Fully-Convolutional Siamese Networks for Object Tracking

Luca Bertinetto Jack Valmadre Joao F.Henriques
Andrea Vedaldi Philip H. S. Torr
Department of Engineering Science, University of Oxford

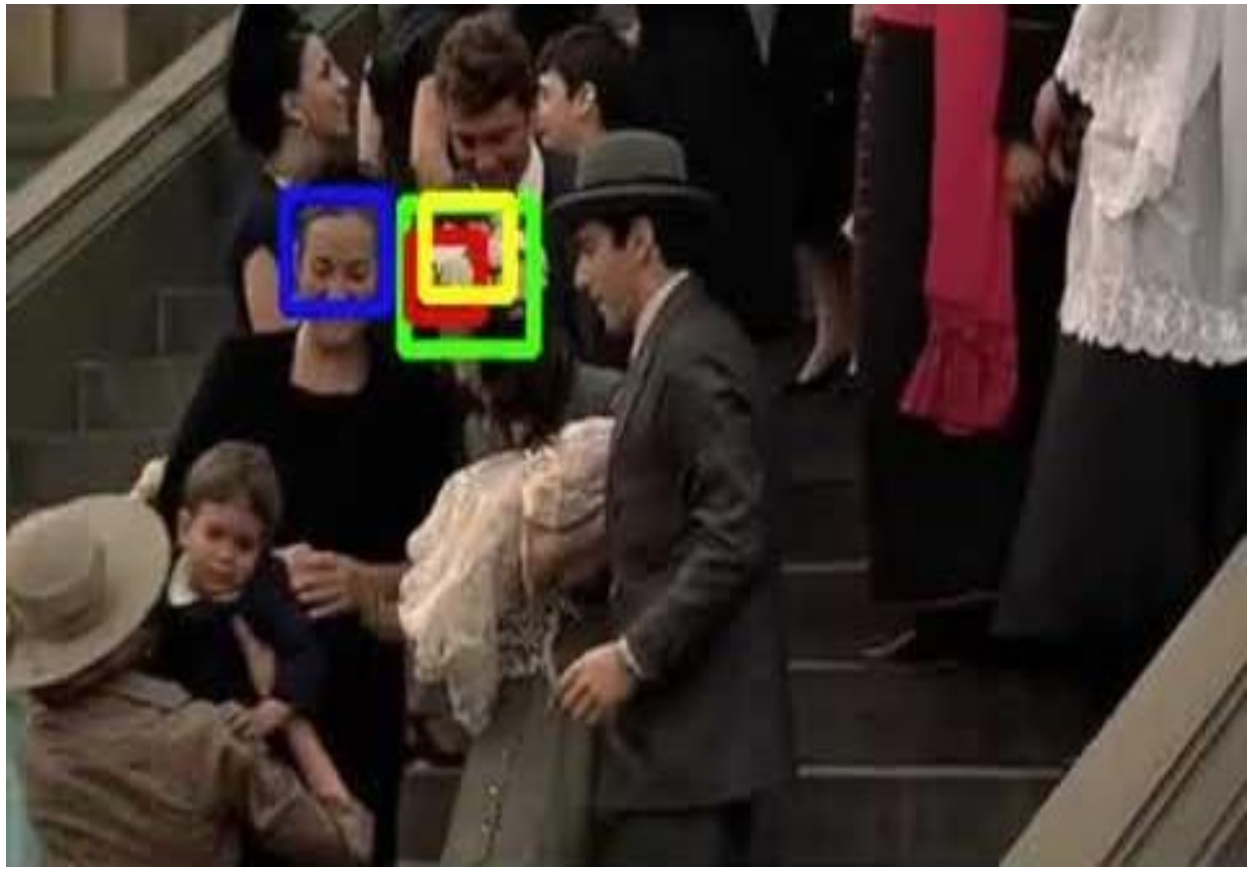
swkim@dongguk.edu
sunwoo KIm

Index

dongguk UNIVERSITY

Introduction - What is the Visual Object Tracking?

