Fully-Convolutional Siamese Networks for Object Tracking

Luca Bertinetto*, Jack Valmadre*,
João Henriques, Andrea Vedaldi and
Philip Torr

www.robots.ox.ac.uk/~luca

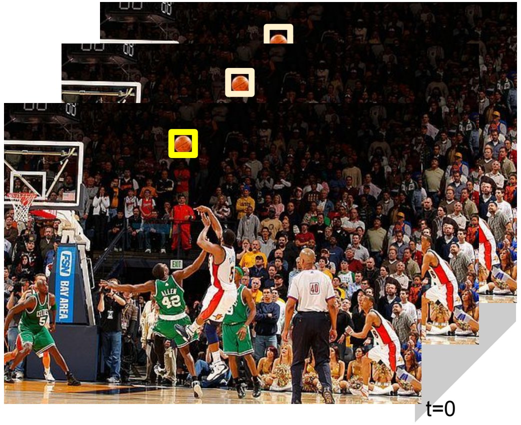
luca.bertinetto@eng.ox.ac.uk

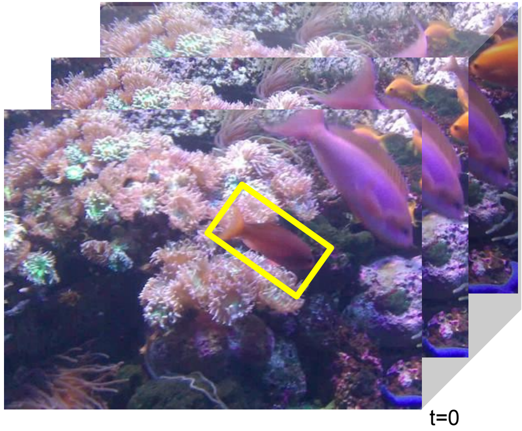
UNIVERSITY OF OXFORD

Tracking of single, arbitrary objects

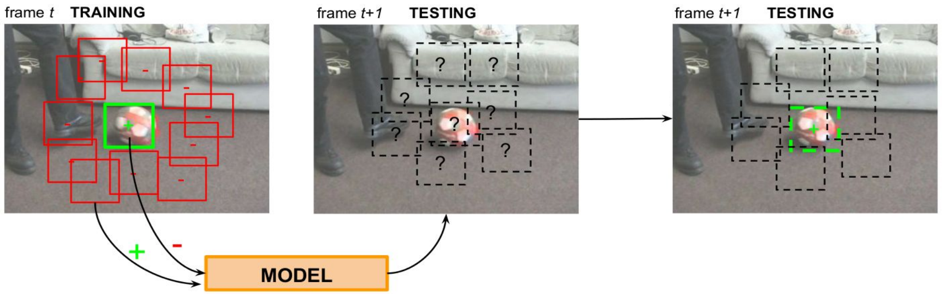
Problem. Track an arbitrary object with the sole input of a single bounding box in the first frame of the video.

Challenge: we need to be *class-agnostic.*





Tracking-by-detection paradigm

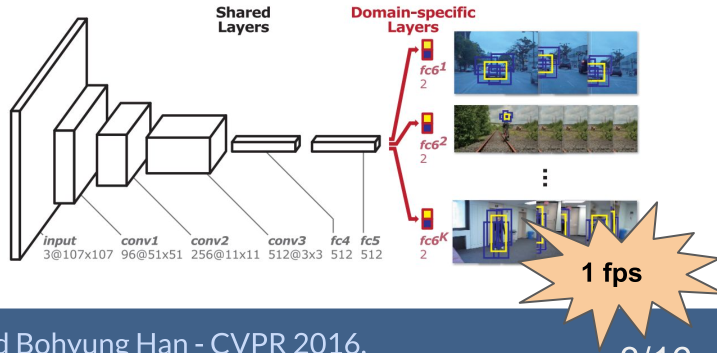


- Learn <u>online</u> a binary classifier (**+** is object, **-** is background).
- Re-detect the object at every frame + update the classifier
    - Online training and testing.

