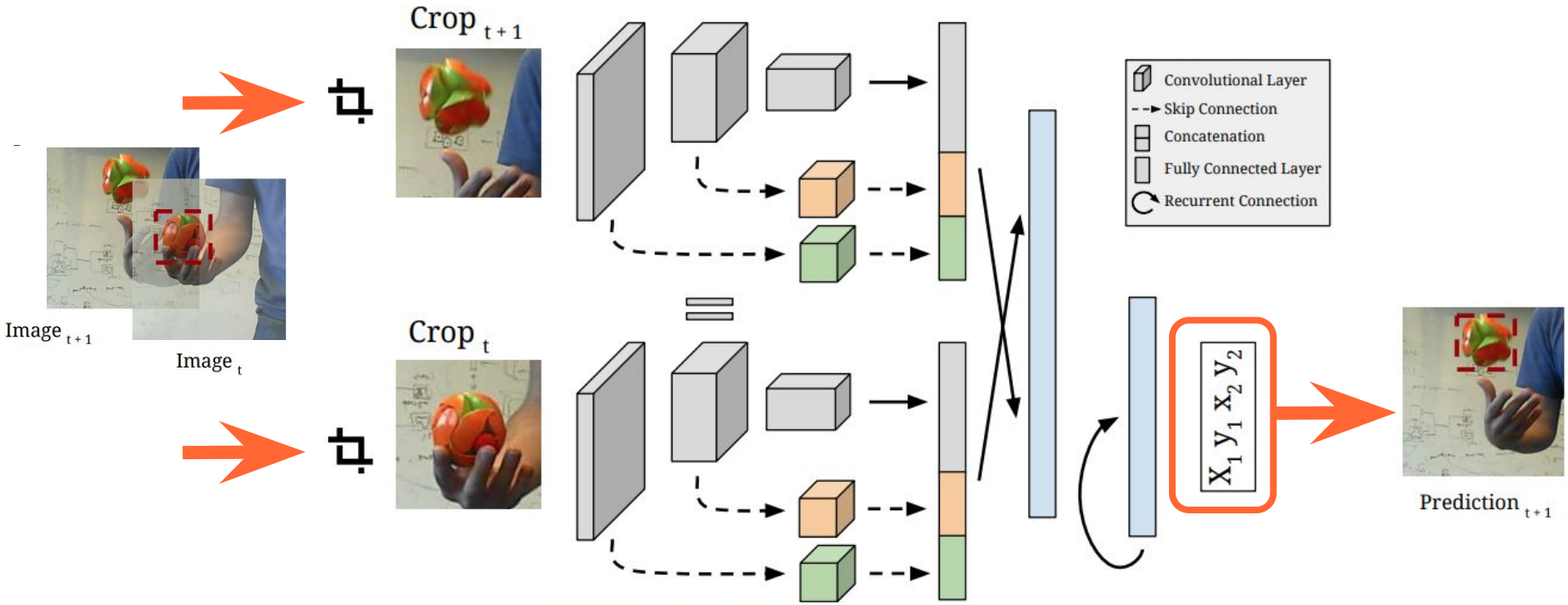
The bounding box to be tracked is not explicitly encoded: it determines the position of the **crops** to be fed into the network. Fast inference at **100 fps**.



#GOTURN Held, David, Sebastian Thrun, and Silvio Savarese. "Learning to track at 100 fps with deep regression networks." ECCV 2016.

Regression Network + RNN

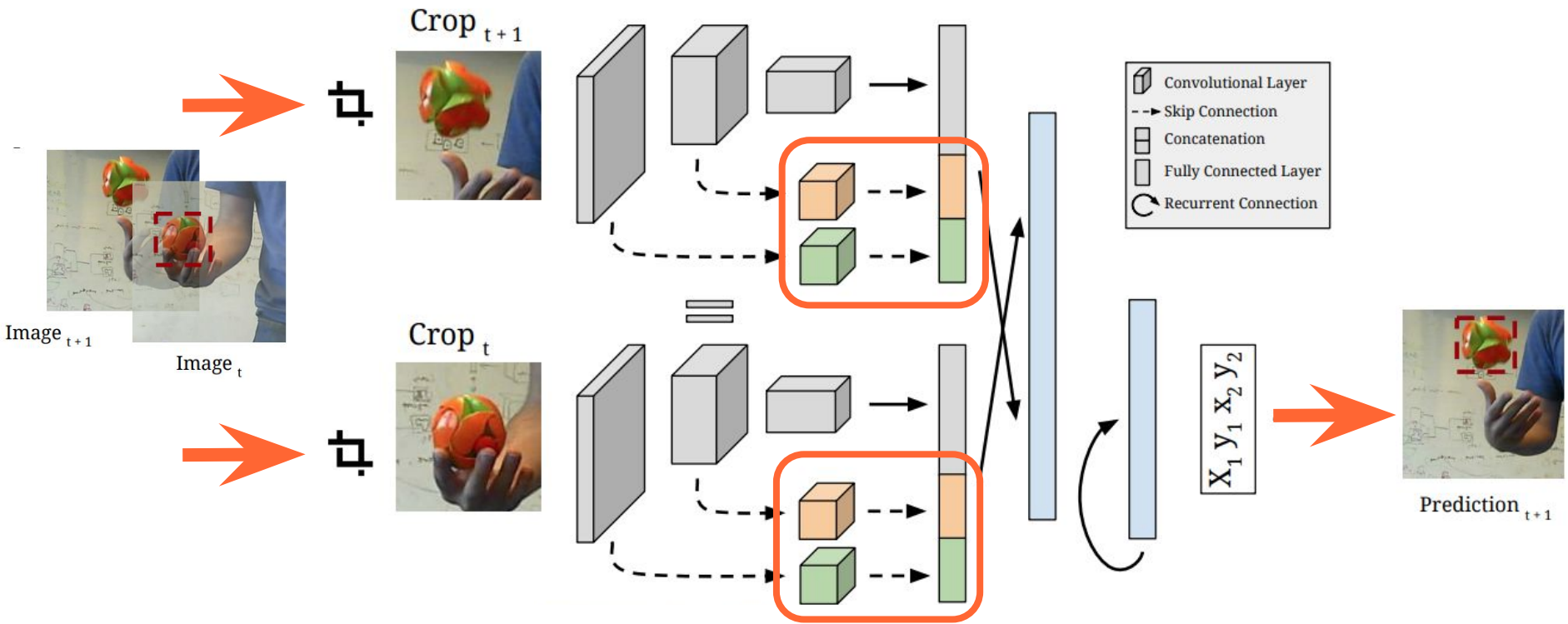
A RNN is added on top of a regression network for long-term temporal consistency.



Gordon, Daniel, Ali Farhadi, and Dieter Fox. "Re^3: Re al-Time Recurrent Regression Networks for Visual Tracking of Generic Objects." ICRA 2018. [code]

Regression Network + RNN

Skip connections help in maintaining the spatial resolution at deeper layers.



Gordon, Daniel, Ali Farhadi, and Dieter Fox. "Re^3: Re al-Time Recurrent Regression Networks for Visual Tracking of Generic Objects." ICRA 2018. [code]

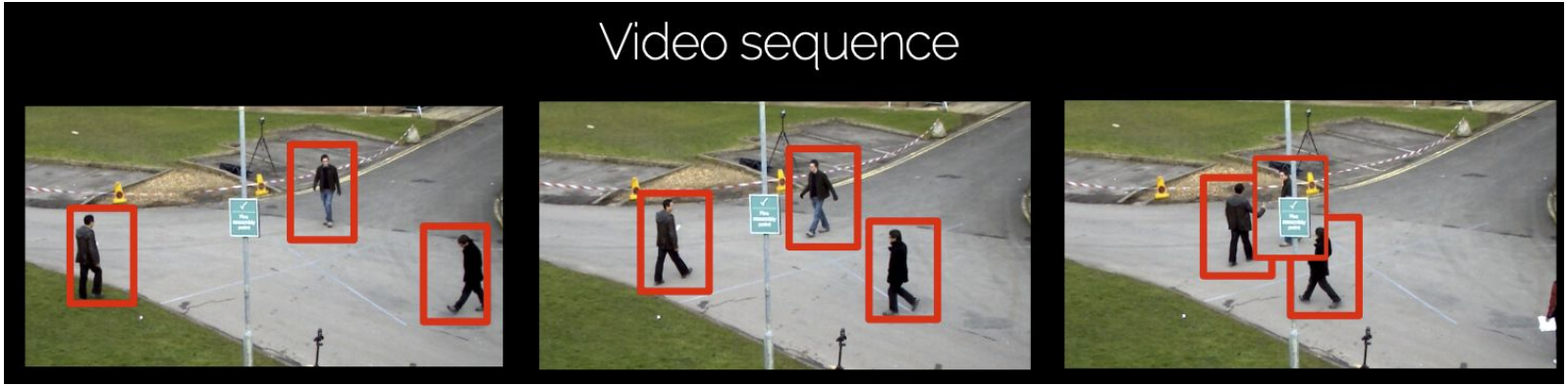
Gordon, Daniel, Ali Farhadi, and Dieter Fox. "Re^3: Re al-Time Recurrent Regression Networks for Visual Tracking of Generic Objects." ICRA 2018.

Outline

- Motivation
- First steps with NN/DNN
- Correlation filters
- Box regressors
- **Tracking by detection**
- Tracking with Language

Tracking-by-Detection

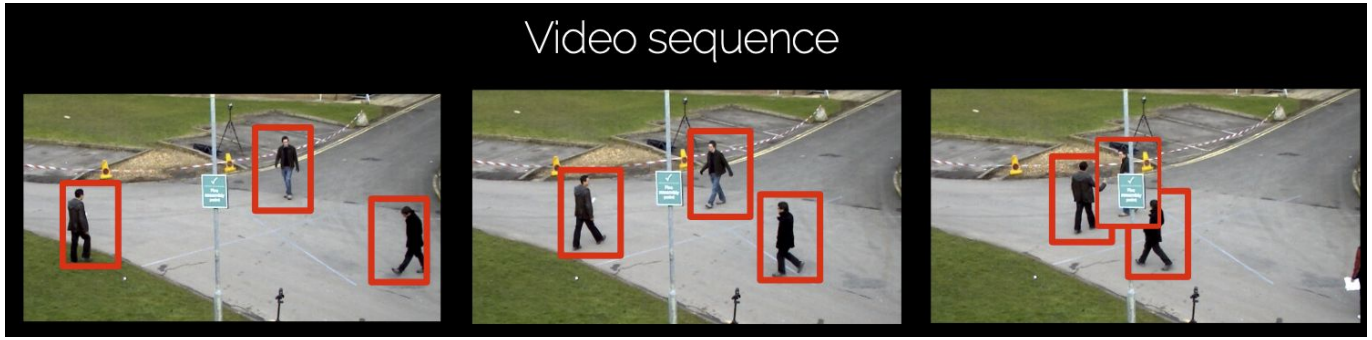
Step 1: Frame-by-frame object detection



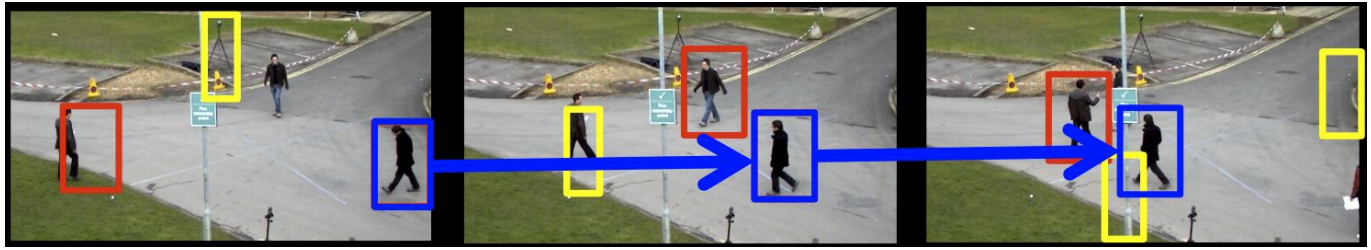
Slide credit: Laura Leal-Taixé (UPC TelecomBCN 2018) [slides] [video]

Tracking-by-Detection

Step 1: Frame-by-frame object detection



Step 2: Link detections to form trajectories.



