SiameseFC

Fully-Convolutional Siamese Networks for Object Tracking

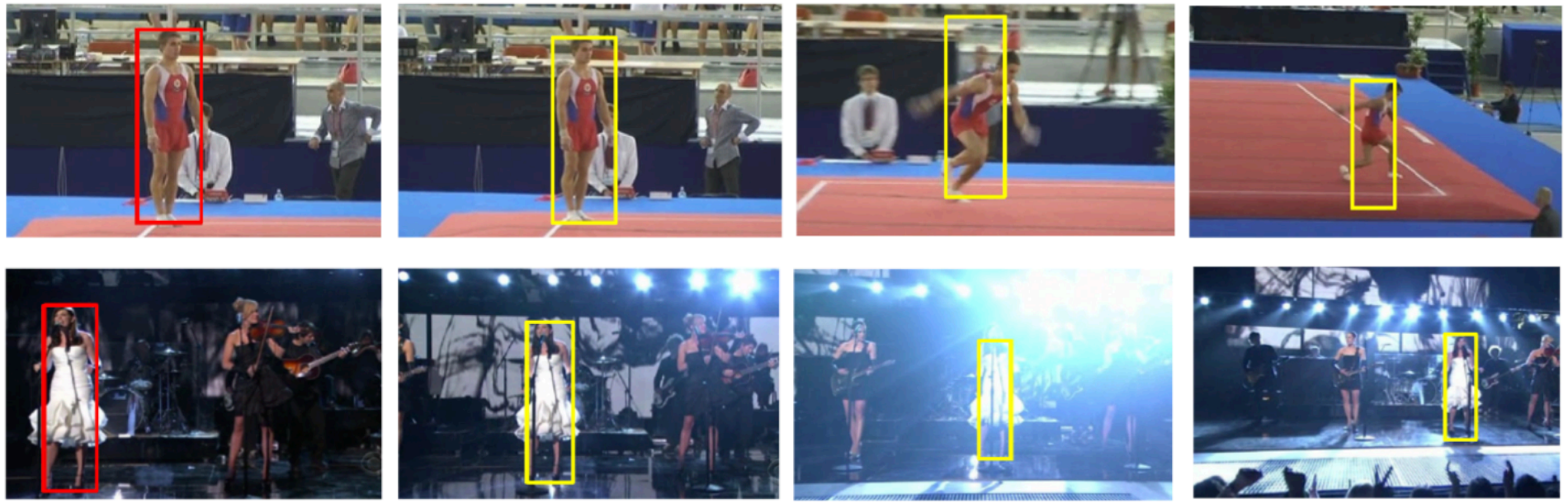
2019. 3. 4.

Jihun **Kim**, Hanyang Univ.

Introduction

Bertinetto et al. Fully-Convolutional Siamese Networks for Object Tracking. ECCV 2016.