What about the deep learning frenzy?

- In tracking community, deep-nets took more time to become mainstream.
  - CVPR’15 - not a single tracker was using deep-nets as a core component and not even deep features.
  - CVPR’16 - 50% were.

- Sometimes better performance than legacy features, **but** …
- Training on benchmarks → controversial.
- Slow



Learning Multi-Domain Convolutional Neural Networks for Visual Tracking - Hyeonseob Nam and Bohyung Han - CVPR 2016.

Our work

UNIVERSITY OF OXFORD

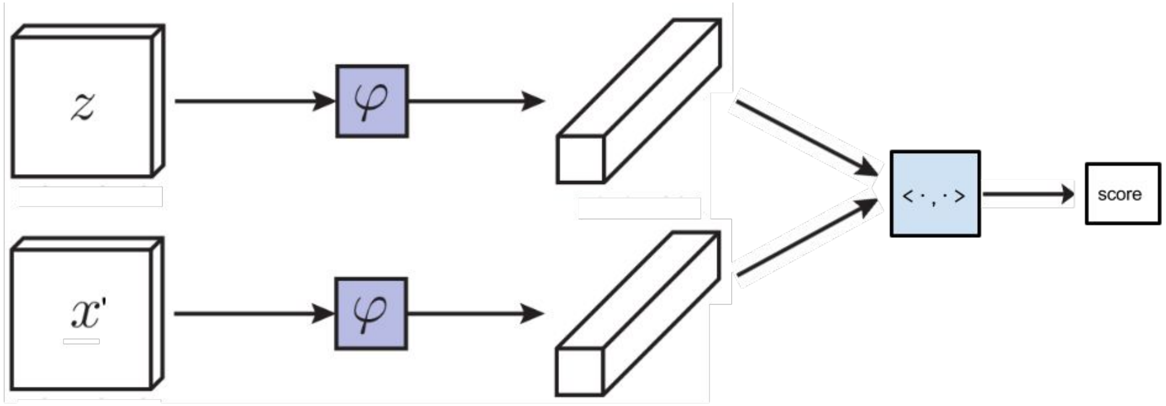
Conv-nets for arbitrary object tracking, with three constraints.

1. Real-time.
2. No benchmark videos for training.
3. Simple.

Vanilla siamese conv-net

UNIVERSITY OF OXFORD

- Trains a model to address a similarity learning problem.
- Function compares an exemplar z to a candidate of the same size x'.
- Output score tell us how similar are the two image patches.



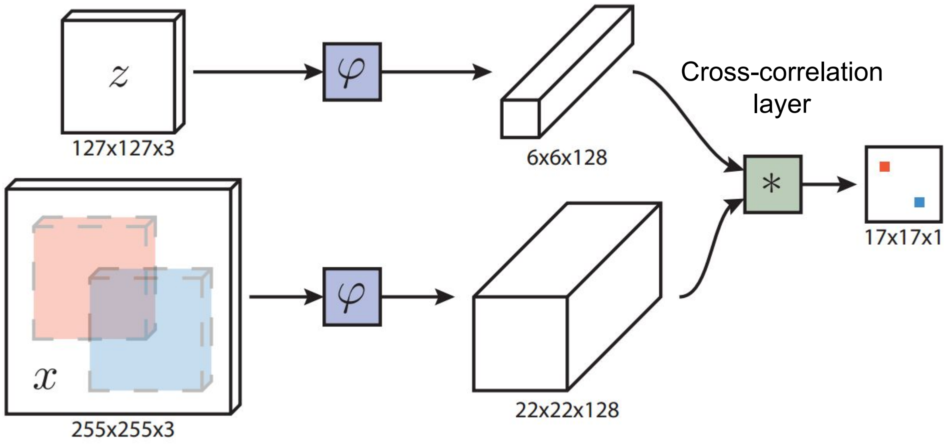
Our architecture

- Our network is **fully convolutional**.