Step 3: Track management to make decisions on starting and terminating tracks.

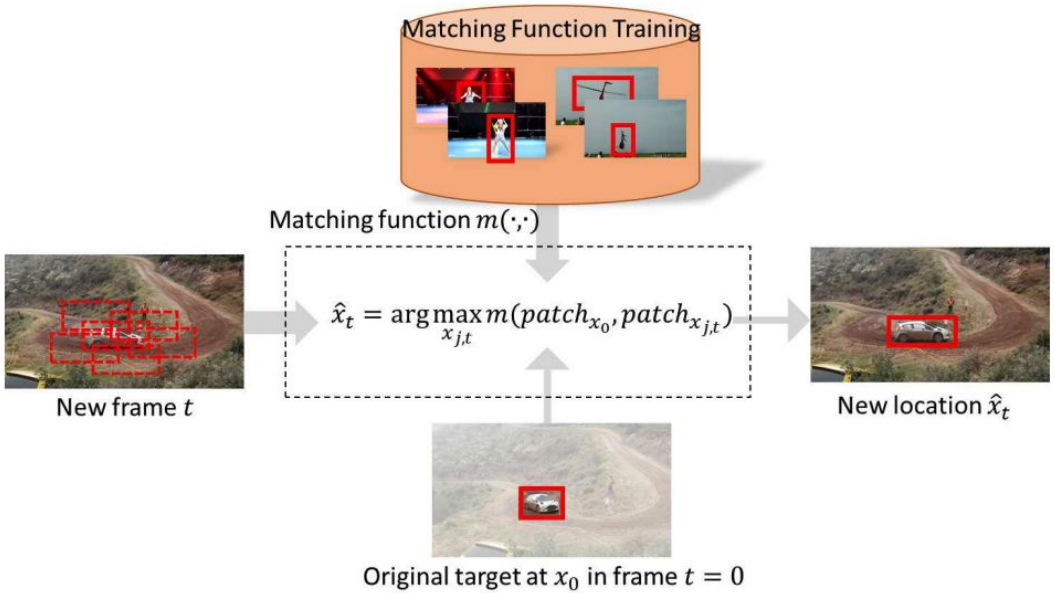
Figures: Laura Leal-Taixé (UPC TelecomBCN 2018) [slides] [video]

Outline

- Motivation
- First steps
- Correlation filters
- Box regressors
- **Tracking by detection**
  - **Learn appearance**
  - + Box regression
  - Graph partitioning
- Tracking with Language

Feature Learning

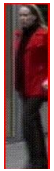
The tracker simply matches the initial patch of the target in the first frame with candidates in a new frame and returns the most similar patch by a **learned matching function**.

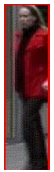


#SINT Tao, Ran, Efstratios Gavves, and Arnold WM Smeulders. "Siamese instance search for tracking." CVPR 2016.

Feature Learning

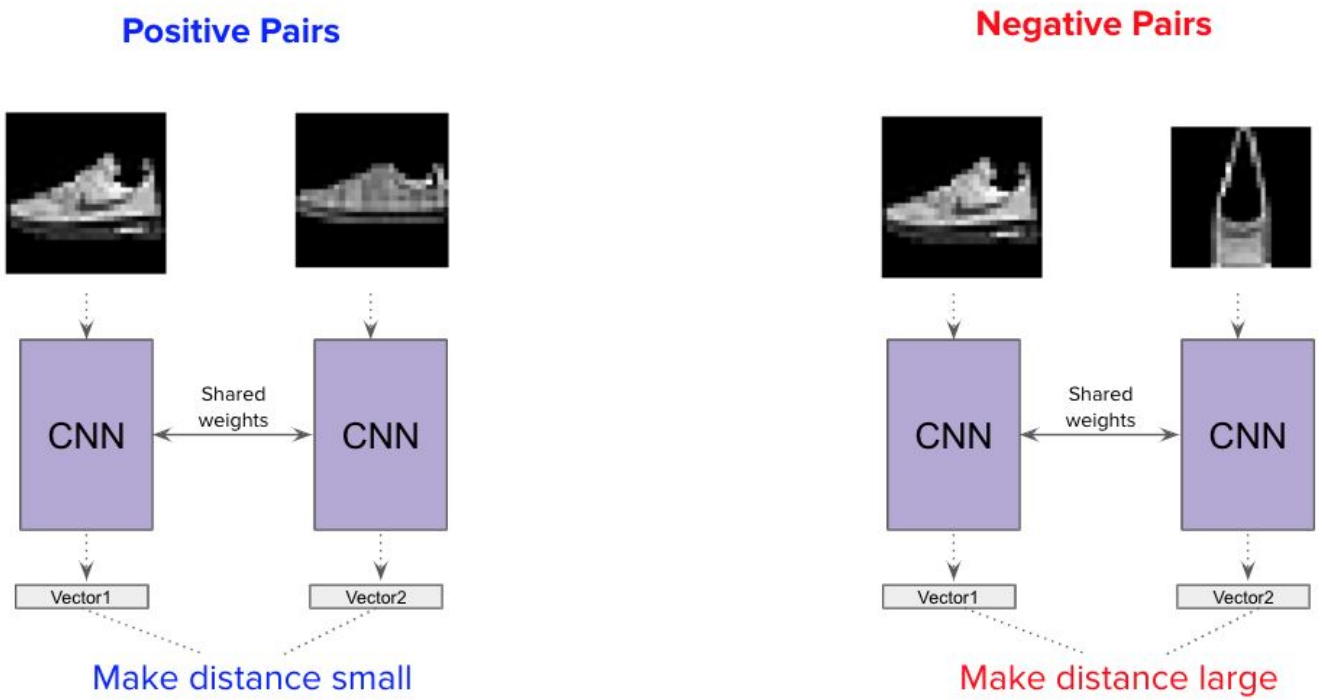
Distance between same objects should be **small**, distance between different objects should be **BIG**.

d( , ) → 0

d( , ) >> 0

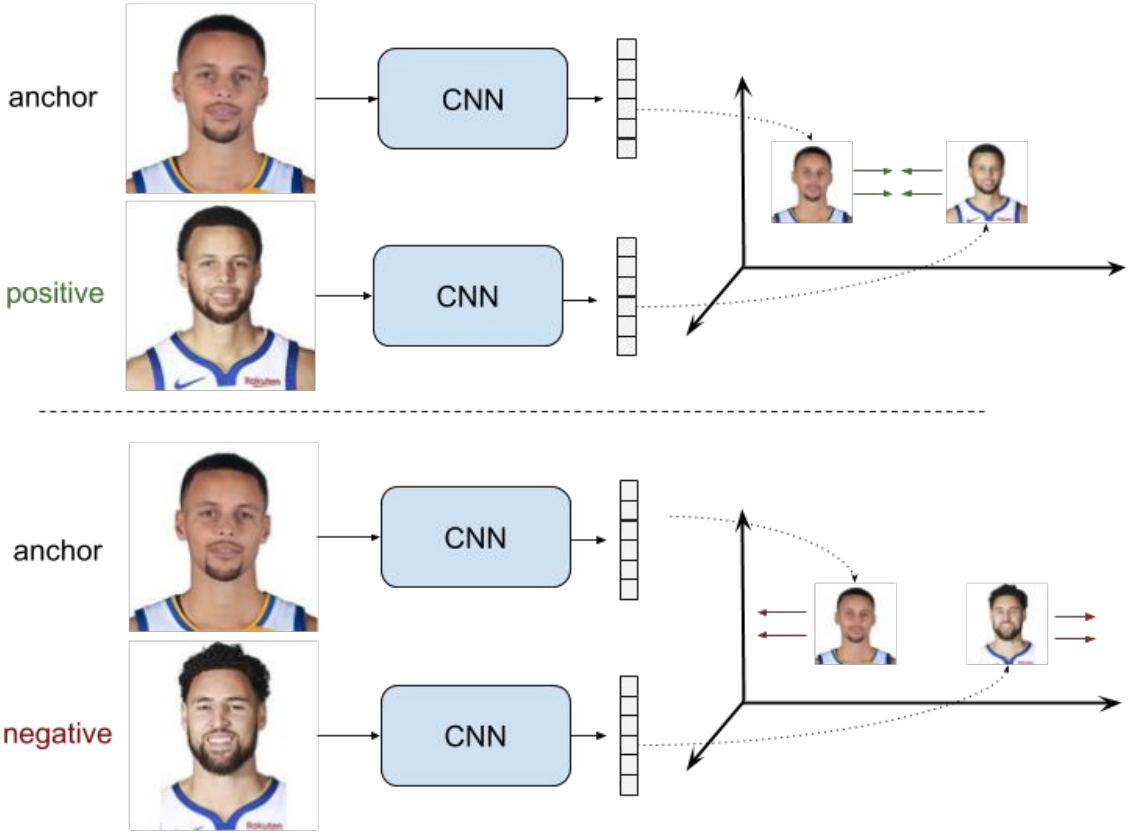
Feature Learning: Siamese networks

Siamese Networks are two identical DNN that share their weights.



$$l(\mathbf{x}_1, \mathbf{x}_2, \delta) = \delta \cdot l_P(d_D(\mathbf{x}_1, \mathbf{x}_2)) + (1 - \delta) \cdot l_N(d_D(\mathbf{x}_1, \mathbf{x}_2))$$

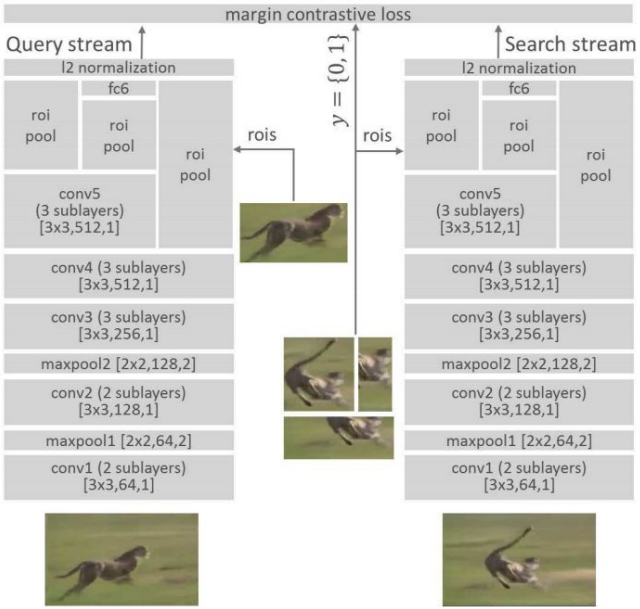
Feature Learning: Siamese networks



Source: Raul Gómez, "Understanding Ranking Loss, Contrastive Loss, Margin Loss, Triplet Loss, Hinge Loss and all those confusing names" (2019)

Feature Learning: Siamese networks

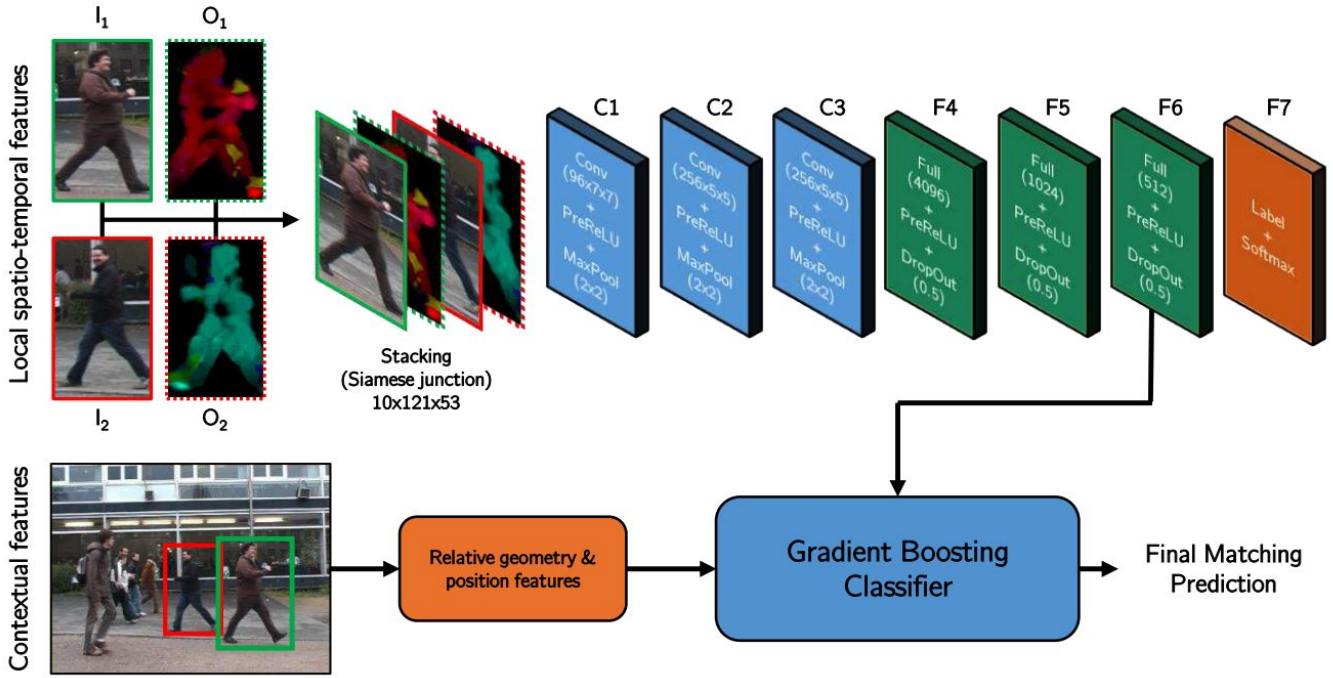
The matching function for object tracking can learned with a **siamese architecture** (shared weights) with a margin contrastive loss.