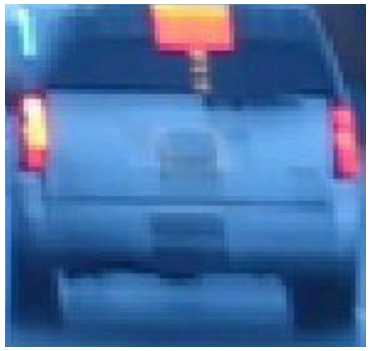
Tracking an arbitrary object in video



Frame 1 (init.) Frame 50 Frame 100 Frame 200

Similarity function

exemplar image z



candidate image x

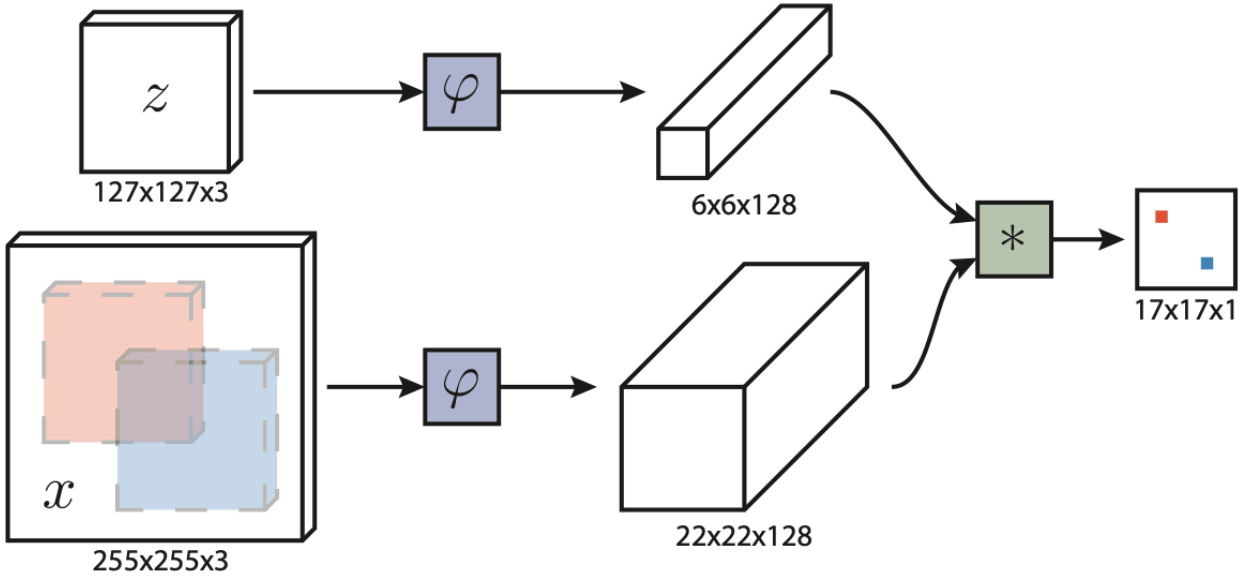
$f(z,x)$ = (high score)

$f(z,x)$ = (low score)

Structure

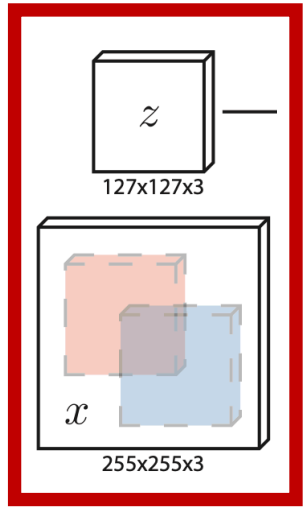
Bertinetto et al. Fully-Convolutional Siamese Networks for Object Tracking. ECCV 2016.

Deep similarity learning for tracking



$$f(z, x) = g(\varphi(z), \varphi(x))$$

Exemplar image and search image



exemplar image z

search images (mini batch)

