UNIVERSITY OF
OXFORD

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

Tracking of single, arbitrary objects

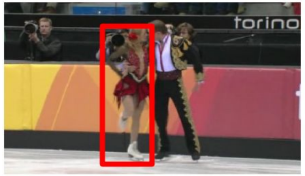
UNIVERSITY OF OXFORD

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

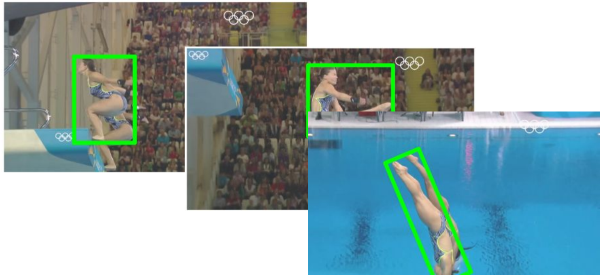
Challenges.

- We need to be *class-agnostic.*
- Stability-Plasticity dilemma[Grossberg87]

> *"How can a learning system remain plastic in response to significant new events, yet also remain stable in response to irrelevant events?"*

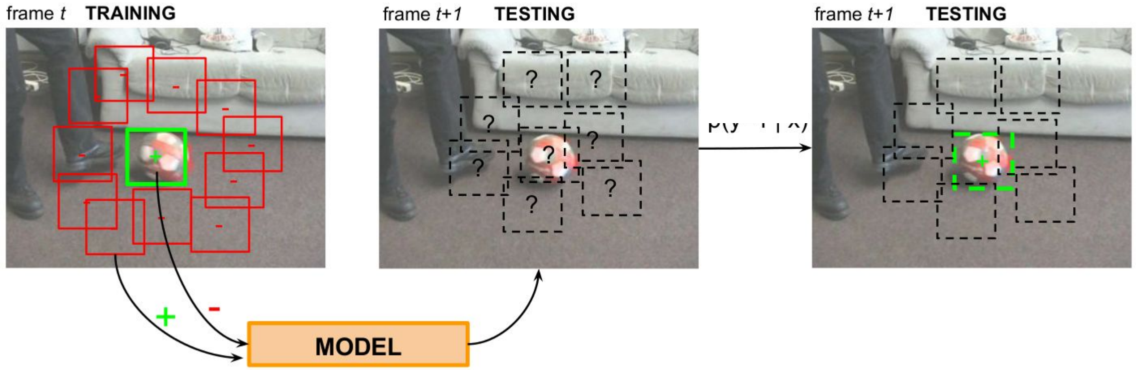






Recent history of object tracking [2010 - today]

Tracking-by-detection paradigm



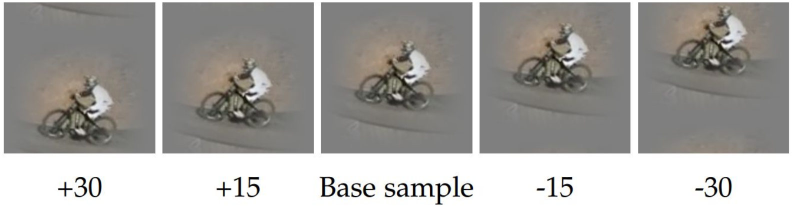
- Learn <u>online</u> a binary classifier (**+** is object, **-** is background).
- Re-detect the object at every frame + update the classifier.

Recent history of object tracking [2014 - today]

UNIVERSITY OF OXFORD

Correlation filters become the most popular choice

- Sampling space is loosely a circulant matrix → diagonalized with Discrete Fourier Transform.



From [Henriques15]

- Fast training and evaluation of linear classifier in the Fourier Domain.
- Mostly used with HOG features.