#SINT Tao, Ran, Efstratios Gavves, and Arnold WM Smeulders. "Siamese instance search for tracking." CVPR 2016.

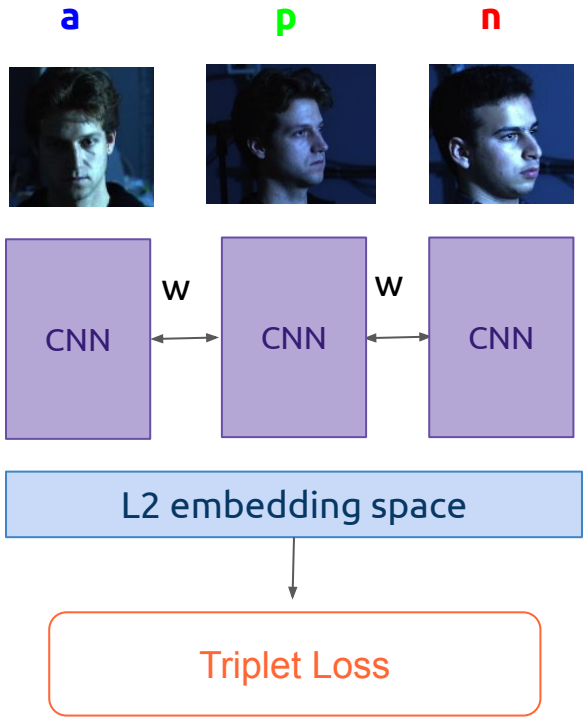
Feature Learning: Siamese networks

The matching function may be completed with **contextual features** derived from the position and size of the compared input patches.



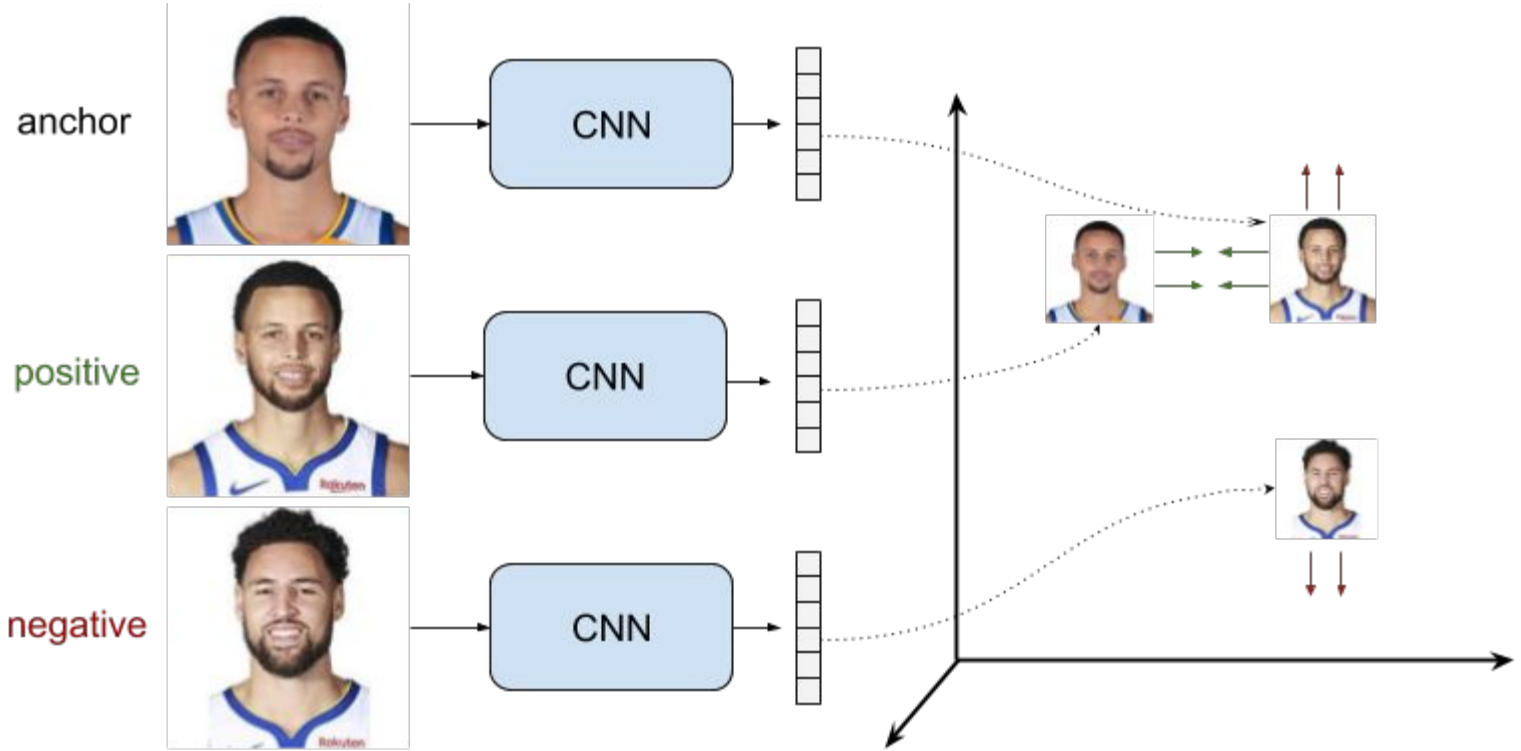
Leal-Taixé, Laura, Cristian Canton-Ferrer, and Konrad Schindler. "Learning by tracking: Siamese CNN for robust target association." CVPRW. 2016. [video]

Feature Learning: Triplet Loss



#FaceNet Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." CVPR 2015.

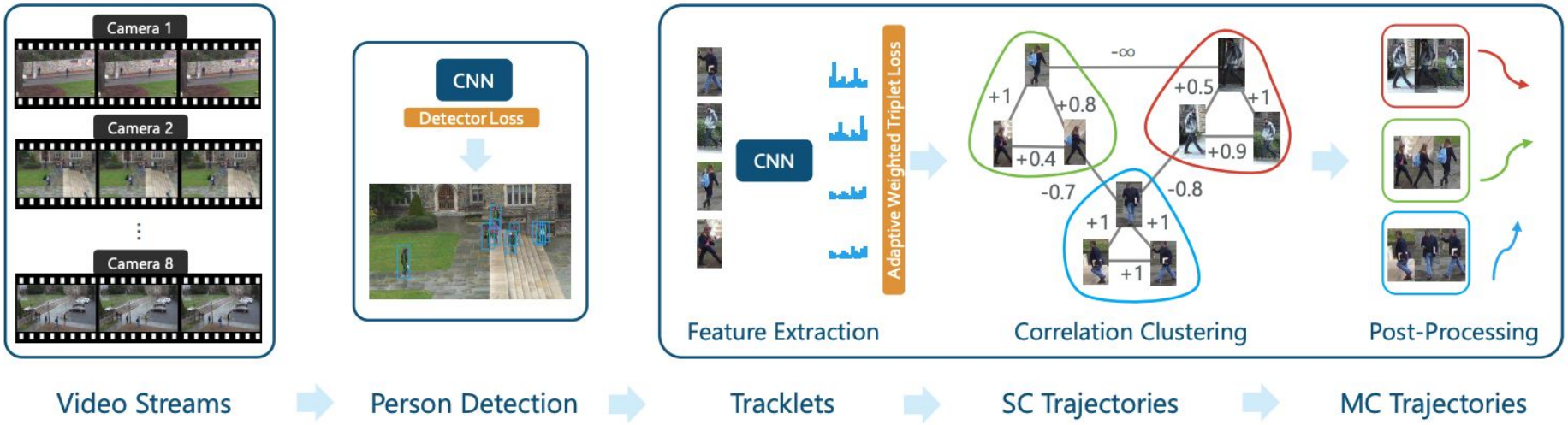
Feature Learning: Triplet Loss



Source: Raul Gómez, "Understanding Ranking Loss, Contrastive Loss, Margin Loss, Triplet Loss, Hinge Loss and all those confusing names" (2019)

Feature Learning: Triplet Loss

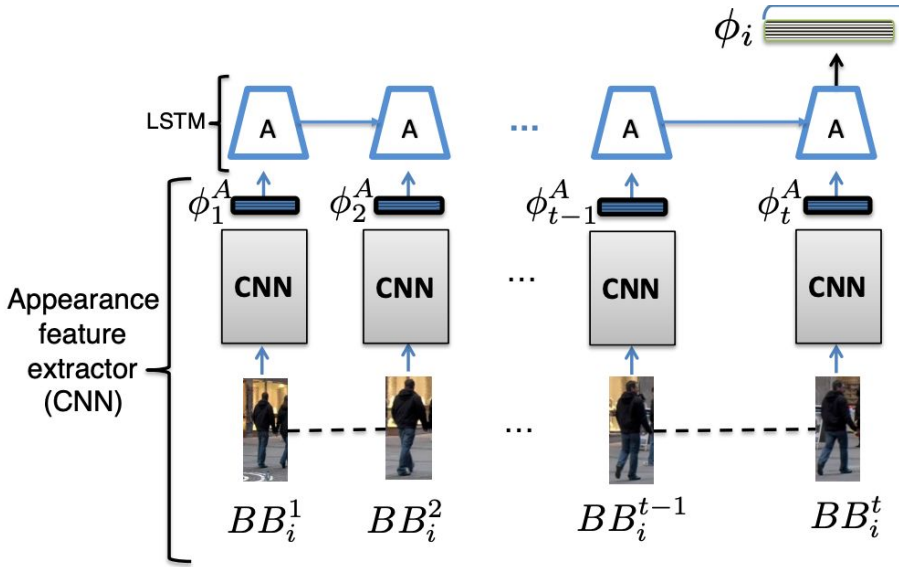
CNN is trained with a triplet loss.



Ristani, Ergys, and Carlo Tomasi. "Features for multi-target multi-camera tracking and re-identification." CVPR 2018.