- Two inputs of different sizes:
  - smaller (exemplar / target-object).
  - bigger (search area).

- Cross-correlation layer: computes the similarity at all translated sub-windows on a dense grid in a single evaluation.

- Output is a score map.



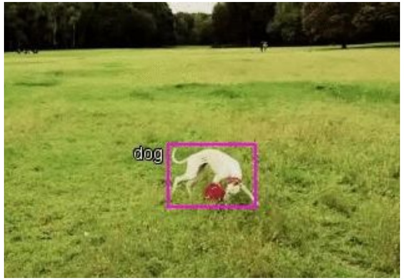
Forward pass: **>100Hz**

ILSVRC15-VID (ImageNet Video)

UNIVERSITY OF OXFORD

- So far tracking community could not rely on large labelled dataset.
  - ALOV+OTB+VOT in total have less than 600 video, with some overlap.
  - Not all labelled per frame.

- ImageNet Video
  - Official task is object detection from video - can be easily adapted to arbitrary object tracking.
  - Almost **4,500 videos** and **1,200,000 bounding boxes**!
  - 30 classes: mostly animals (~75%) and some vehicles (~25%)



Training

- Dataset build by extracting two patches with different amount of context for every labelled object. Then resized to 127x127 and 255x255.

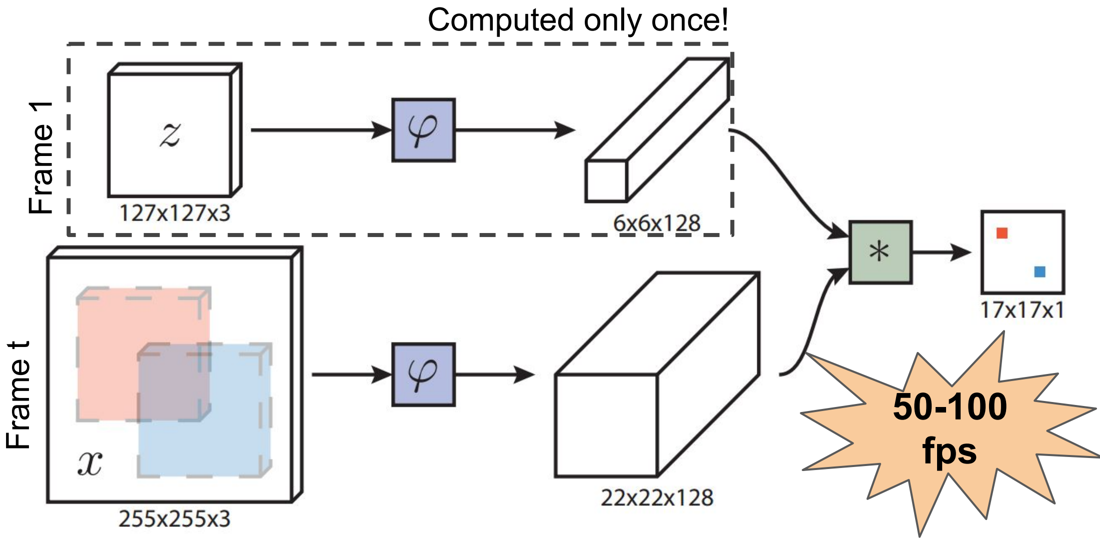
- Pick random video and random pair of frames within the video (max N frames apart).
    - N controls the “difficulty” of the problem.