Convolutional embedding function

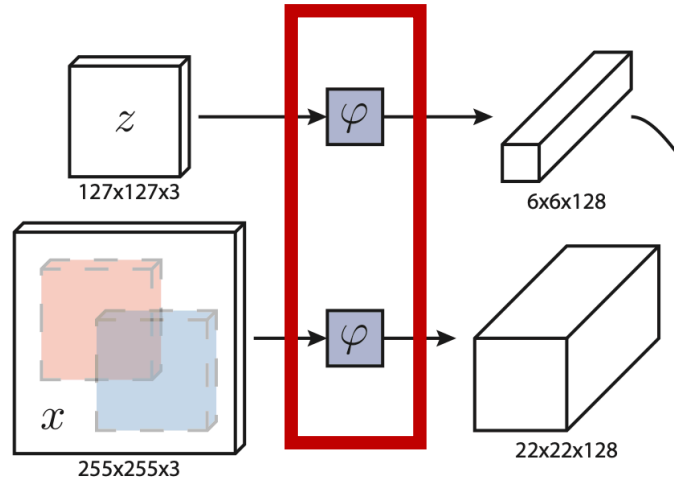
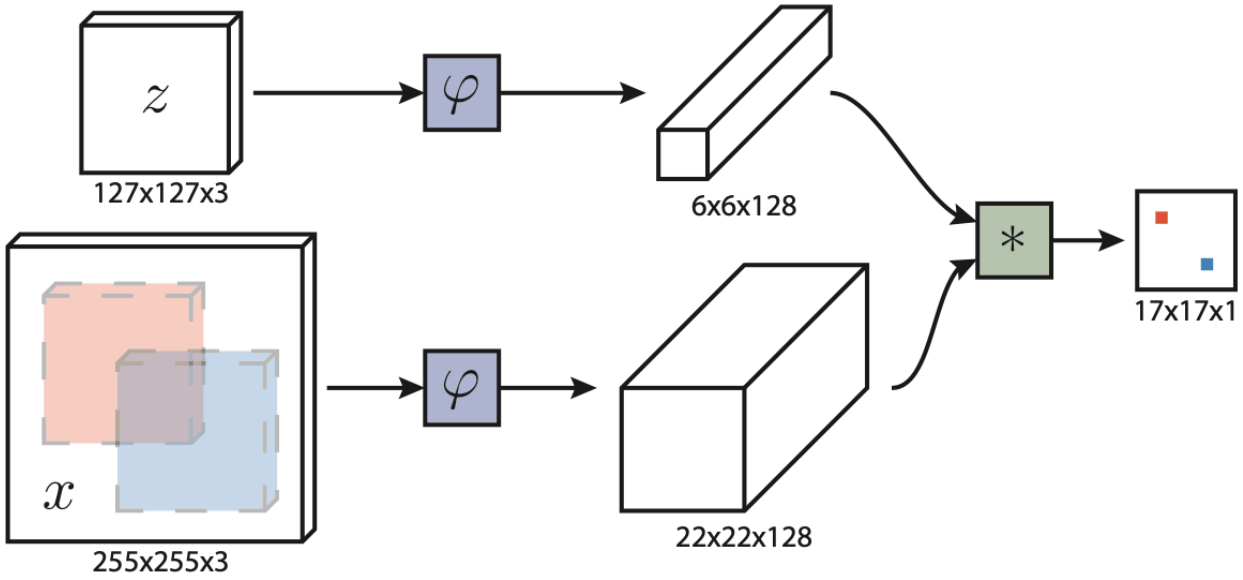


Table 1: Architecture of convolutional embedding function, which is similar to the convolutional stage of the network of Krizhevsky et al. [16]. The channel map property describes the number of output and input channels of each convolutional layer.

				Activation size		
Layer	Support	Chan. map	Stride	for exemplar	for search	chans.
				127×127	255×255	$\times 3$
conv1	11×11	96×3	2	59×59	123×123	$\times 96$
pool1	3×3		2	29×29	61×61	$\times 96$
conv2	5×5	256×48	1	25×25	57×57	$\times 256$
pool2	3×3		2	12×12	28×28	$\times 256$
conv3	3×3	384×256	1	10×10	26×26	$\times 192$
conv4	3×3	384×192	1	8×8	24×24	$\times 192$
conv5	3×3	256×192	1	6×6	22×22	$\times 128$