Feature Learning: Temporal aggregation

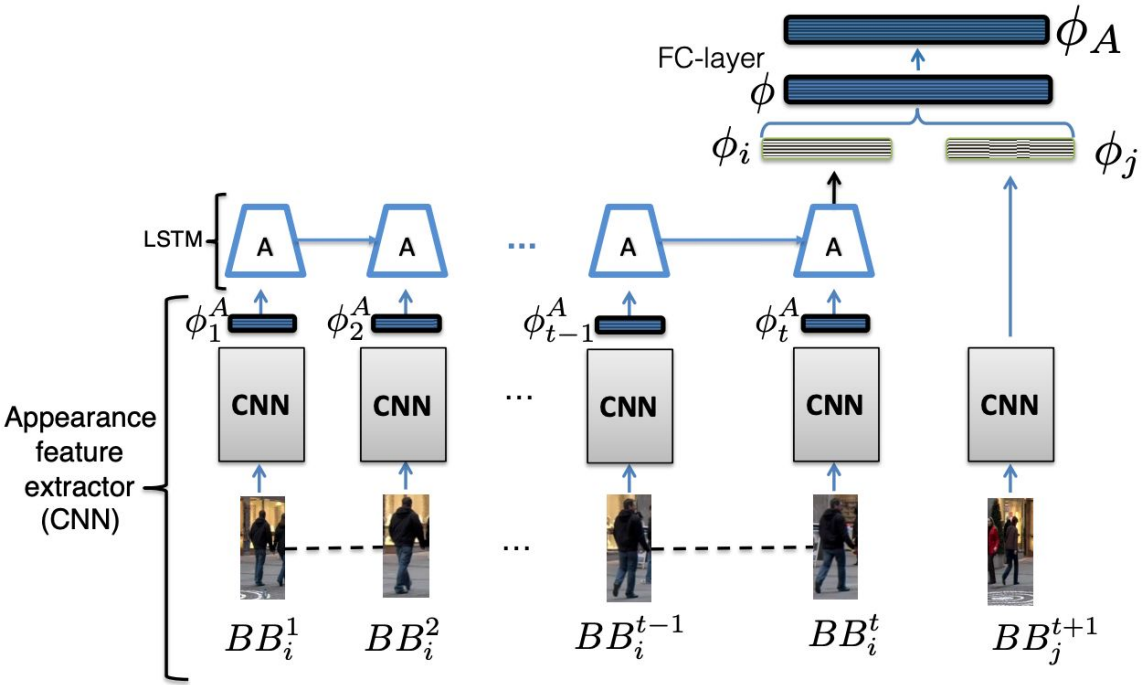
The **temporally aggregated appearance** of target BB_i from t to t+1 is encoded with an LSTM.



Sadeghian, Amir, Alexandre Alahi, and Silvio Savarese. "Tracking the untrackable: Learning to track multiple cues with long-term dependencies." ICCV 2017.

Feature Learning: Temporal aggregation

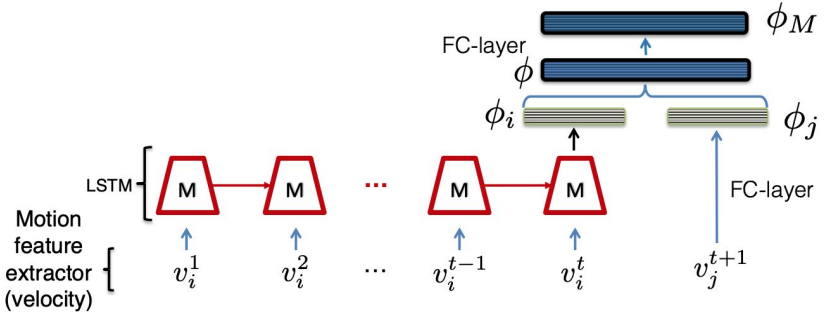
A FC layer assessed whether the appearance of BB_i matches the one of detection BB_j in t+1.

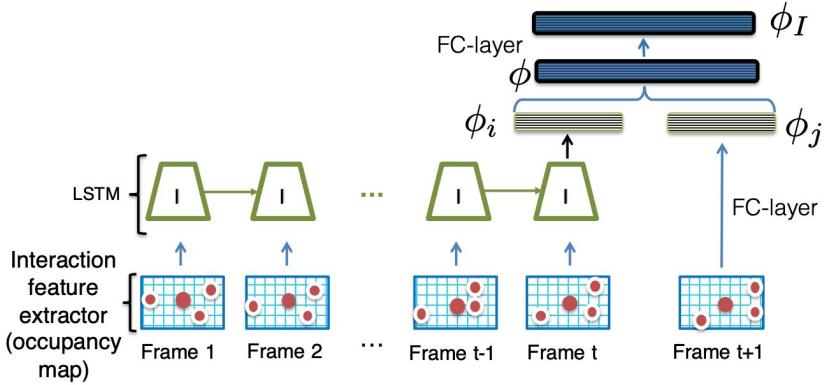


Sadeghian, Amir, Alexandre Alahi, and Silvio Savarese. "Tracking the untrackable: Learning to track multiple cues with long-term dependencies." ICCV 2017.

Feature Learning: Temporal aggregation

A similar approach is used to assess the similarity between targets and detections in terms of **motion** and **interaction** (spatial layout).

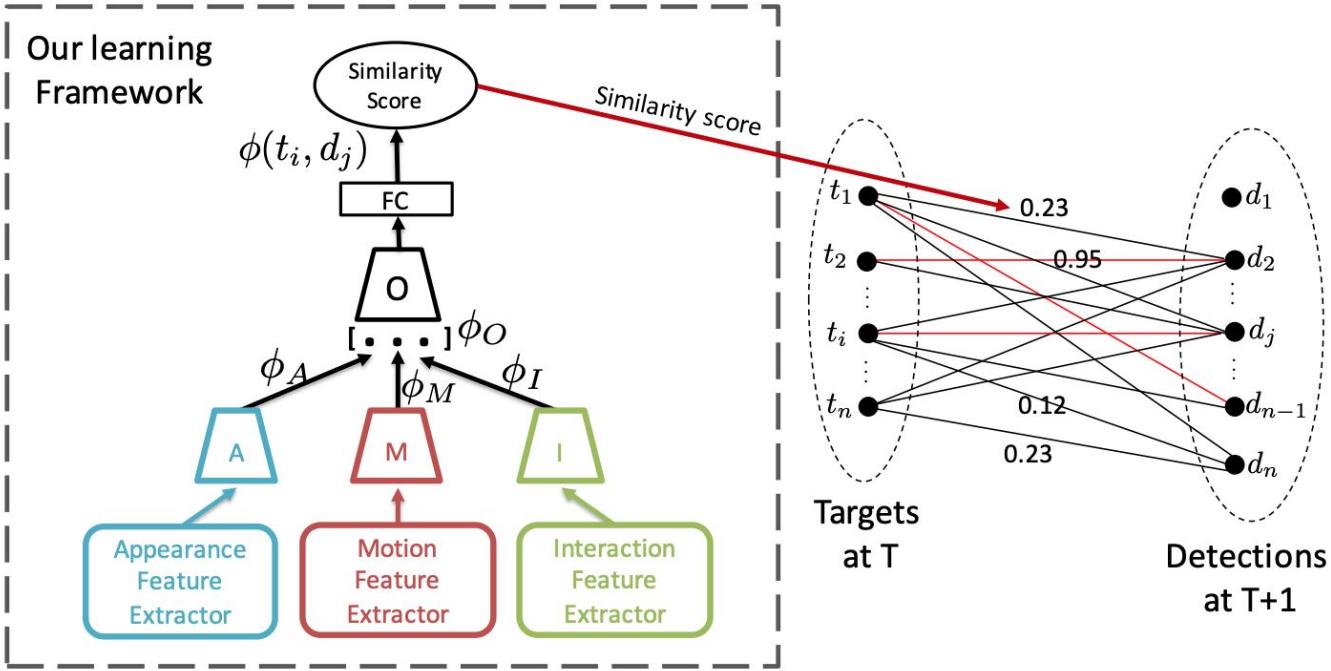




Sadeghian, Amir, Alexandre Alahi, and Silvio Savarese. "Tracking the untrackable: Learning to track multiple cues with long-term dependencies." ICCV 2017.

Feature Learning: Temporal aggregation

The similarities in terms of appearance, motion and interaction are fused with another fully connected (FC) layer to compute a similarity score between each target at T and detections at T+1.



Sadeghian, Amir, Alexandre Alahi, and Silvio Savarese. "Tracking the untrackable: Learning to track multiple cues with long-term dependencies." ICCV 2017.

Outline

- Motivation
- First steps
- Correlation filters
- Box regressors
- **Tracking by detection**
  - Learn appearance
  - **+ Box regression**
  - Graph partitioning
- Tracking with Language

Tracking by detection + Box regression

Background: Faster-RCNN

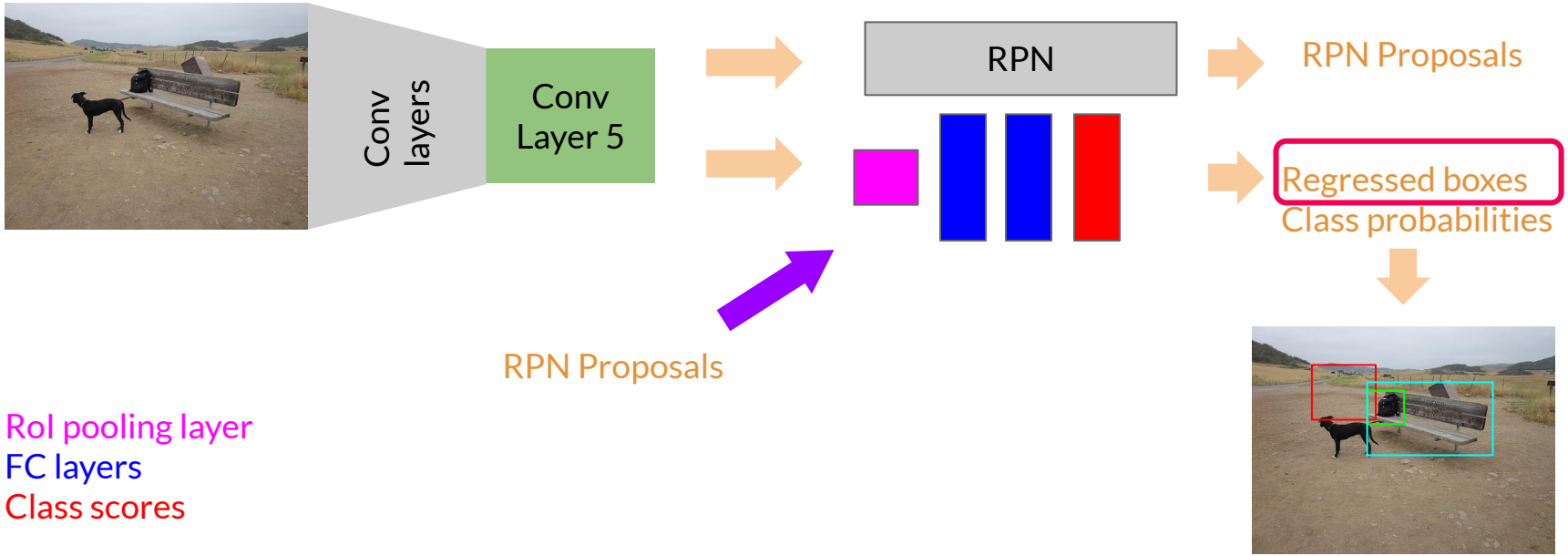
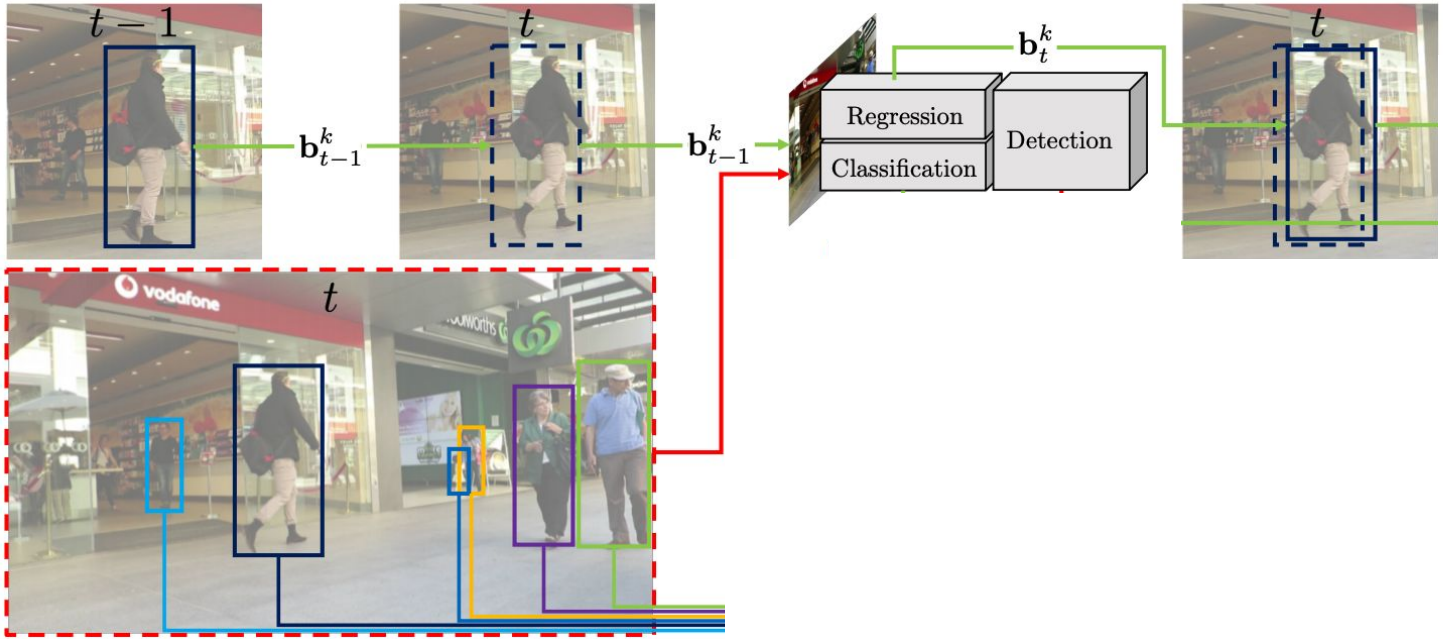