Recent history of object tracking [2015 - today]

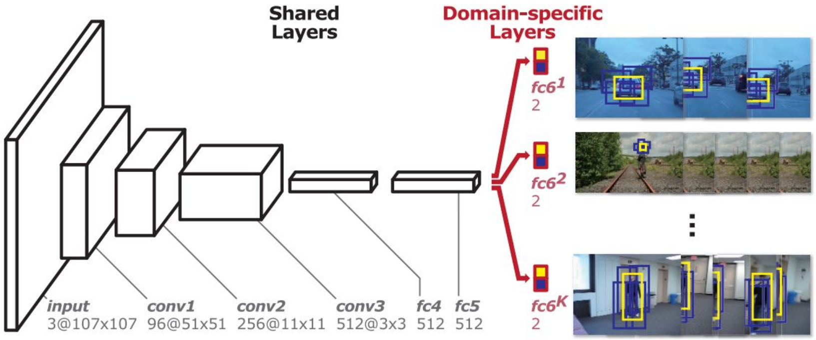
UNIVERSITY OF OXFORD

What about the deep learning frenzy?

- In tracking, deep-nets took more time to become mainstream.
  - CVPR’15 - not a single tracker was using deep-nets as a core and not even deep features.
  - CVPR’16 - 50% were.
- Not clear advantage
  - Slow
  - Similar performance to methods based on legacy features.
- Training on benchmarks → controversial.
  - Benchmarks propose very similar scenarios. Risk to overfit and lack of generalization.

MDNet [CVPR16, winner of VOT15]

- Best results so far.
- Rationale: separate domain-independent (e.g. the concept of “objectness”) to domain-dependent (video-specific) information.
- **Training**. fixed common part (3conv+2fc) and several “one-hot” fc branches.
- **Tracking**. fine-tuning of several layers, hard-negative mining, bbox regression.



- Best results so far.
- Trained from benchmarks video.
- Very slow.

1 fps

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

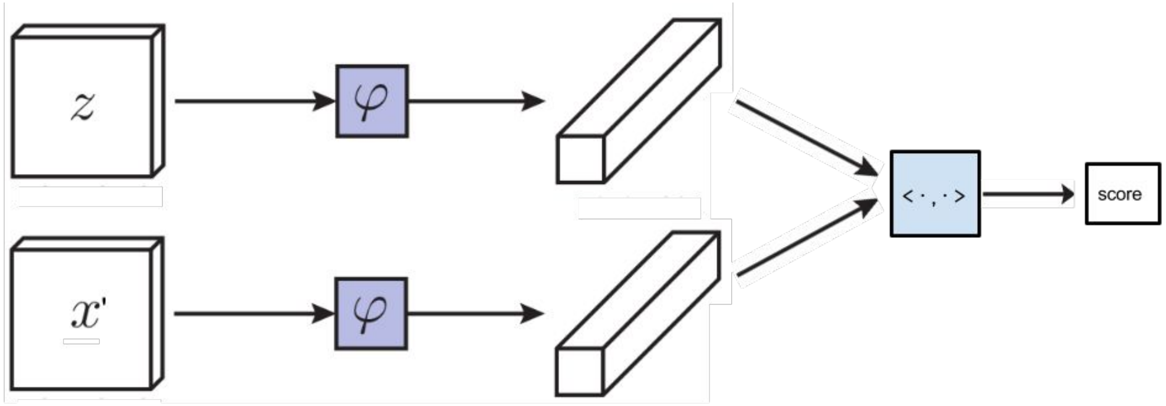
Our work

- We wanted to use conv-nets for arbitrary object tracking

- Three constraints
    - No below real-time (at least 20-25 frames per second).
    - No benchmark videos for training.
    - Simplicity.