- Mean of logistic loss over all positions,

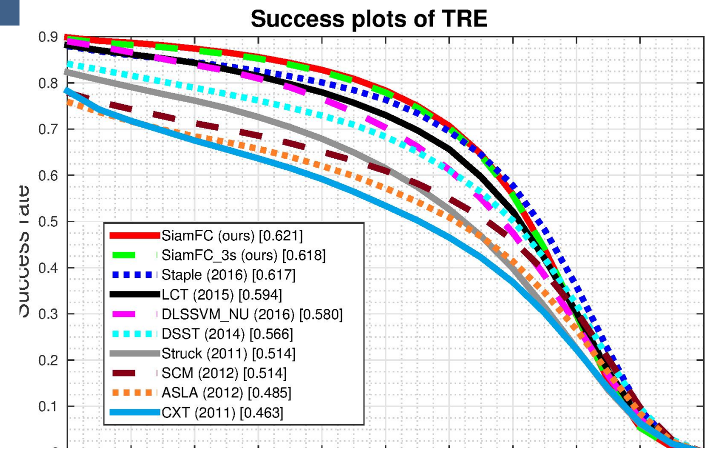
$$\ell(y, v) = \log(1 + \exp(-yv))$$

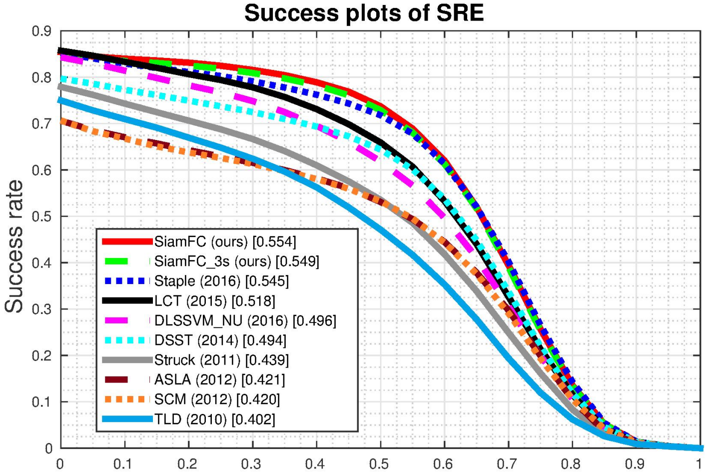
Tracking pipeline

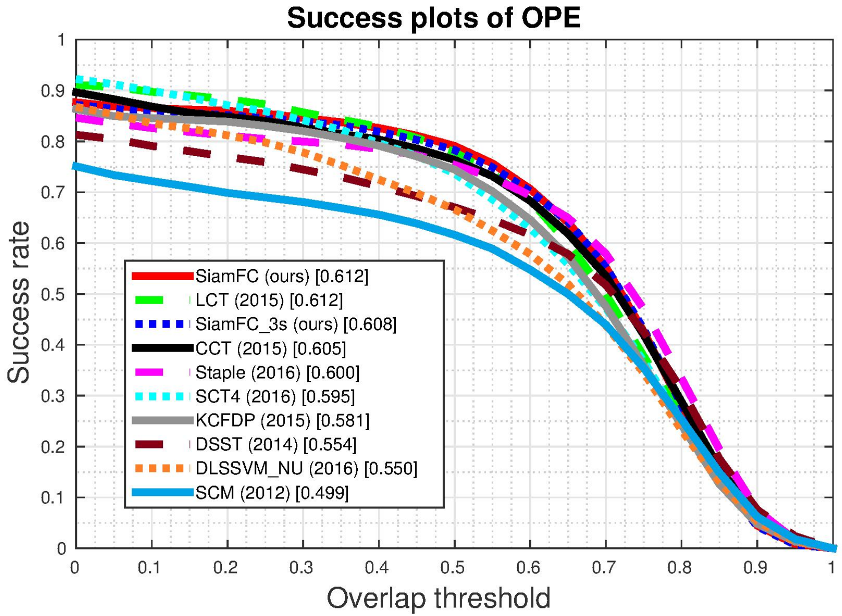
- Activations for the exemplar **z** only computed for first frame.
- Subwindow of **x** with max similarity sets the new location.
- That's (almost) it!
    - No update of target representation.
    - No bbox regression.
    - No fine-tuning → fast!
- Only three little tricks:
    - Pyramid of 3 scales.
    - Response upsamped with bi-cubic interpolation.
    - Cosine window to penalize large displacements.