Figure: Amaia Salvador (DLCV 2016)

#FasterRCNN Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks." NIPS 2015. [Lecture by Amaia Salvador] [Lecture by Míriam Bellver]

Tracking by detection + Box regression

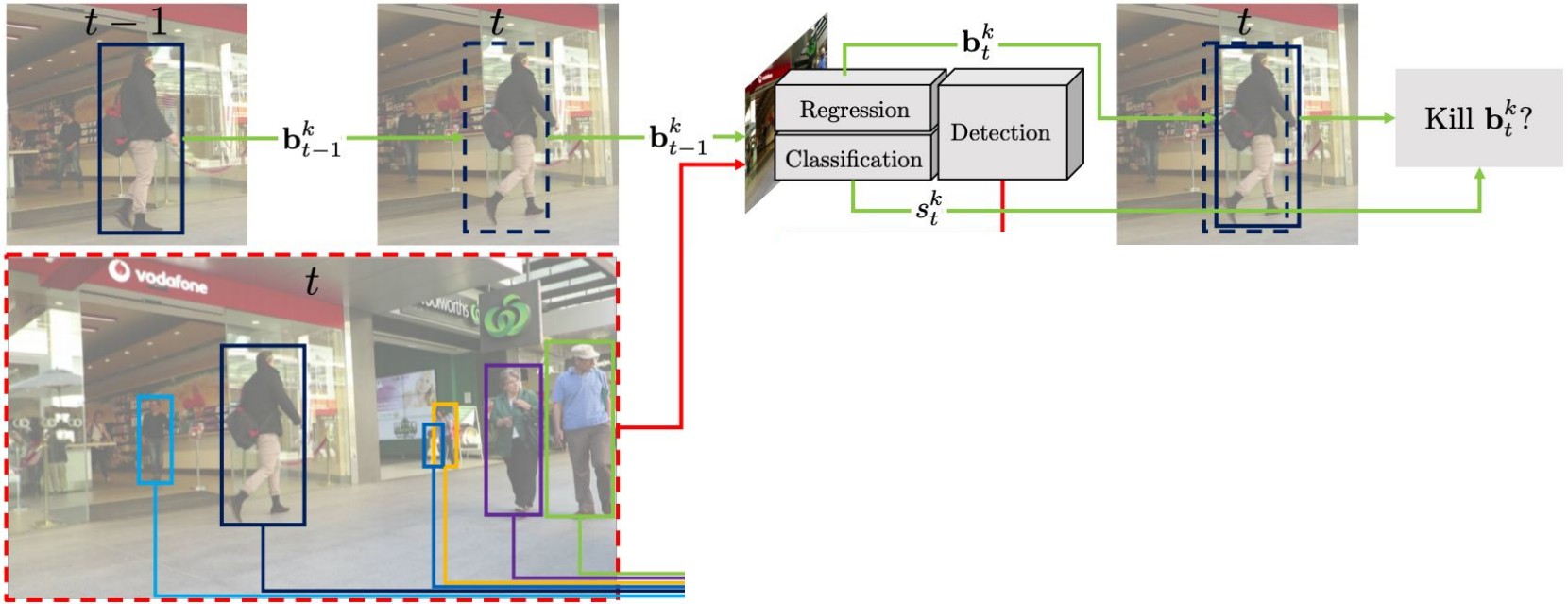
Green: The **regressor** from the Faster R-CNN detector regresses all bounding boxes from previous frames (b^k_{t-1}) to the object's new position in frame t (b^k_t).



#Tracktor Bergmann, Philipp, Tim Meinhardt, and Laura Leal-Taixe. "Tracking without bells and whistles." arXiv (2019). [Slides]

Tracking by detection + Box regression

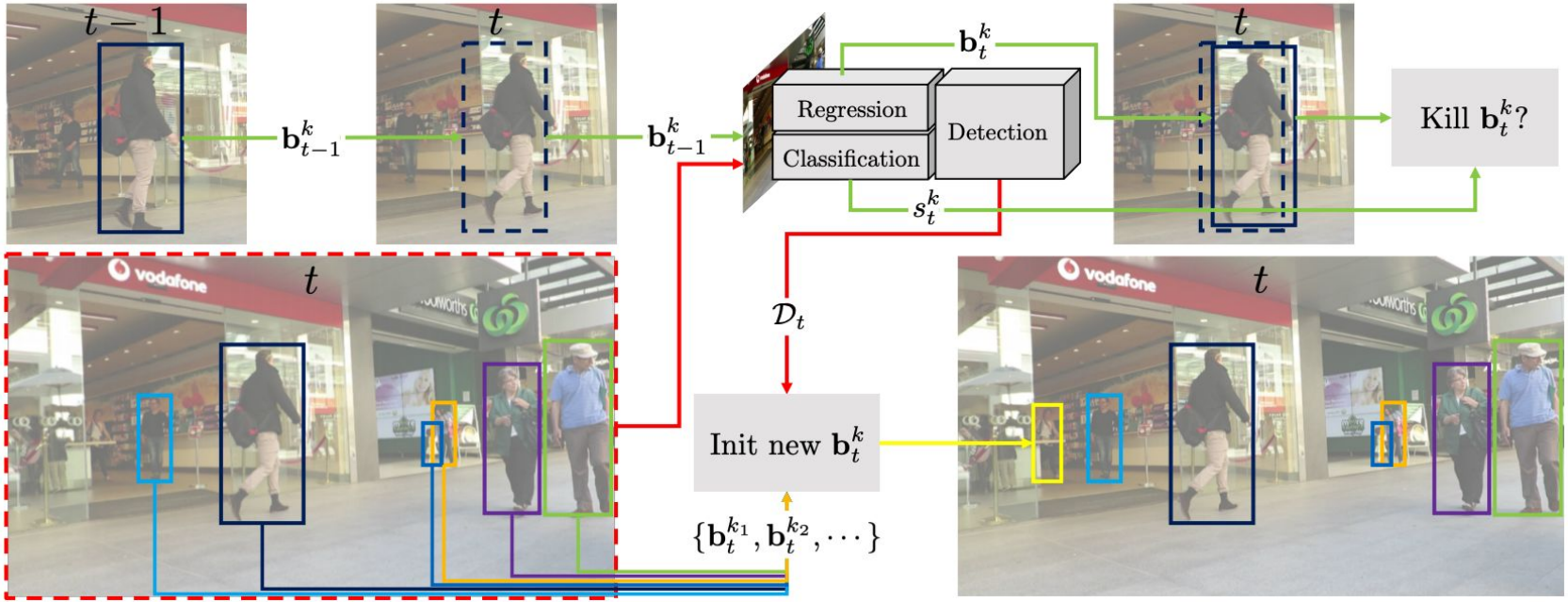
Green: The corresponding object classification scores s^k_t of the new bounding box positions are then used to kill potentially occluded tracks.



#Tracktor Bergmann, Philipp, Tim Meinhardt, and Laura Leal-Taixe. "Tracking without bells and whistles." arXiv (2019). [Slides]

Tracking by detection + Box regression

Red: The object detector provides a set of object detections D_t at frame t (not only the ones tracked from t-1). This allows initializing new tracks if non of the detections has an IoU larger than a certain threshold.

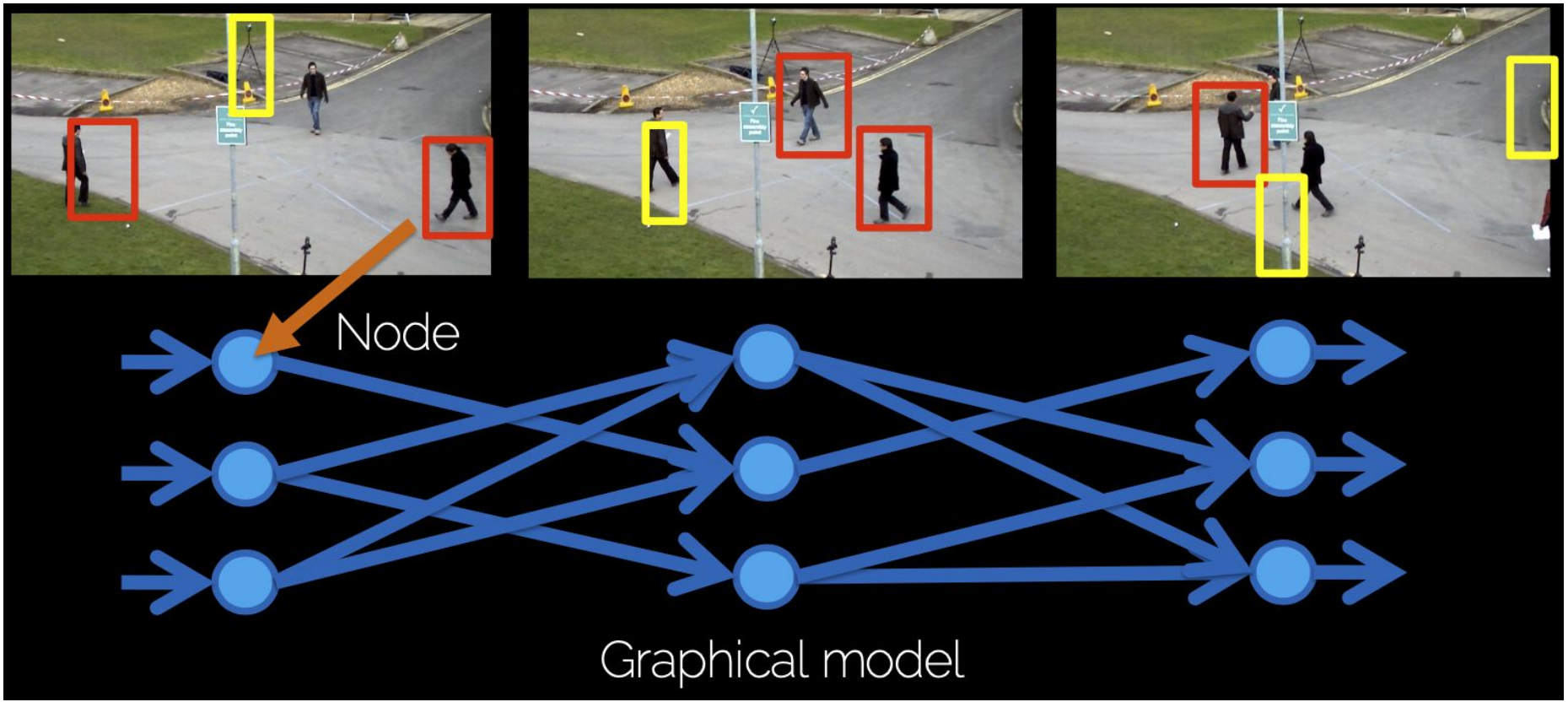


#Tracktor Bergmann, Philipp, Tim Meinhardt, and Laura Leal-Taixe. "Tracking without bells and whistles." arXiv (2019). [Slides]

