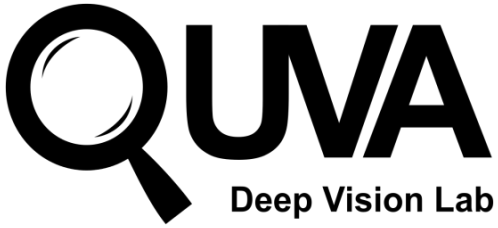
Siamese Instance Search for Tracking

Ran Tao, Efstratios Gavves, Arnold Smeulders



UNIVERSITEIT VAN AMSTERDAM

Ran Tao, Efstratios Gavves, Arnold Smeulders. *Siamese Instance Search for Tracking*. CVPR, 2016, Las Vegas.

(Single) Visual Object Tracking

Track the target's positions over time in a video, given a starting box in 1st frame

Applications

- Surveillance
- Robotics
- Human-computer Interaction
- Autonomous Driving
- Drones

Tracking is hard

- Start from 1 snapshot of the target
- But the target may change its appearance significantly due to illumination variation, scale change, rotation, etc. *[Smeulders et al, TPAMI, 2014: 13 hard aspects]*

- Track the ‘thing’ in the bounding box (i.e. unknown object)
- Unknown environment

How to handle the appearance variations of the target?

Prevalent paradigm in literature

Starting from the 1st frame, learn and update a target model on-the-fly

- **Target model**: target/non-target binary classifier, regressor
- **Update** the model using the data inferred by the tracker itself

Prevalent paradigm in literature

Starting from the 1st frame, learn and update a target model on-the-fly

- **Target model**: target/non-target binary classifier, regressor
- **Update** the model using the data inferred by the tracker itself

The data inferred by the tracker itself are not absolutely reliable → drifting

The proposed tracker: motivation

Since the only reliable data is the initial target region in the first frame, the proposed tracker only relies on the initial target. (no update)

The proposed tracker: motivation

Since the only reliable data is the initial target region in the first frame, the proposed tracker only relies on the initial target. (no update)

Then how to handle the appearance variations?

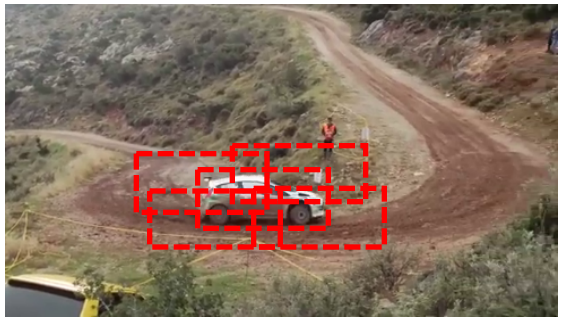
The proposed tracker: motivation

Since the only reliable data is the initial target region in the first frame, the proposed tracker only relies on the initial target. (no update)

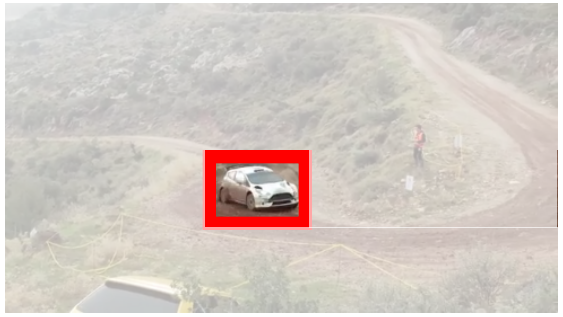
Then how to handle the appearance variations?

Certain objects change appearance over time in a similar way. ➔
Can we learn a comparison mechanism (similarity metric) a priori, that is robust against typical appearance variations an object may have in videos?