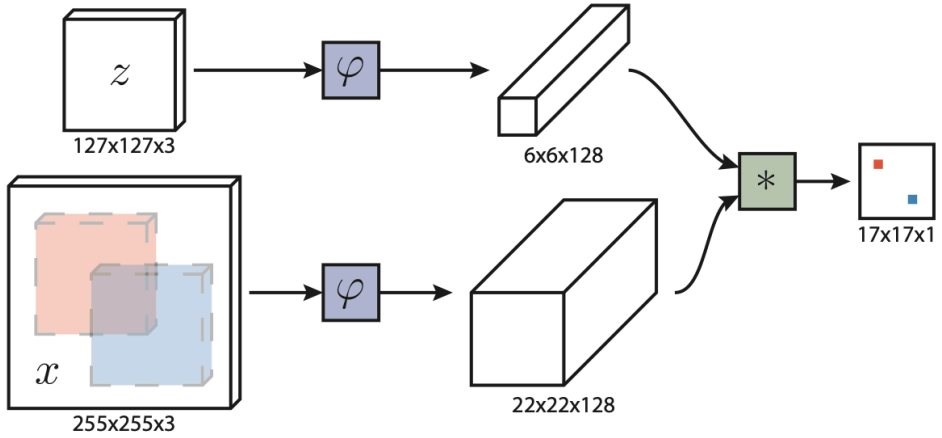
Correlation layer



Training

Bertinetto et al. Fully-Convolutional Siamese Networks for Object Tracking. ECCV 2016.

Logistic loss



Loss of a single score

$$\boldsymbol{l(y,v) = log(1 + exp(-yv))}$$

$\boldsymbol{v}$: Real-valued score of a single exemplar-candidate pair

$\boldsymbol{y}$: $y \in \{+1, -1\}$, Ground truth label

Loss of a score map

$$\boldsymbol{L(y,v) = \frac{1}{|D|}\sum_{u \in D} l(y[u], v[u])}$$

Data

Exemplar image

Corresponding search image

Fig. 2: Training pairs extracted from the same video: exemplar image and corresponding search image from same video. When a sub-window extends beyond the extent of the image, the missing portions are filled with the mean RGB value.

Results

Bertinetto et al. Fully-Convolutional Siamese Networks for Object Tracking. ECCV 2016.

The OTB-13 Benchmark

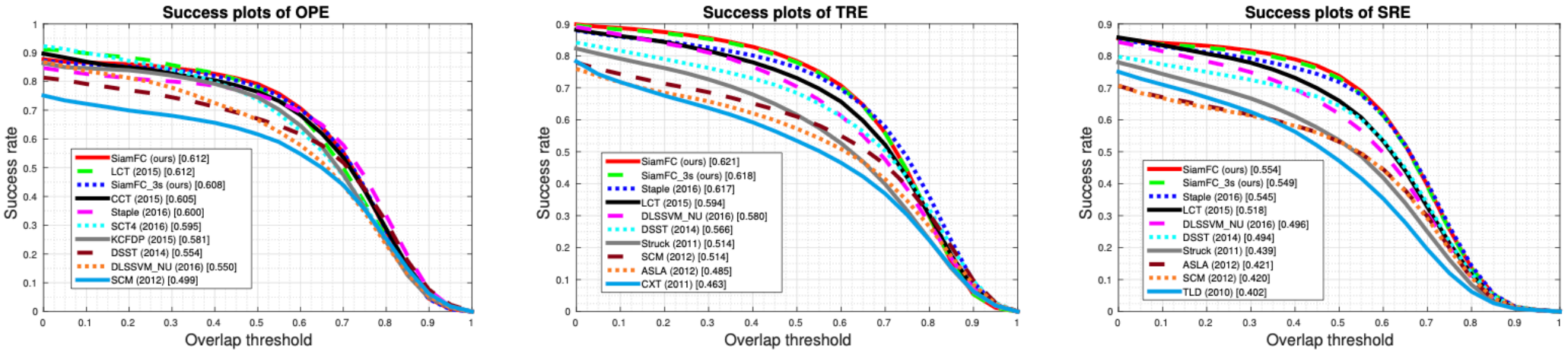


Fig. 3: Success plots for OPE (one pass evaluation), TRE (temporal robustness evaluation) and SRE (spatial robustness evaluation) of the OTB-13 [11] benchmark. The results of CCT, SCT4 and KCFDP were only available for OPE at the time of writing.