Outline

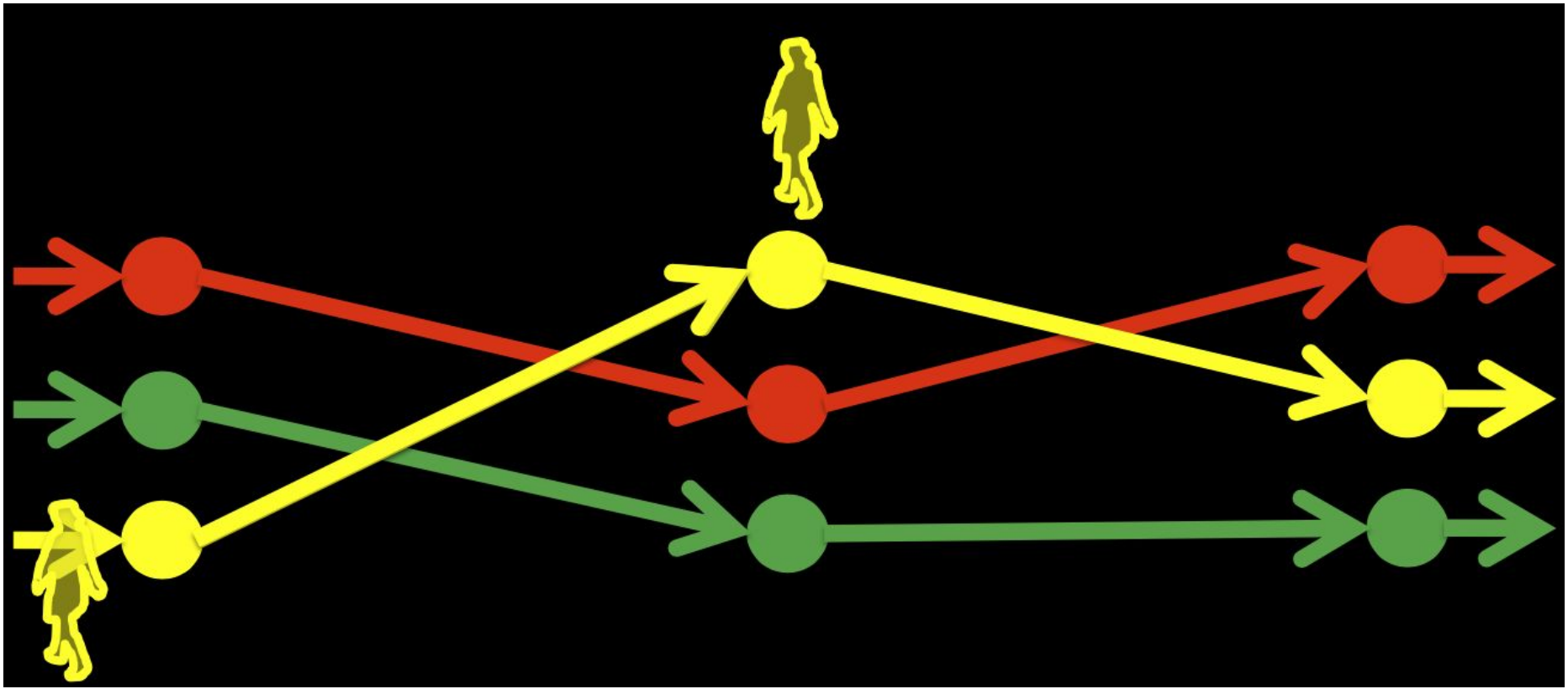
- Motivation
- First steps
- Correlation filters
- Box regressors
- **Tracking by detection**
  - Learn appearance
  - + Box regression
  - **Graph partitioning**
- Tracking with Language

Tracking by detection + Graph partition



Slide credit: Laura Leal-Taixé (UPC TelecomBCN 2018) [slides] [video]

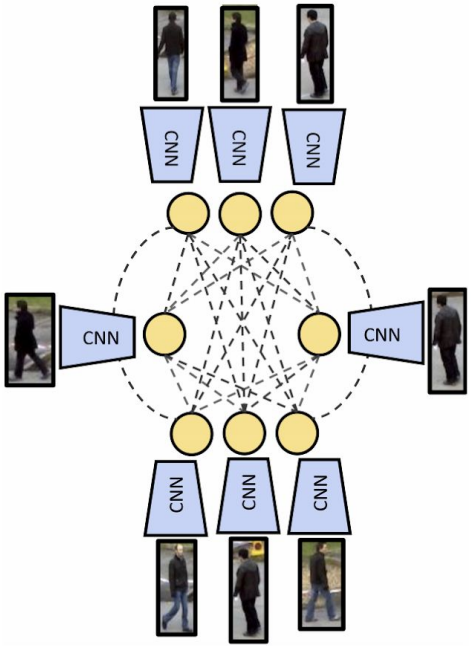
Tracking by detection + Graph partition



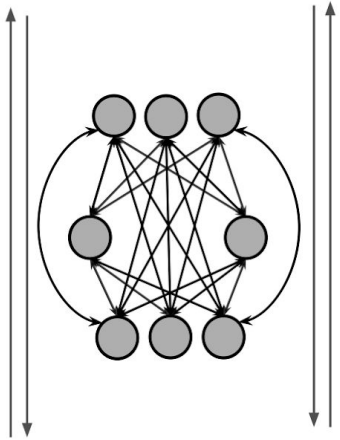
Slide credit: Laura Leal-Taixé (UPC TelecomBCN 2018) [slides] [video]

Tracking by detection + Graph partition

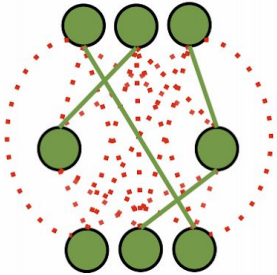
(a) Input



(b) Graph Construction + Feature Encoding



(c) Neural Message Passing



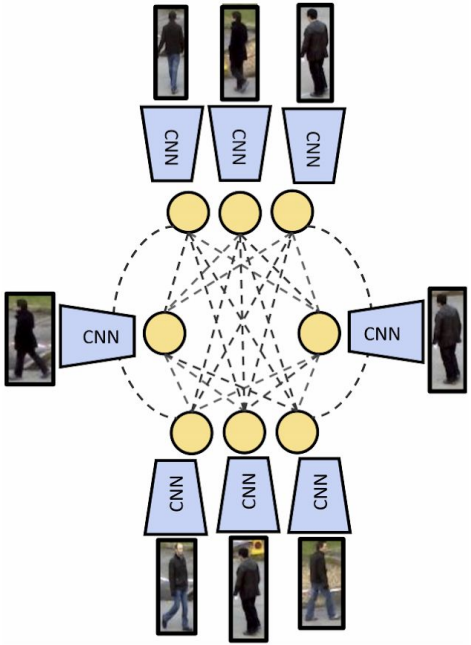
(d) Edge Classification

(e) Output

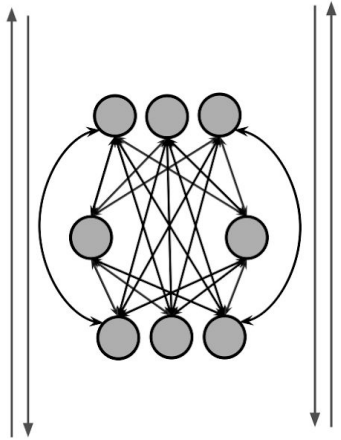
Brasó, Guillem, and Laura Leal-Taixé. "Learning a Neural Solver for Multiple Object Tracking." CVPR 2020.

Tracking by detection + Hungarian

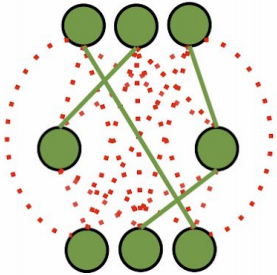
(a) Input



(b) Graph Construction + Feature Encoding



(c) Neural Message Passing



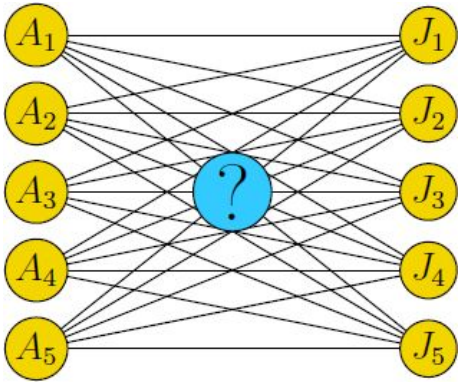
(d) Edge Classification

(e) Output

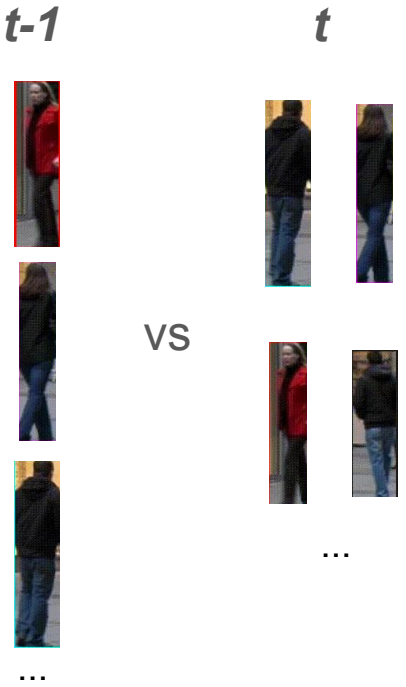
Brasó, Guillem, and Laura Leal-Taixé. "Learning a Neural Solver for Multiple Object Tracking." CVPR 2020.

Tracking by detection + Hungarian

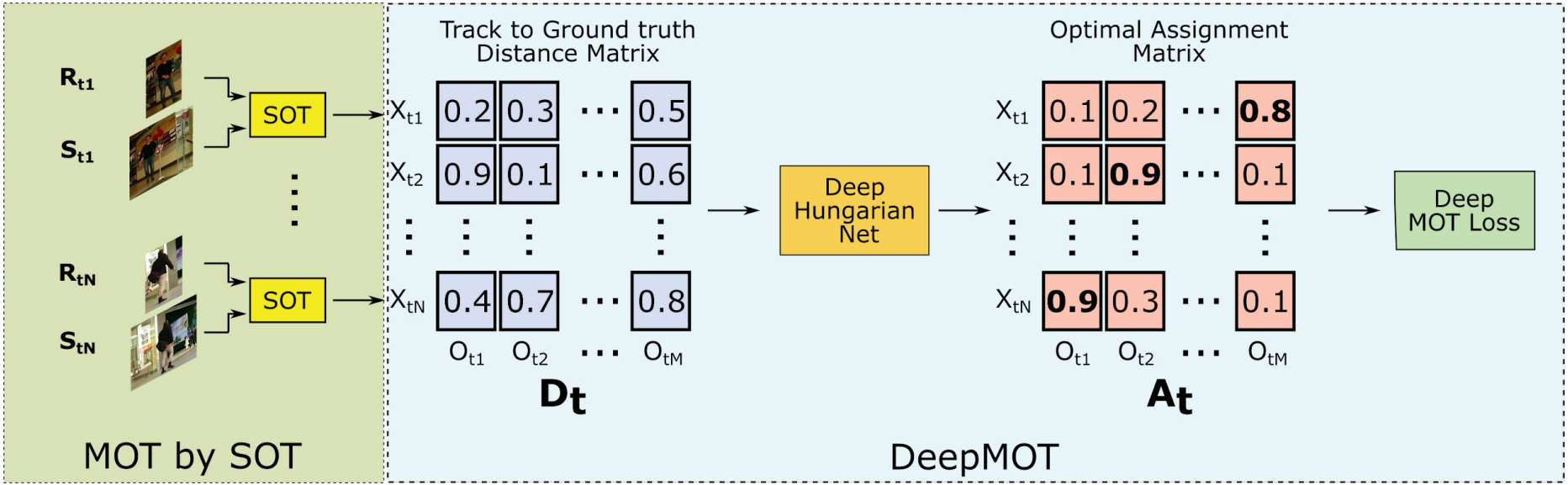
Assignment problem (Hungarian algorithm)



[Hungarian algorithm explained]



Tracking by detection + Hungarian



#DeepMOT Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixe, Xavier Alameda-Pineda, "How To Train Your Deep Multi-Object Tracker" CVPR 2020. [code]

Outline

- Motivation
- First steps
- Correlation filters
- Box regressors
- Tracking by detection
- **Tracking with Language**

Object tracking with Language

Li, Zhenyang, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. "Tracking by natural language specification." CVPR 2017.