- **Localizing the target in the video**
- **Given arbitrary target**
- **Class-agnostic**
- **Hard Negatives**



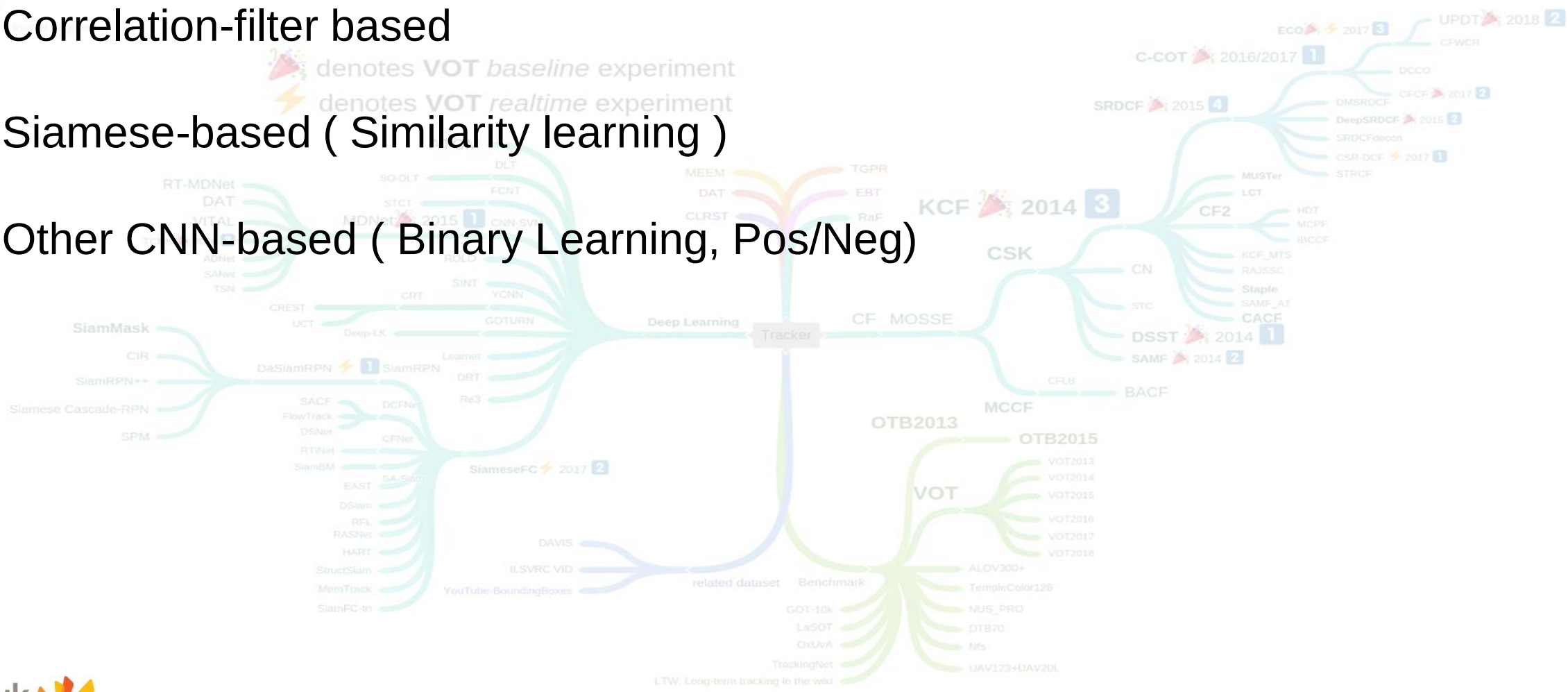
<SiamVGG>

Introduction - How to track the object?