New state-of-the art for real-time trackers (OTB-13)

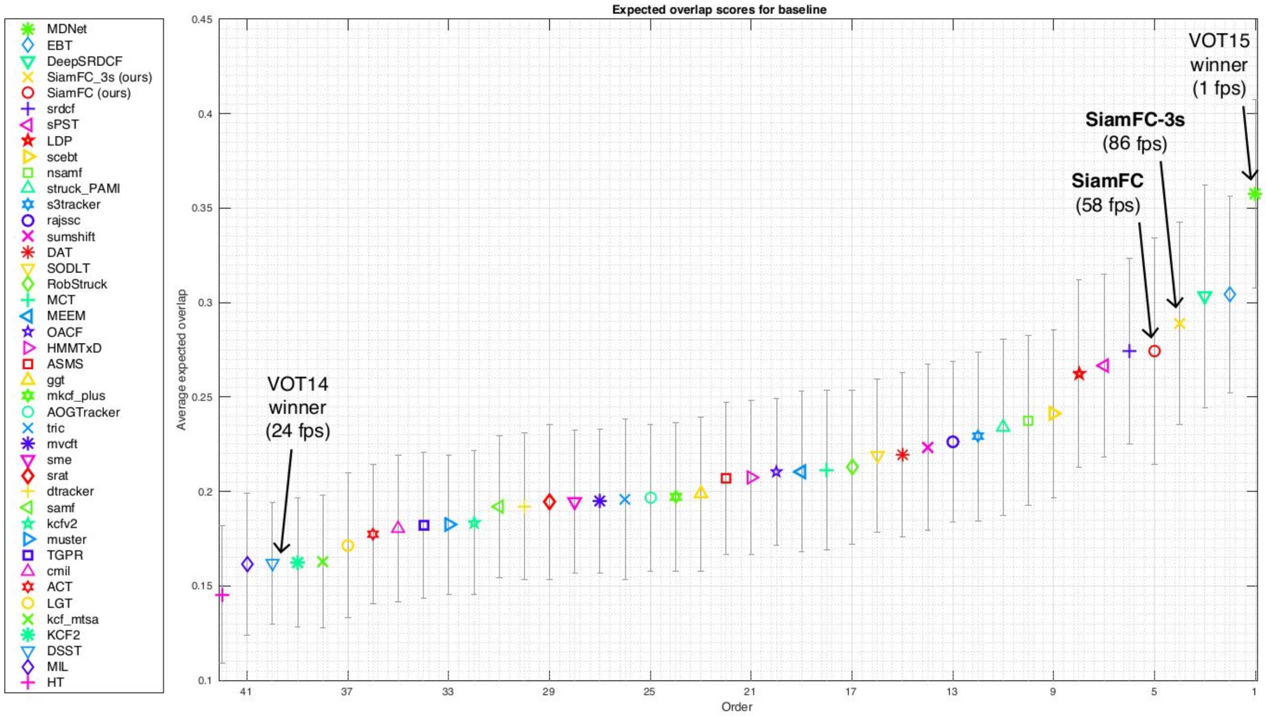




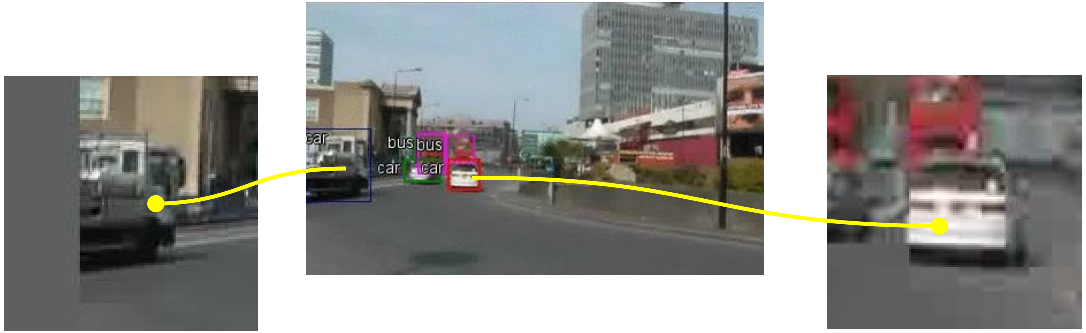


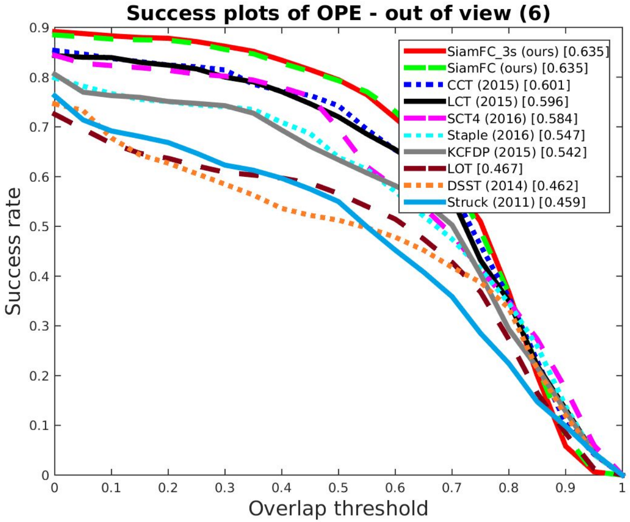
State-of-the-art for general trackers (VOT-15)

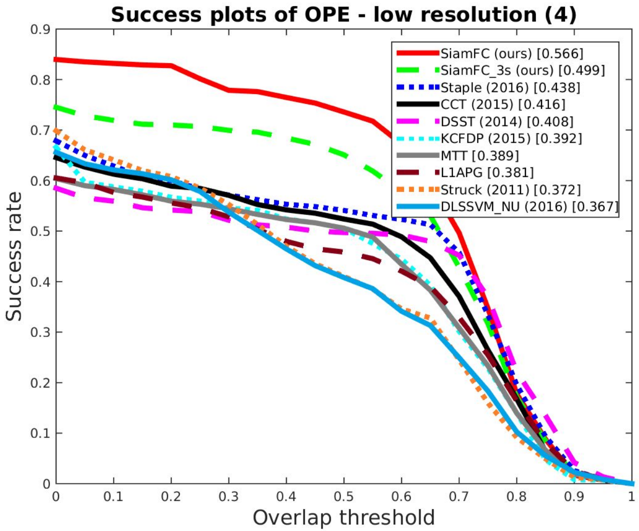