Vanilla siamese conv-net for similarity learning

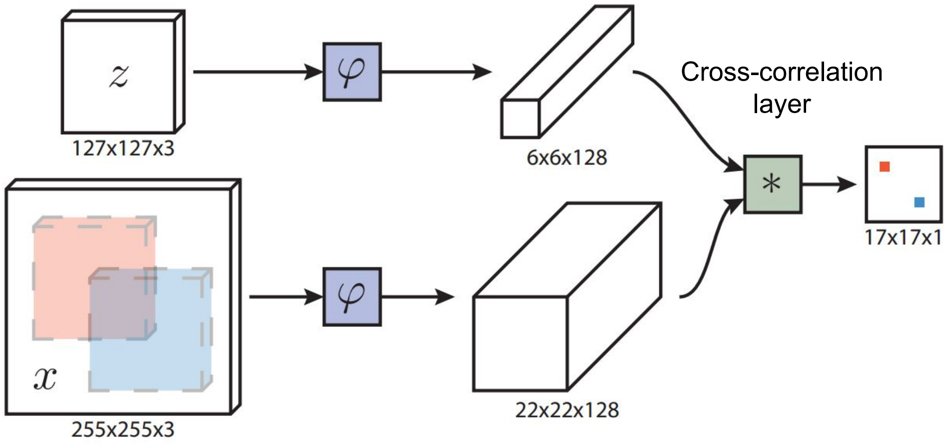
- Siamese conv-net trained to address a similarity learning problem in an offline phase.
- The conv-net learns a function that compares an exemplar z to a candidate of the same size x'.
- Score tell us how similar are the two image patches.



Fully-Convolutional Siamese Networks for Object Tracking

UNIVERSITY OF OXFORD

- Our network is **fully convolutional**.
    - No padding.
    - No fully-connected layers.
- Two inputs of different sizes: smaller is the exemplar (target object during tracking), bigger is the search area.
- Output of embedding function has spatial support.
- 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**

CODE AVAILABLE! `www.robots.ox.ac.uk/~luca/siamese-fc.html`

Training

- Dataset build by extracting two patches with +/- 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 at every position,

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

ILSVRC15-VID (ImageNet Video)

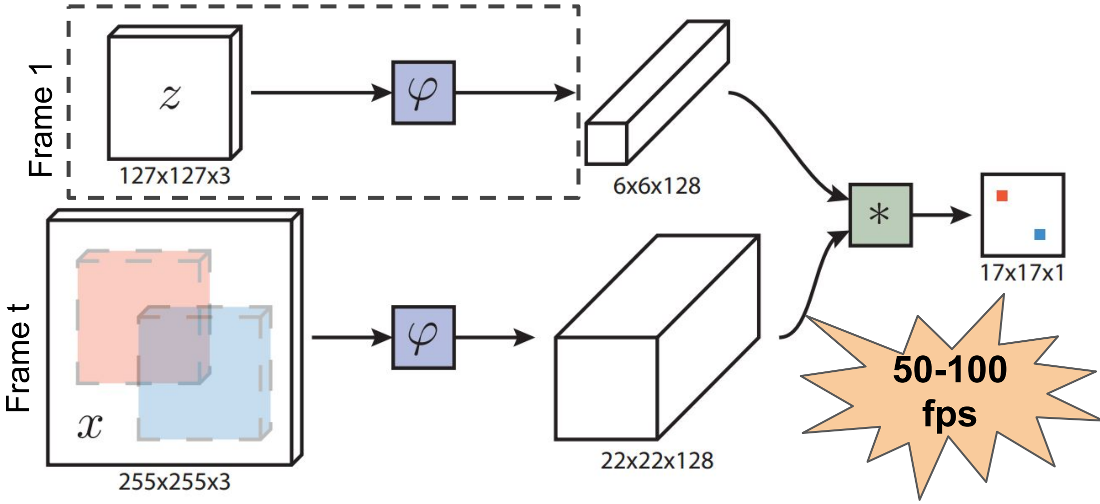
- So far tracking community could not rely on large labelled dataset.
    - ALOV+OTB+VOT in total have less than 600 video, with some overlap.
    - They should be reserved for the purpose of testing.