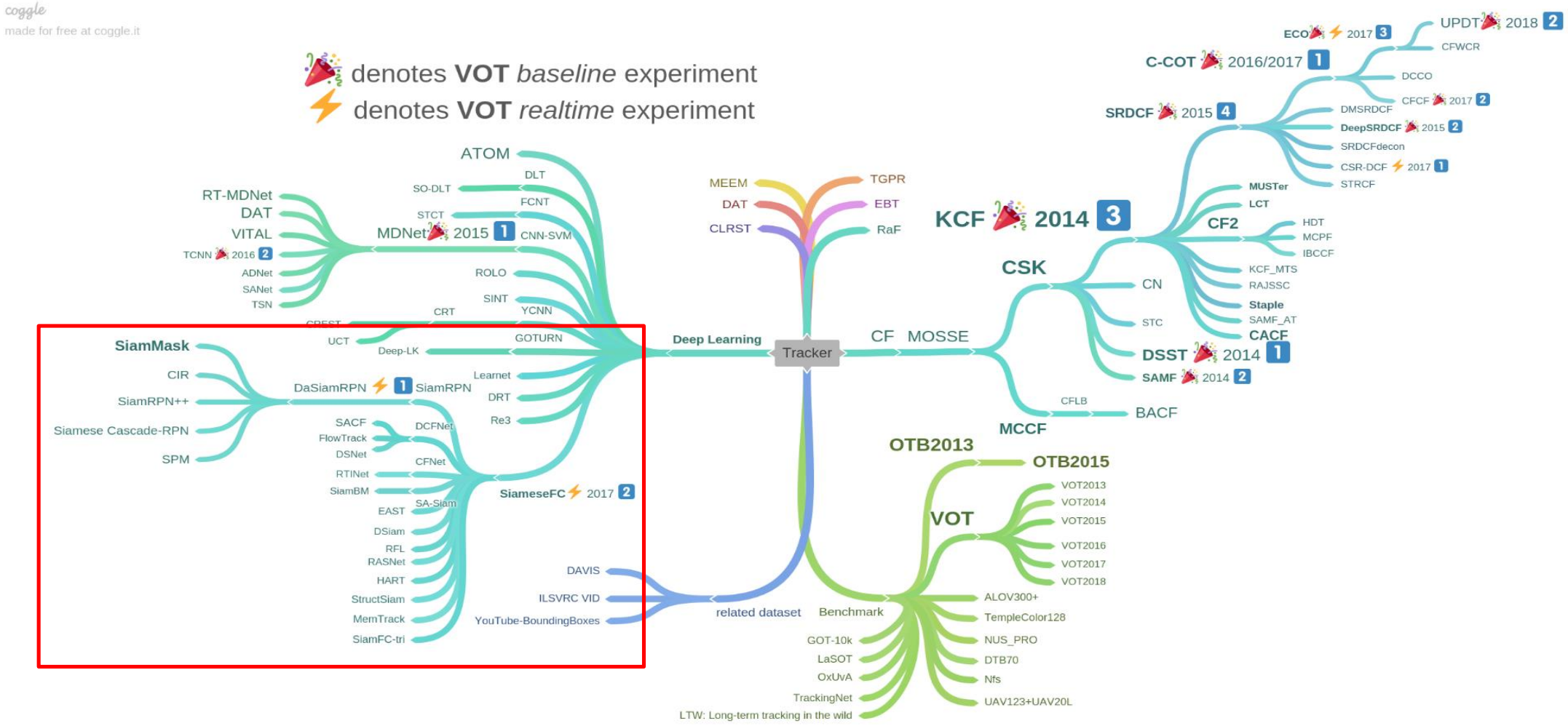
- Find the most ***similar*** patch in ***T*** frame based on ***T-1*** frame target
  - it needs robustness to object deformations.

Introduction - How to define similarity?