VOT-15 Results

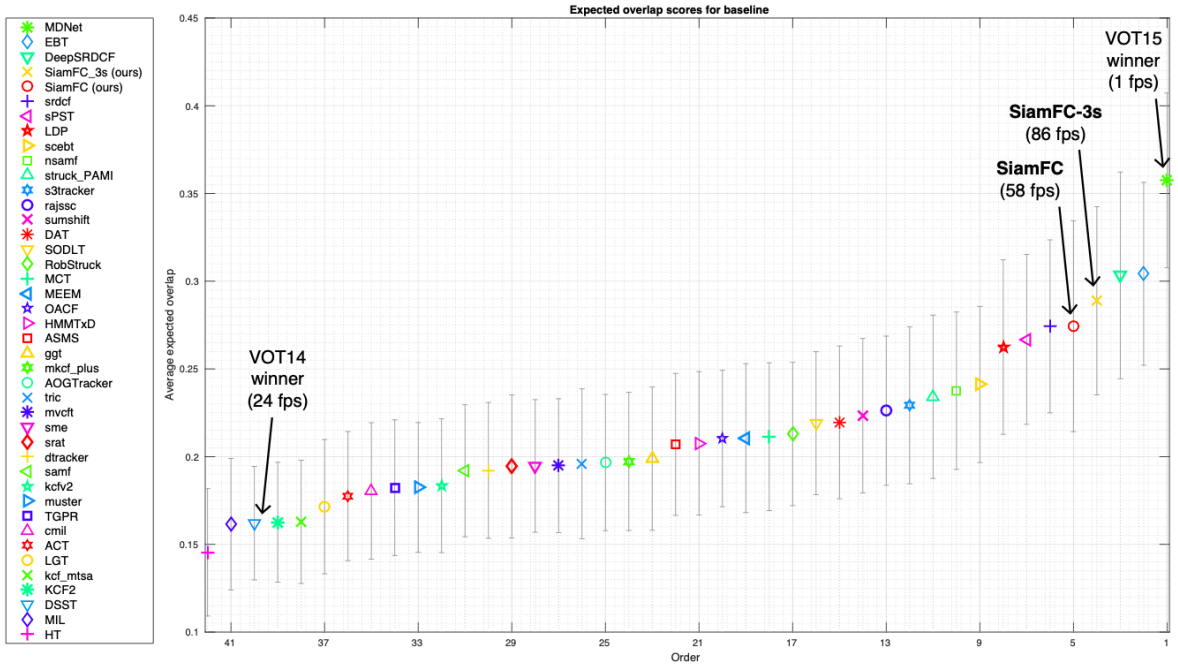


Fig. 5: VOT-15 ranking in terms of expected average overlap. Only the best 40 results have been reported.

< The final ranking in terms of expected average overlap >

Table 2: Raw scores, overlap and reported speed for our proposed method and the best 15 performing trackers of the VOT-15 challenge. Where available, we compare with the speed reported by the authors, otherwise (*) we report the values from the VOT-15 results [12] in EFO units, which roughly correspond to fps (e.g. the speed of the NCC tracker is 140 fps and 160 EFO).

Tracker	accuracy	# failures	overlap	speed (fps)
MDNet [9]	0.5620	46	0.3575	1
EBT [41]	0.4481	49	0.3042	5
DeepSRDCF [6]	0.5350	60	0.3033	< 1 *
SiamFC-3s (ours)	0.5335	84	0.2889	**86**
SiamFC (ours)	0.5240	87	0.2743	58
SRDCF [42]	0.5260	71	0.2743	5
sPST [43]	0.5230	85	0.2668	2
LDP [12]	0.4688	78	0.2625	4 *
SC-EBT [44]	0.5171	103	0.2412	–
NSAMF [45]	0.5027	87	0.2376	5 *
StruckMK [3]	0.4442	90	0.2341	2
S3Tracker [46]	0.5031	100	0.2292	14 *
RAJSSC [12]	0.5301	105	0.2262	2 *
SumShift [46]	0.4888	97	0.2233	17 *
DAT [47]	0.4705	113	0.2195	15
SO-DLT [7]	0.5233	108	0.2190	5

< Scores and speed of the 15 highest ranked trackers >

Snapshots

(from VOT-15)