- ImageNet Video
    - Official task is object detection and classification from video.
        - Step-by-step guide to prepare the data to train our net: https://github.com/bertinetto/siamese-fc/tree/master/ILSVRC15-curation
    - Almost **4,500 videos** and **1,200,000 bounding boxes**!
    - 30 classes: mostly animals (~75%) and some vehicles (~25%)



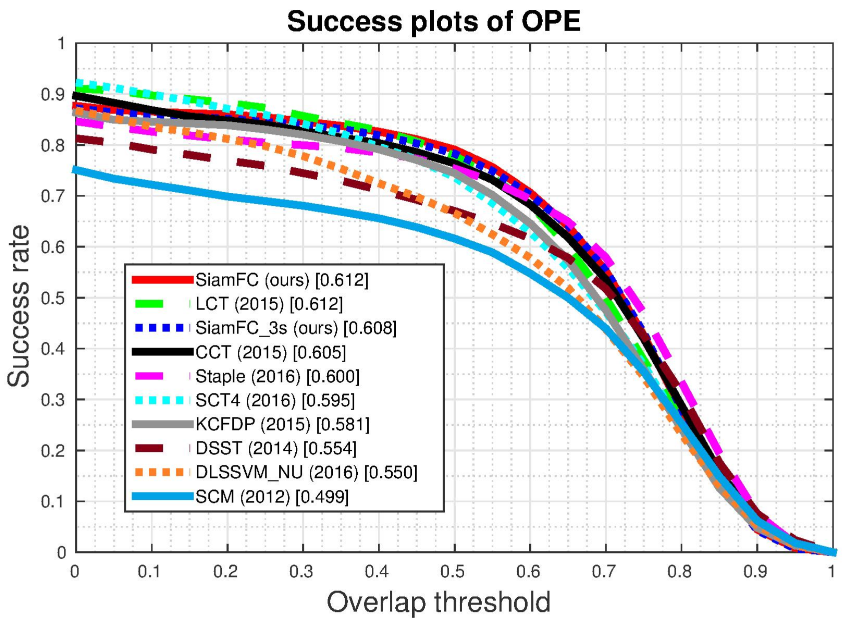
Tracking pipeline

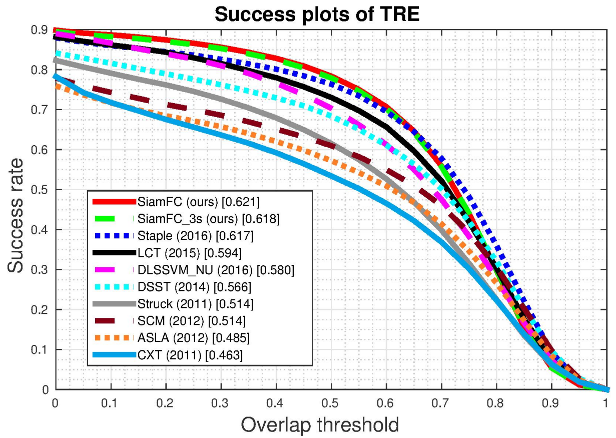
- Activations for the exemplar **z** only computed for first frame.
- Subwindow of **x** with max similarity sets the new location.
- That's (almost) it!
  - No update of target representation.
  - No re-detection.
  - No bbox regression.
  - No fine-tuning → fast!
- Only three little tricks:
  - Pyramid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