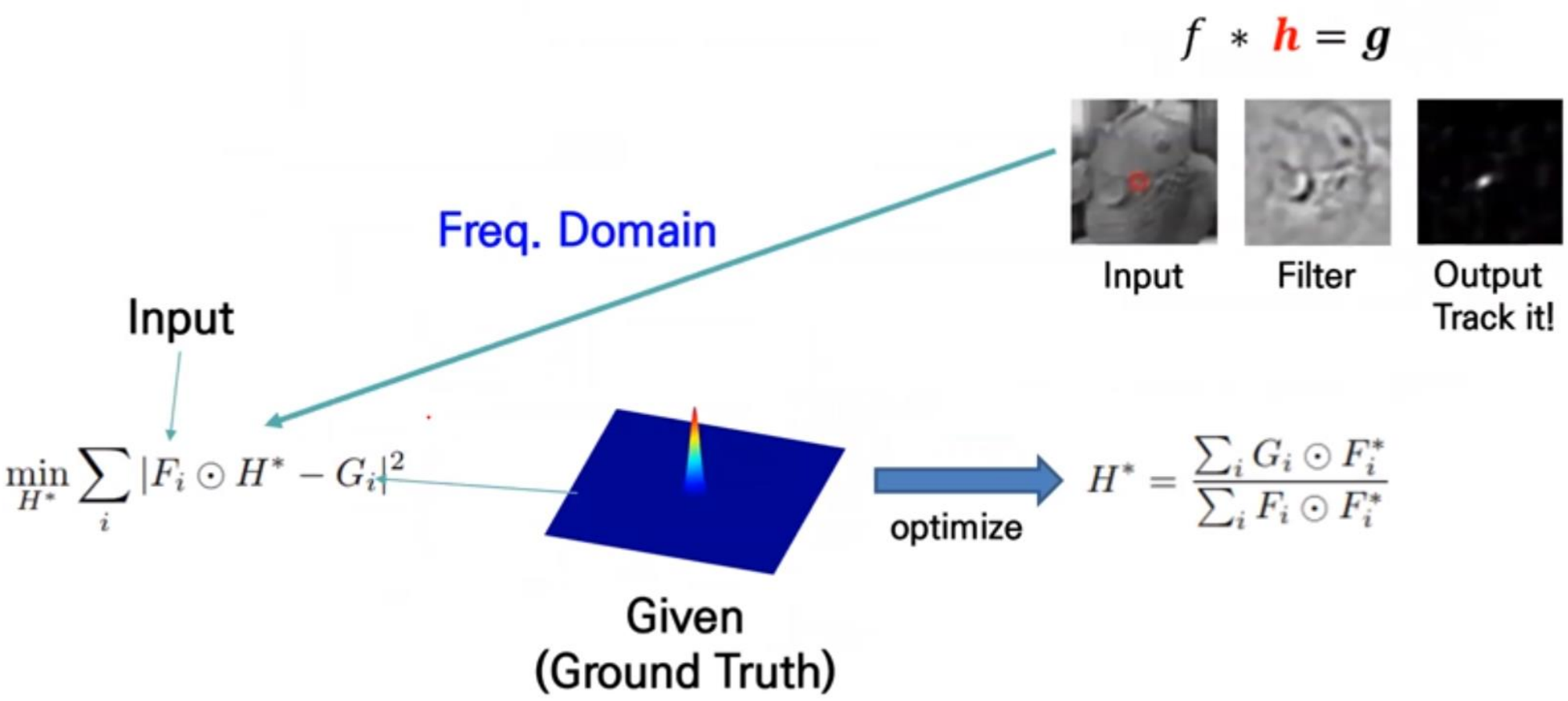
- Correlation-filter based
- Siamese-based (Similarity learning)
- Other CNN-based (Binary Learning, Pos/Neg)