- At 1 fps, the best tracker is almost 2 orders of magnitude slower of our method, which runs at 86 frames per second.
- None among the top-15 trackers operate above 20 frames per second.



Results reflect training dataset bias

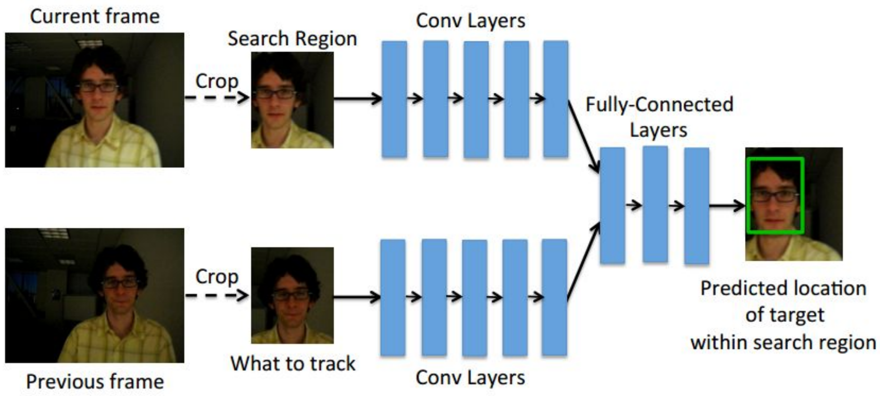






Concurrent work - GOTURN [ECCV `16]