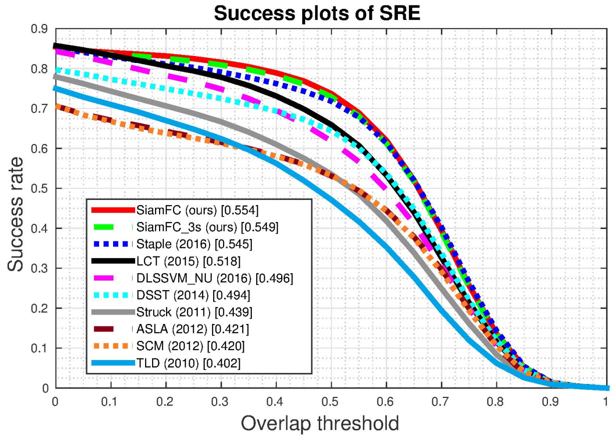


CODE AVAILABLE! `www.robots.ox.ac.uk/~luca/siamese-fc.html`

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



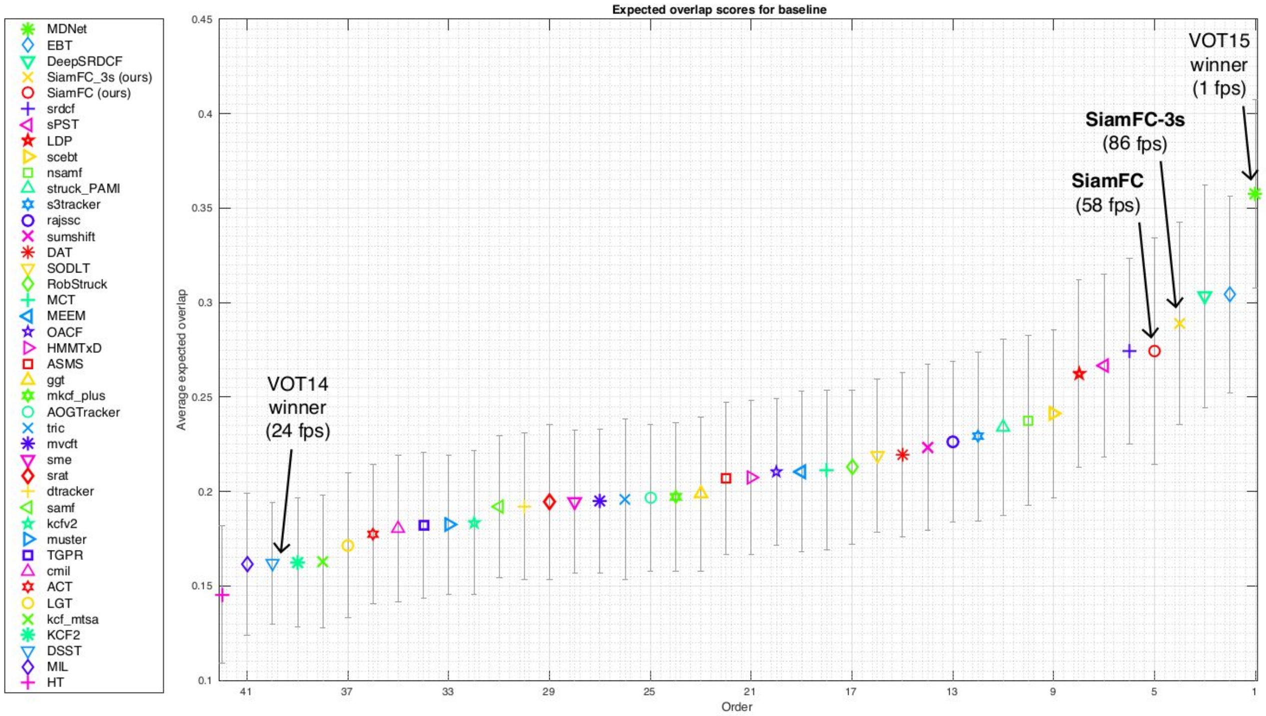




State-of-the-art for general trackers (VOT-15)

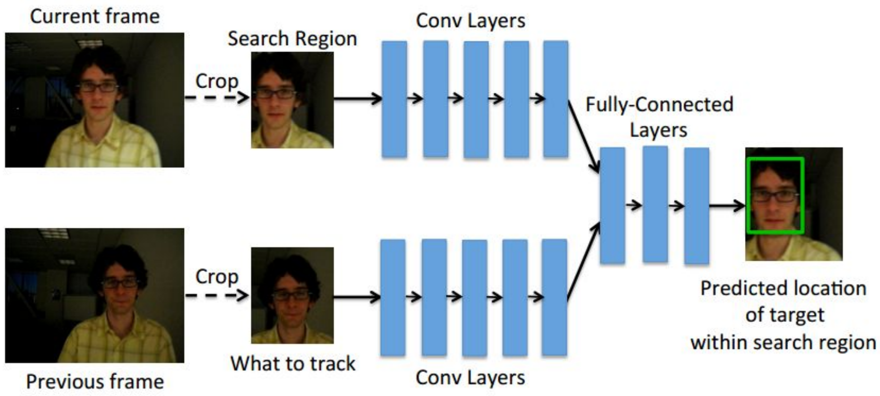
- At 1 fps, the best tracker is almost 2 orders of magnitude slower of our method, which runs at 86 frames per second.