Introduction - VOT roadmap



Introduction - Correlation-filter based



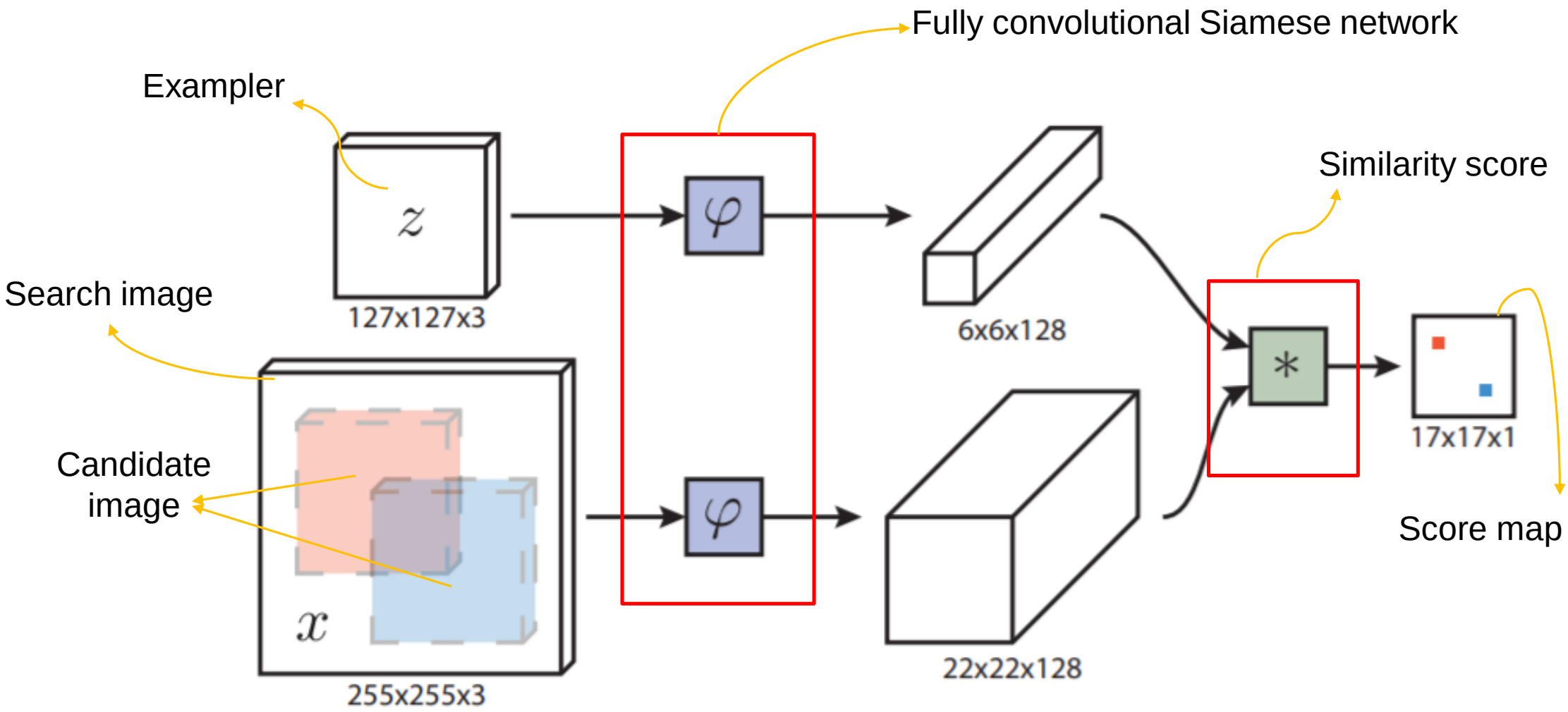
Introduction - Correlation-filter based

- First step
  - Find h(filter) from i(input image) & g(given output)