Siamese INstance search Tracker (SINT)



candidate patches



Initial patch

$$\underset{j}{\text{argmax}}\ m(initialpatch, patch_j)$$

predicted patch

Siamese INstance search Tracker (SINT)

Simply tracks the target by retrieving in every frame the candidate most similar to the initial target in the first frame

- No online updating
- No occlusion detection
- No geometric matching
- No combination of trackers

But still delivers state-of-the-art tracking performance (at the publication time).

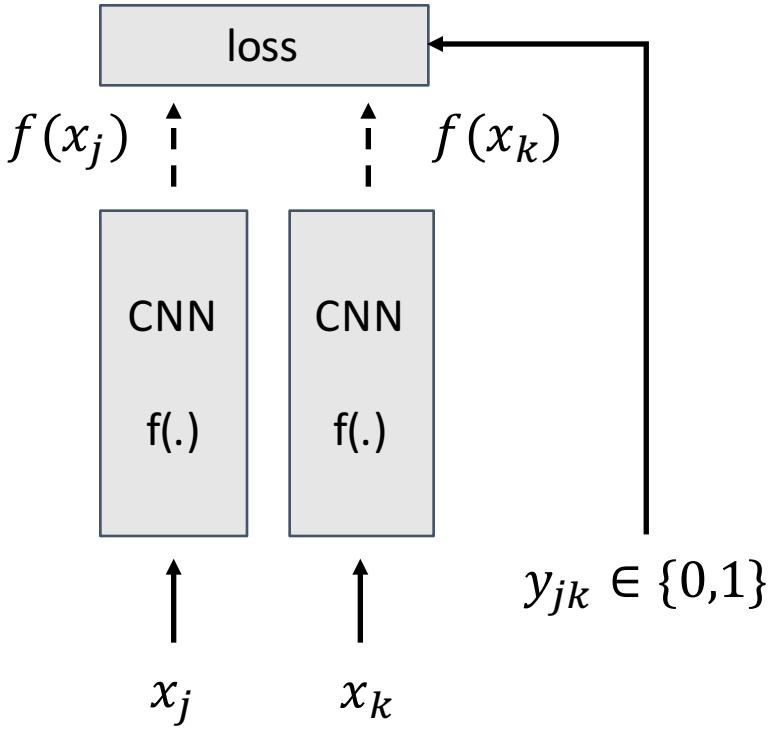
Strength is from the similarity function $m(\cdot,\cdot)$ learned offline using **Siamese network**.

Siamese INstance search Tracker (SINT)

Learn once on a rich video dataset with box annotations following an object.

Once learned, it is applied as is, without any further adapting, to track any previously unseen targets.

Similarity Function Learning



Marginal Contrastive Loss:

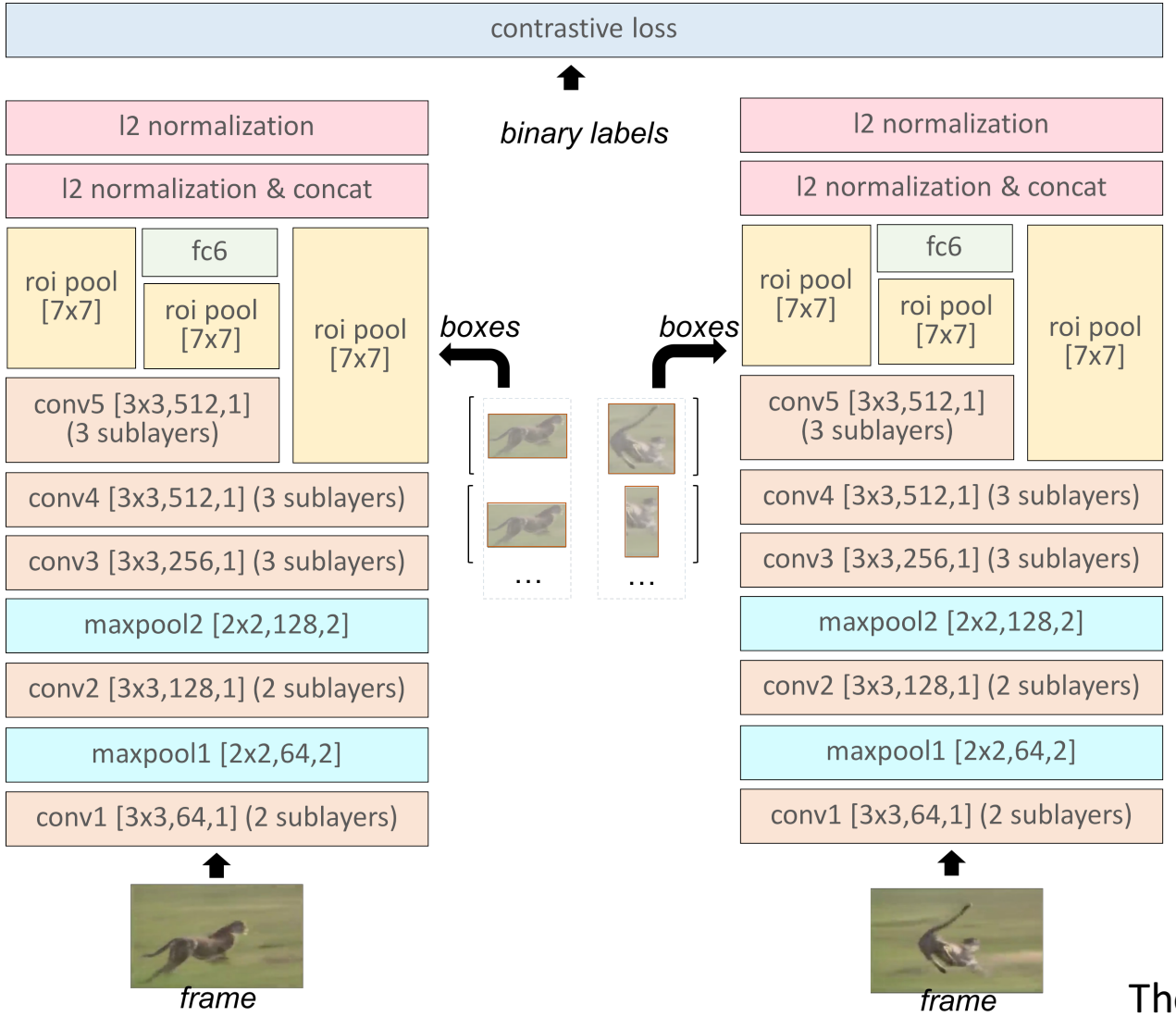
$$L(x_j, x_k, y_{jk}) = \frac{1}{2} y_{jk} D^2 + \frac{1}{2}(1 - y_{jk}) \max(0, \sigma - D^2)$$

$$D = \left\| f(x_j) - f(x_k) \right\|_2$$

Similarity function (after learning):

$$m(x_j, x_k) = f(x_j) \cdot f(x_k)$$

Network Architecture

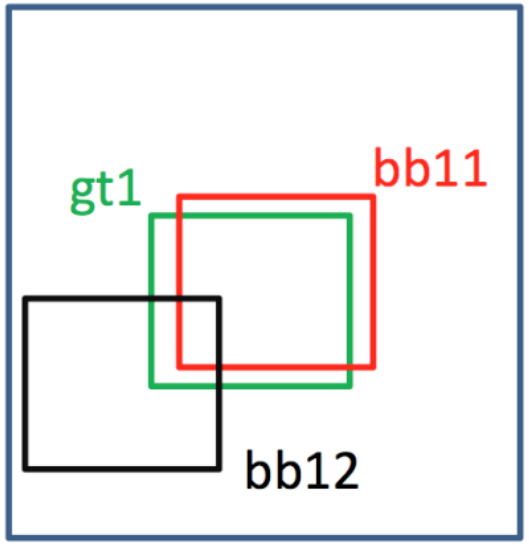


- Region-of-interest (ROI) pooling → process all boxes in a frame in one single pass through the network
- Very few max pooling → improve localization accuracy
- Use outputs of multiple layers (conv4_3, conv5_3, fc6) → to be robust in various situations (unknown environment)