Drastic change of appearance

motion blur

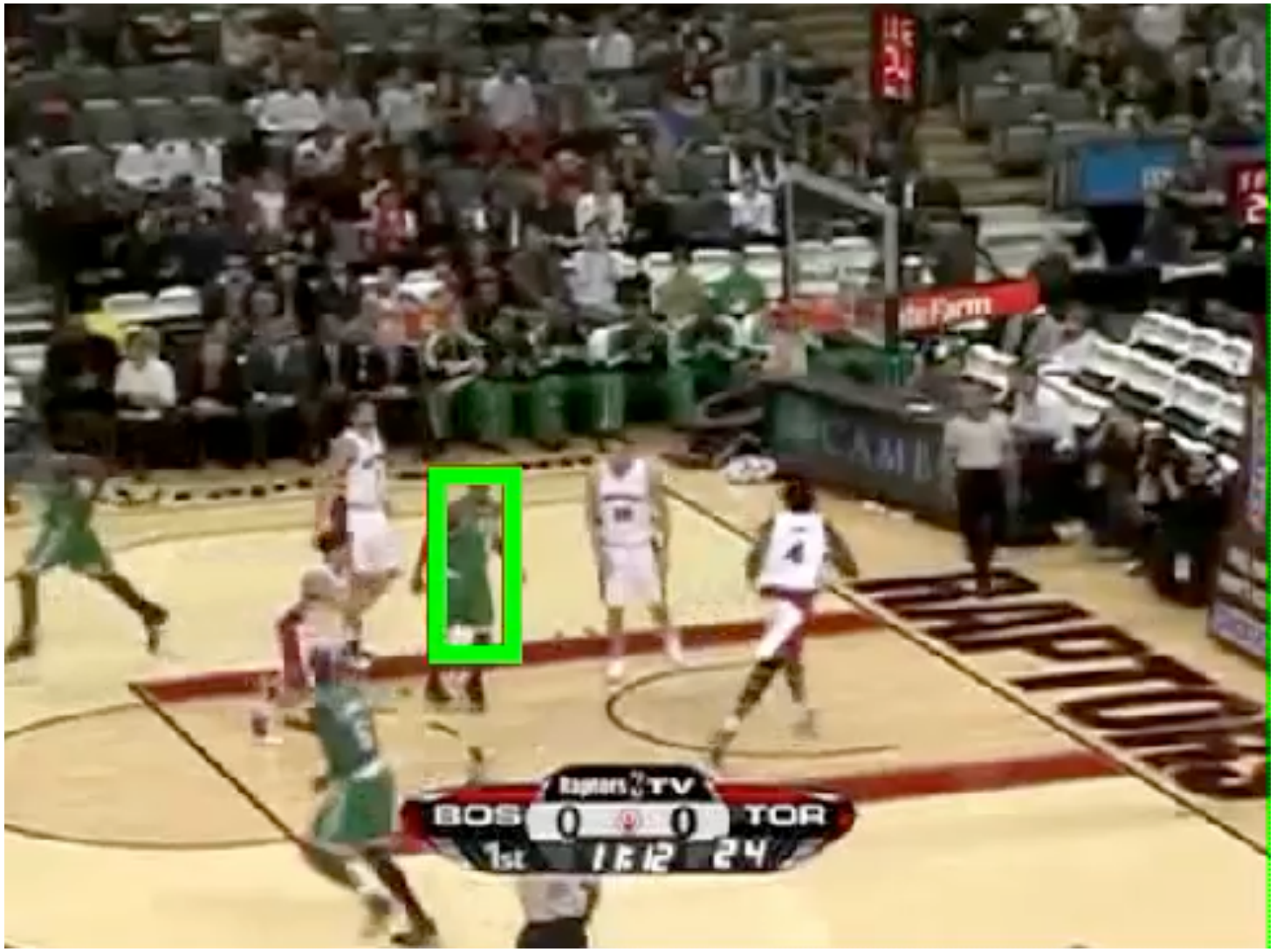
Drastic change of appearance

Drastic change of appearance

Scenes with confusion

Poor illumination & Scale change

Visualization with OTB-15: Basketball



Ground truth

SiamFC tracker