Object tracking with Langauge

Feng, Qi, Vitaly Ablavsky, Qinxun Bai, Guorong Li, and Stan Sclaroff. "Real-time Visual Object Tracking with Natural Language Description." WACV 2020.

Outline

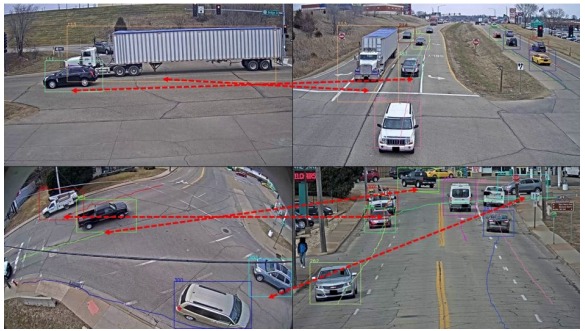
- Motivation
- First steps
- Correlation filters
- Box regressors
- Tracking by detection
- Tracking with Language

Learn more

- Papers with Code: [Multiple Object Tracking]

[5th BMTT MOTChallenge Workshop: Multi-Object Tracking and Segmentation]

[AI City Challenge 2020]

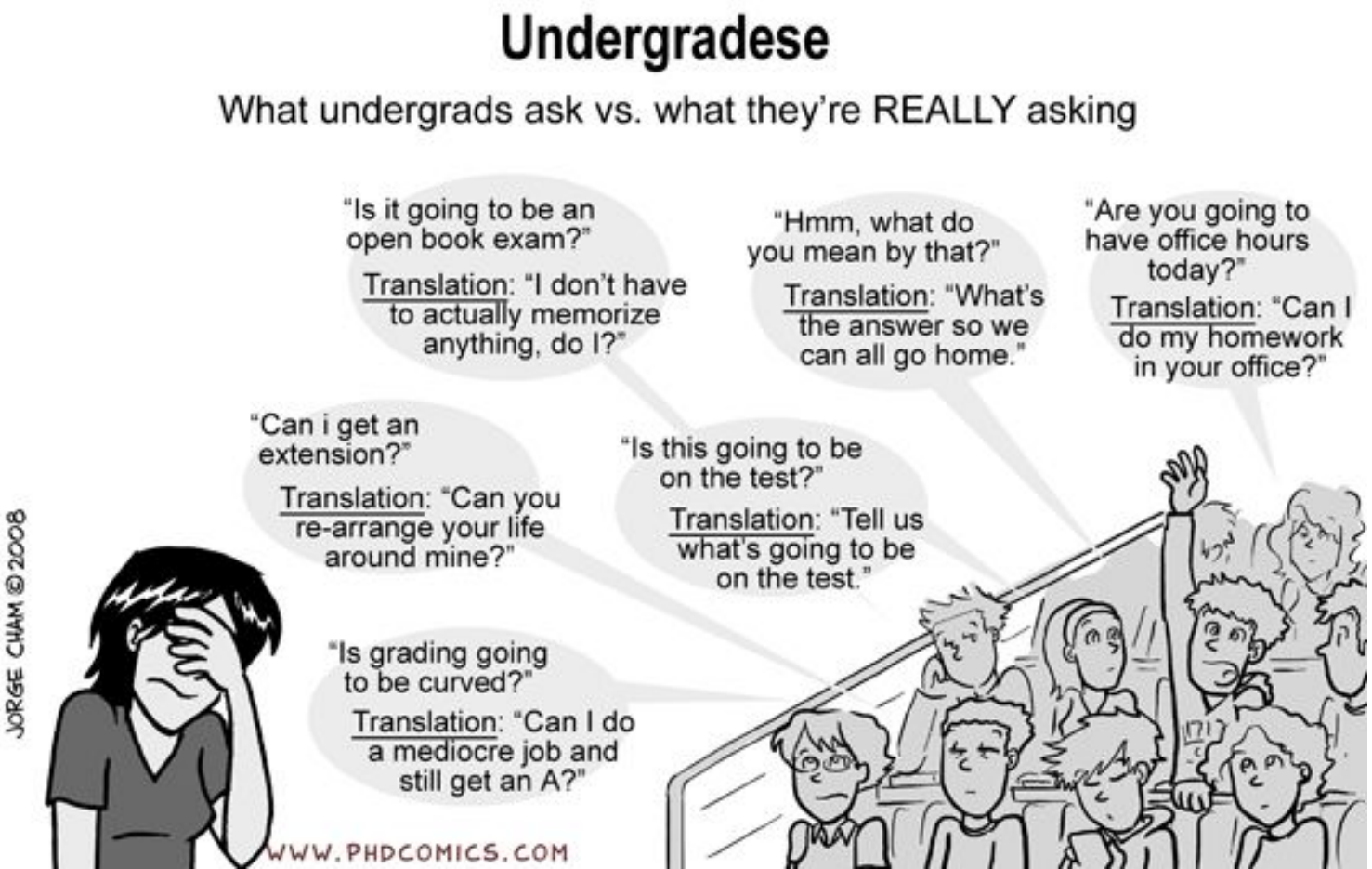


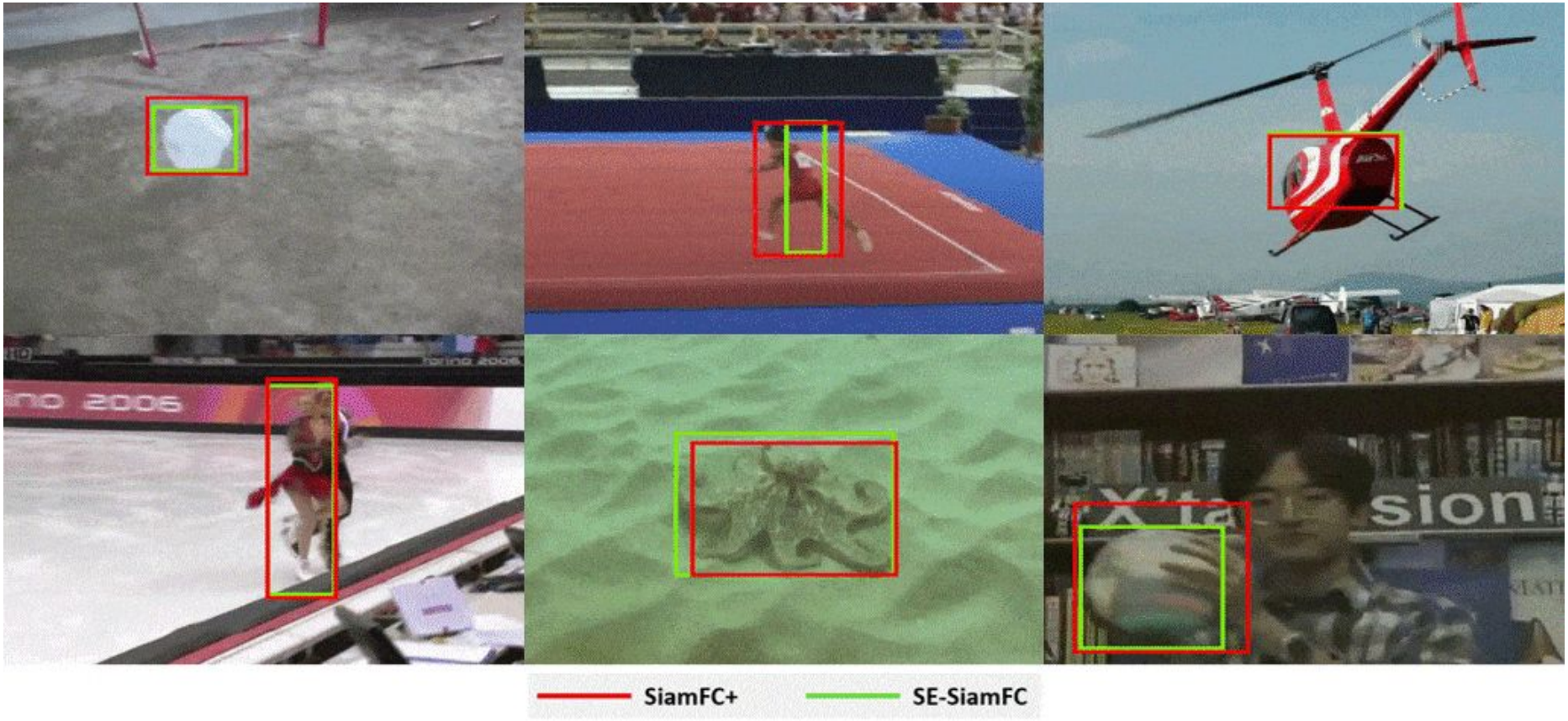
[2nd Workshop and Challenge on Target Re-identification and Multi-Target Multi-Camera Tracking]

Other references

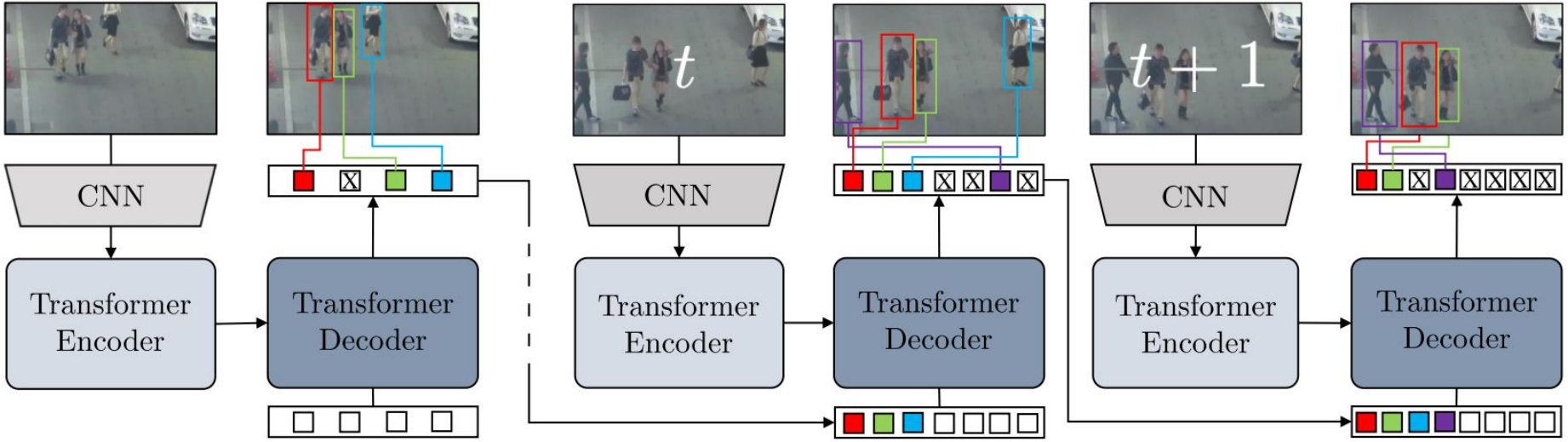
- Kieritz, Hilke, Wolfgang Hubner, and Michael Arens. "Joint detection and online multi-object tracking." CVPRW 2018.
- PTAV: Heng Fan and Haibin Ling. "Parallel Tracking and Verifying: A Framework for Real-Time and High Accuracy Visual Tracking." ICCV (2017).
- Zhu Z, Huang G, Zou W, et al. Uct: Learning unified convolutional networks for real-time visual tracking CVPR 2017.
- **SiamRPN:** Bo Li, Wei Wu, Zheng Zhu, Junjie Yan. "High Performance Visual Tracking with Siamese Region Proposal Network." CVPR2018
- FlowTrack: Zheng Zhu, Wei Wu, Wei Zou, Junjie Yan. "End-to-end Flow Correlation Tracking with Spatial-temporal Attention." CVPR (2018).
- **DaSiamRPN**: Zheng Zhu, Qiang Wang, Bo Li, Wu Wei, Junjie Yan, Weiming Hu."Distractor-aware Siamese Networks for Visual Object Tracking." ECCV (2018).
- Milan, Anton, S. Hamid Rezatofighi, Anthony Dick, Ian Reid, and Konrad Schindler. "Online multi-target tracking using recurrent neural networks." AAAI 2017.
- **#SiamSE** Sosnovik, Ivan, Artem Moskalev, and Arnold WM Smeulders. "Scale Equivariance Improves Siamese Tracking." WACV 2021. [tweet] [code]
- **#TrackFormer** Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, Christoph Feichtenhofer, "TrackFormer: Multi-Object Tracking with Transformers" arXiv 2021.