- None among the top-15 trackers operate above 20 frames per second.



Concurrent work - GOTURN [ECCV `16]

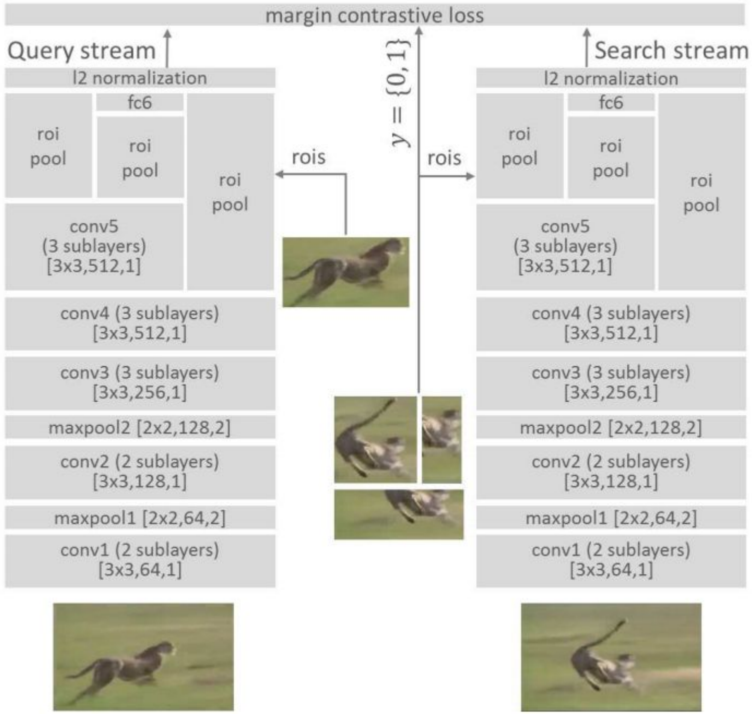
- Siamese architecture trained to solve Bounding Box regression problems.
- Differently, network is not fully convolutional.
- Trained from consecutive frames.
- They are not strictly learning a similarity function - method works (albeit worse) also with a single branch.
- Fast (100fps), but significantly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



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

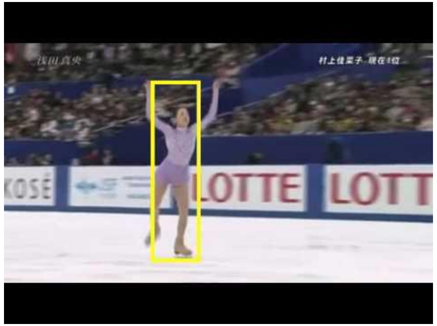
Concurrent work - SINT [CVPR `16]

- 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: ~2% better than our method.
- BBox regression to improve tracking performance.
- Much slower: only **2 fps** vs **50-85 fps** of our method.



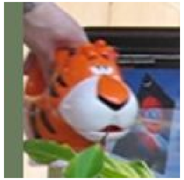
Few examples

UNIVERSITY OF OXFORD








Conclusions

- ImageNet Video: new standard for training tracking algorithms?
- Siamese networks allow simplistic trackers to achieve state-of-the-art results.
- **Fully-convolutional siamese:** allows very high frame-rates, still achieving state-of-the-art performance.
- **Fully-convolutional siamese:** simple and fast building block for future work: e.g. online update of representation.

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

Thank you.

UNIVERSITY OF
OXFORD