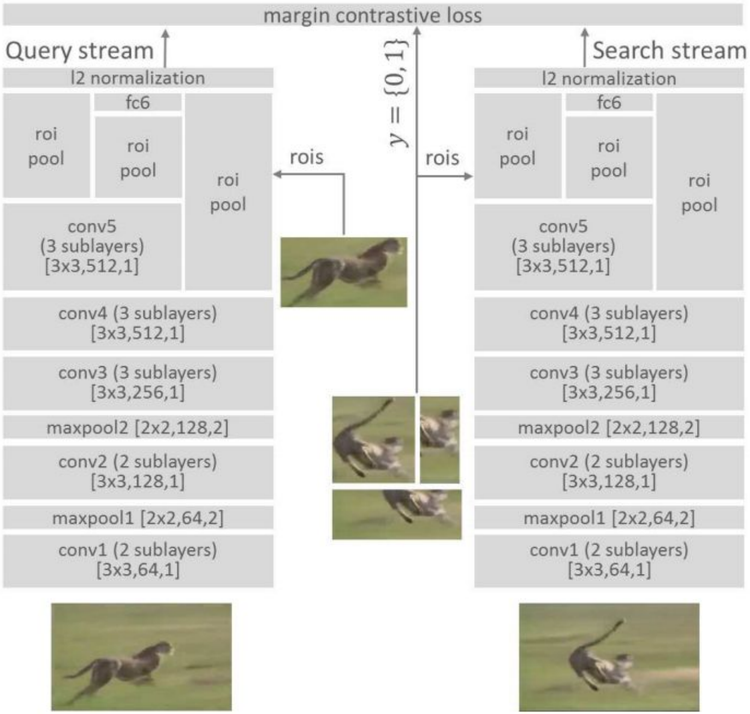
- Siamese architecture trained to solve Bounding Box regression problems.
- Differently, network is not fully convolutional.
- Trained from consecutive frames.
- They are not strictly learning a similarity function - method works (albeit worse) also with a single branch.
- Fast (100fps), but much lower results compared to our method (only VOT-14 available).



Learning to Track at 100 FPS with Deep Regression Networks - David Held, Sebastian Thrun, Silvio Savarese - ECCV 2016.

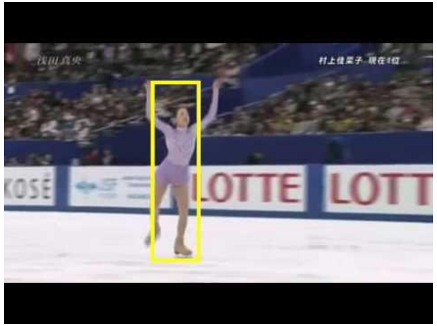
Concurrent work - SINT [CVPR `16]

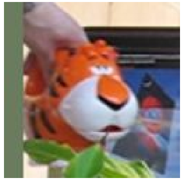
- Siamese architecture trained to learn a generic similarity function.
- Differently, their network is not fully convolutional and they recur instead to **ROI pooling** to sample candidates.
- Results reported only on OTB-13: relative +2% better than our method.
- BBox regression to improve tracking performance.
- Much slower: only **2 fps** vs **85 fps** of our method.



Siamese Instance Search for Tracking - Ran Tao, Efstratios Gavves, Arnold W.M. Smeulders - CVPR 2016.

Few examples





Conclusions

- ImageNet Video: new standard for training tracking algorithms?

- **Fully-convolutional siamese:**
    - Generalizes well (trained on ImageNet Video).
    - allows very high frame-rates, still achieving state-of-the-art performance.
    - Simple+efficient building block for future work.

→ **Code available:** `www.robots.ox.ac.uk/~luca/siamese-fc.html`

Thank you.