Questions ?

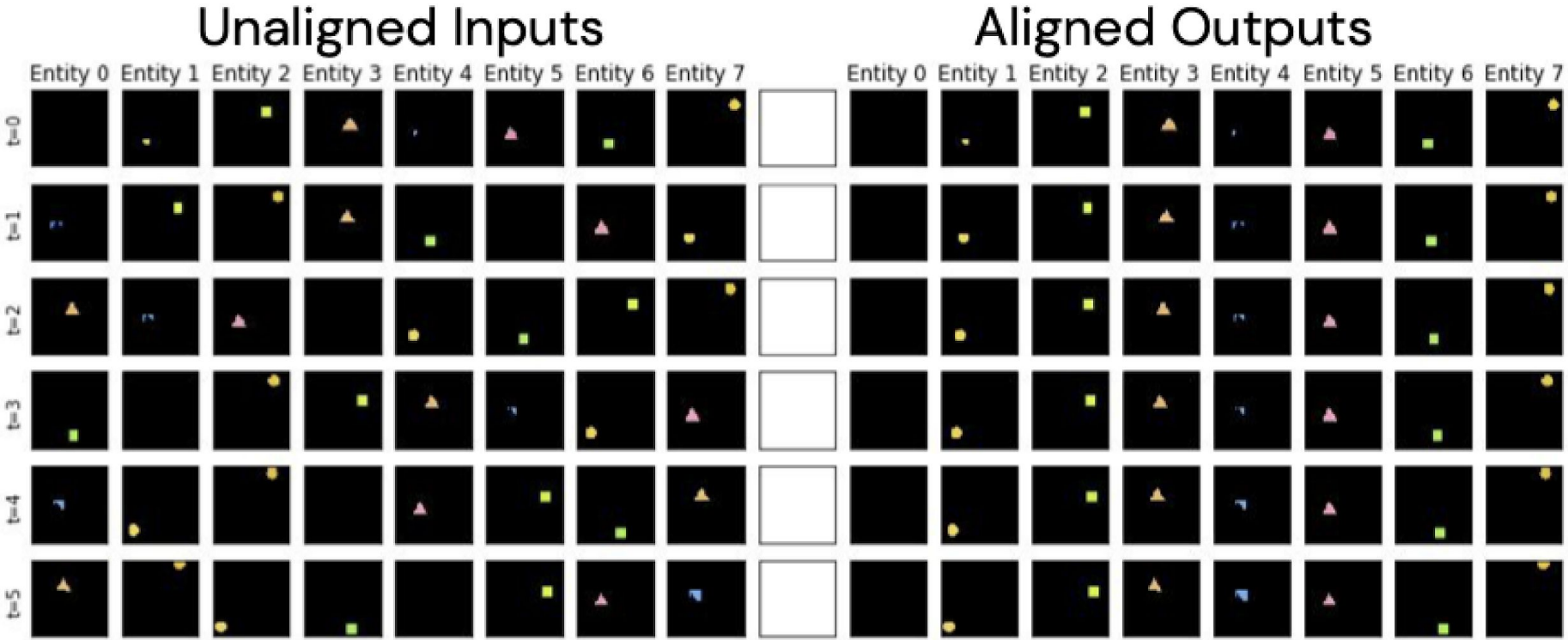

Undergradese
What undergrads ask vs. what they're REALLY asking
"Is it going to be an open book exam?"
Translation: "I don't have to actually memorize anything, do I?"
"Hmm, what do you mean by that?"
Translation: "What's the answer so we can all go home."
"Are you going to have office hours today?"
Translation: "Can I do my homework in your office?"
"Can i get an extension?"
Translation: "Can you re-arrange your life around mine?"
"Is this going to be on the test?"
Translation: "Tell us what's going to be on the test."
"Is grading going to be curved?"
Translation: "Can I do a mediocre job and still get an A?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM



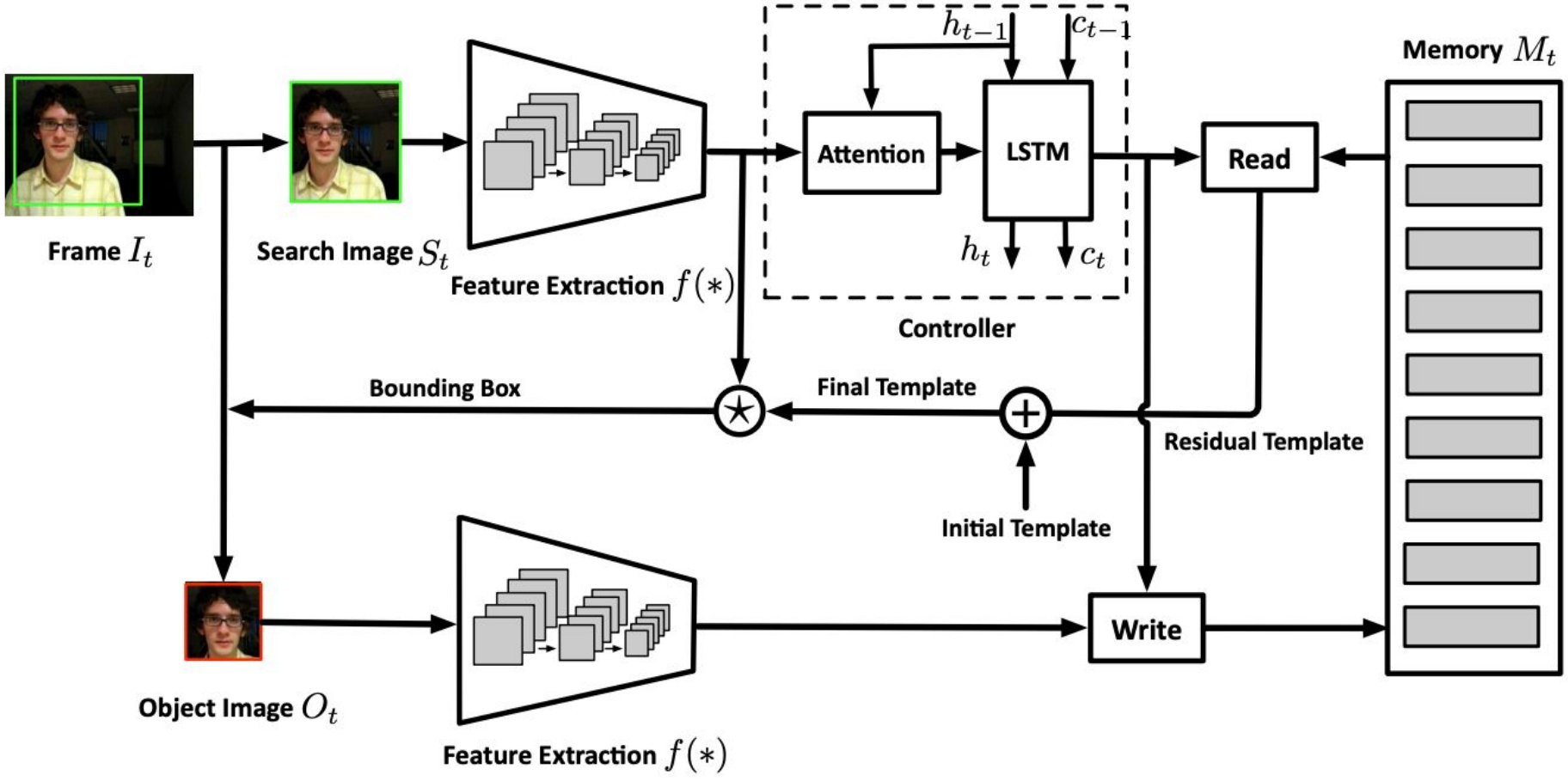
#SiamSE Sosnovik, Ivan, Artem Moskalev, and Arnold WM Smeulders. "Scale Equivariance Improves Siamese Tracking." WACV 2021. [tweet] [code]



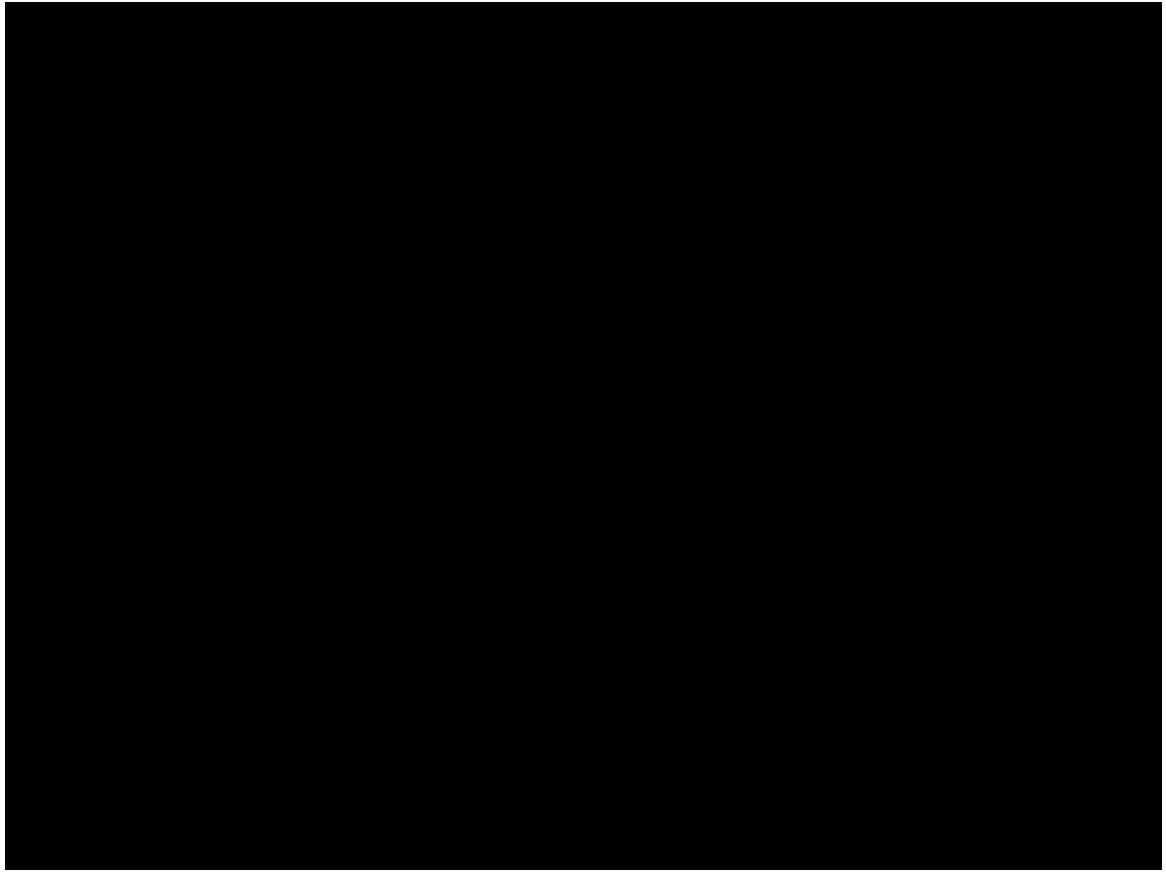
#TrackFormer Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, Christoph Feichtenhofer, "TrackFormer: Multi-Object Tracking with Transformers" arXiv 2021.



#AlignNet Creswell, A., Piloto, L., Barrett, D., Nikiforou, K., Raposo, D., Garnelo, M., ... & Shanahan, M. (2019). AlignNet: Self-supervised Alignment Module. Openreview 2020.



Yang, T., & Chan, A. B. (2018). Learning dynamic memory networks for object tracking. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 152-167).



#SE-SiamFC Scale Equivariance Improves Siamese Tracking. arXiv 2020https://twitter.com/DocXavi/likes. [tweet]