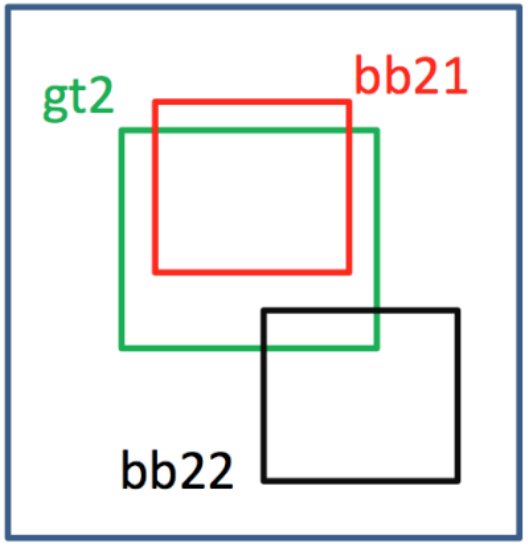
The two branches share the parameters.

Training Pairs

Data: videos of objects with BBox annotation (ALOV)



frame 1



frame 2

(gt1, gt2, 1)
(gt1, bb21, 1)
(gt1, bb22, 0)
(gt2, bb11, 1)
(gt2, bb12, 0)
...

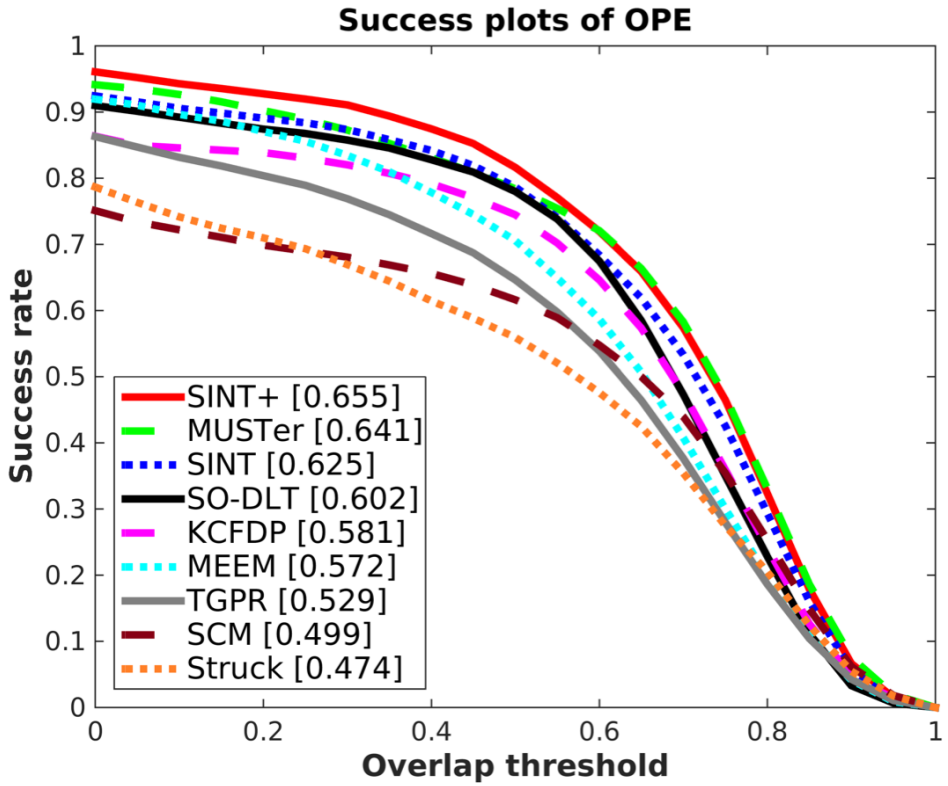
>0.7, 1
<0.5, 0

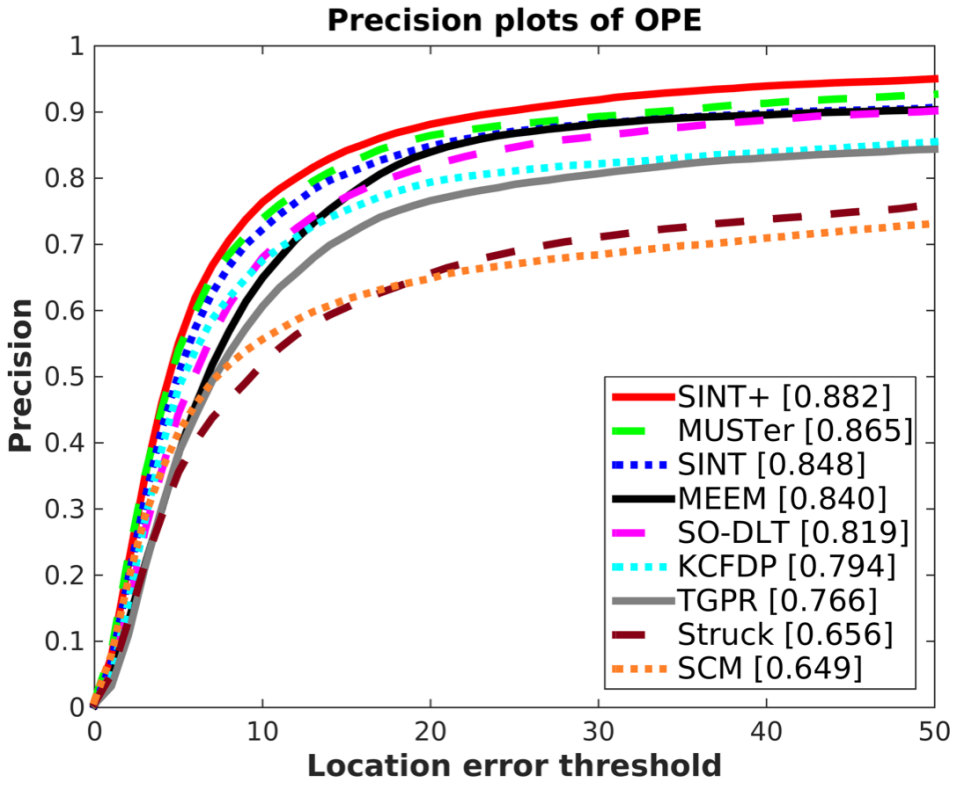
Smeulders *et al*, Visual tracking: An experimental survey, *TPAMI, 2014.*

Training Pairs

- 60,000 pairs of frames for training, 2,000 pairs for validation
- 128 pairs of boxes per pair of frames

Results on OTB





SINT+: adaptive sampling range [Want et al, ICCV15] & optical flow to remove motion inconsistent samples

Large potential to improve SINT by integrating advanced online components

Qualitative Results

Can handle various types of appearance variations

The performance on subsequent frames will not be affected by the mistake made on the current frame.

Target Re-identification

- In the absent of any drifting, SINT allows for target re-identification after the target was absent for a long period of time, provided with a sampling over the whole image.

https://youtu.be/knaxUljyY_Q

Summary

- Siamese INstance search Tracker (SINT)
  - Retrieves in every frame the patch most similar to the 1 original patch of the target, nothing else
  - The strength is from the matching function, learned offline *generically*
- Allows target re-identification after the target was absent for a complete shot
- Establish **a new tracking framework**: it only requires one-time offline learning, and once learned, it is ready to track any new, previously unseen, targets, without any online learning.