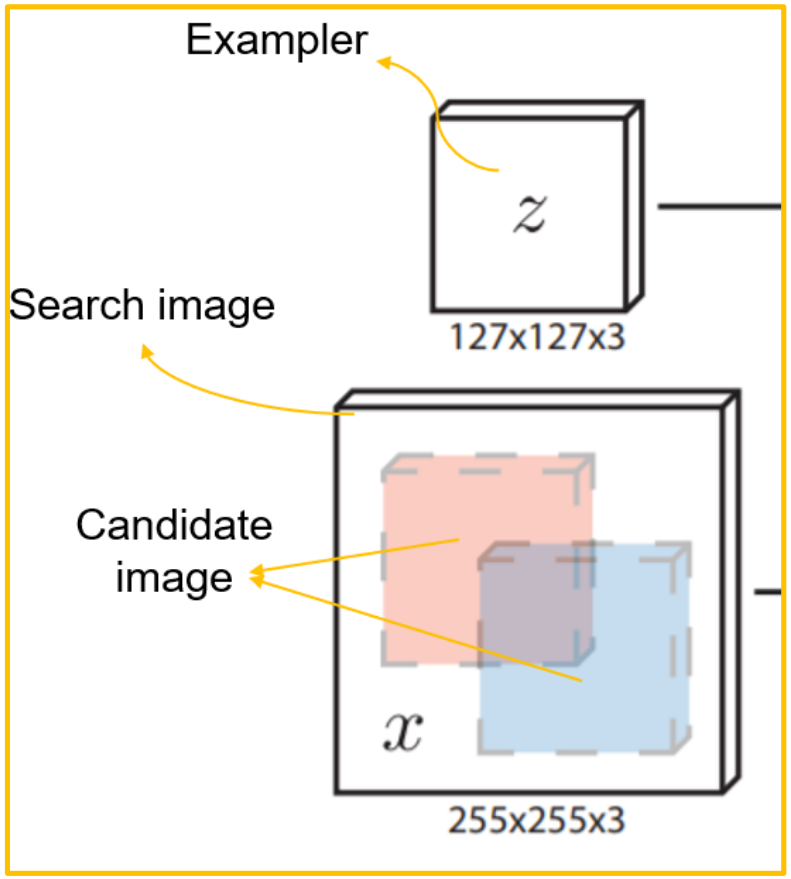
- N step
  - Filtering input image & find output which has the highest response
  - Update filter
  - Iteration…

$$H_i^* = \eta \frac{G_i \odot F_i^*}{F_i \odot F_i^*} + (1-\eta) H_{i-1}^*$$

Model Architecture



Model Architecture - Input



Two inputs

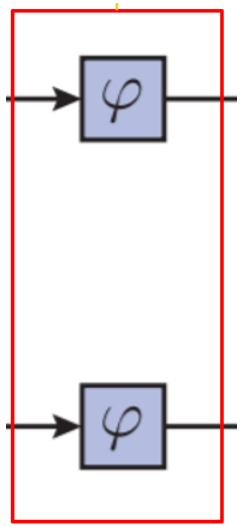
- **Exampler (z)**
  -Randomly choiced in labled objects ***at first frame***
  (It is not important to detect first location of the object. Only tracking the object is the main task.)

  -size : 127x127x3 (fixed in training)

- **Search image (x)**
  -It is extracted with center of the tracked image of T-1 frame. Make 255x255 patch with center of that point.

  -size : 255x255x3

Model Architecture – Siamese Net



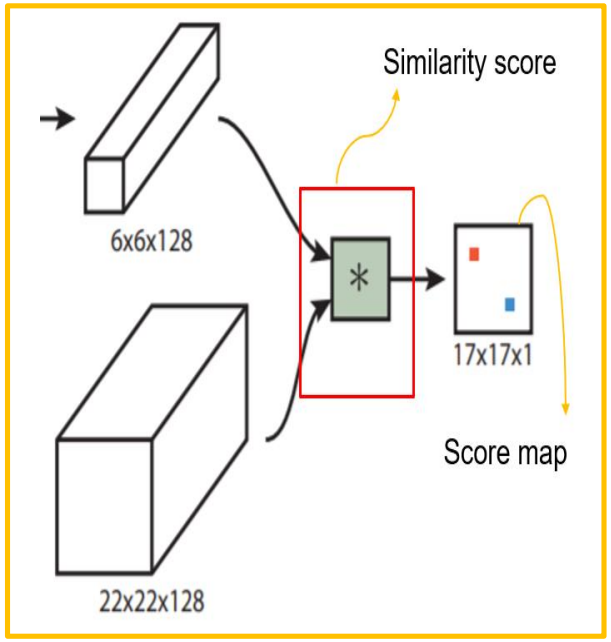
Layer	Support	Chan. map	Stride	Activation size for exemplar	Activation size for search	chans.
				127×127	255×255	$\times 3$
conv1	11×11	96×3	2	59×59	123×123	$\times 96$
pool1	3×3		2	29×29	61×61	$\times 96$
conv2	5×5	256×48	1	25×25	57×57	$\times 256$
pool2	3×3		2	12×12	28×28	$\times 256$
conv3	3×3	384×256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

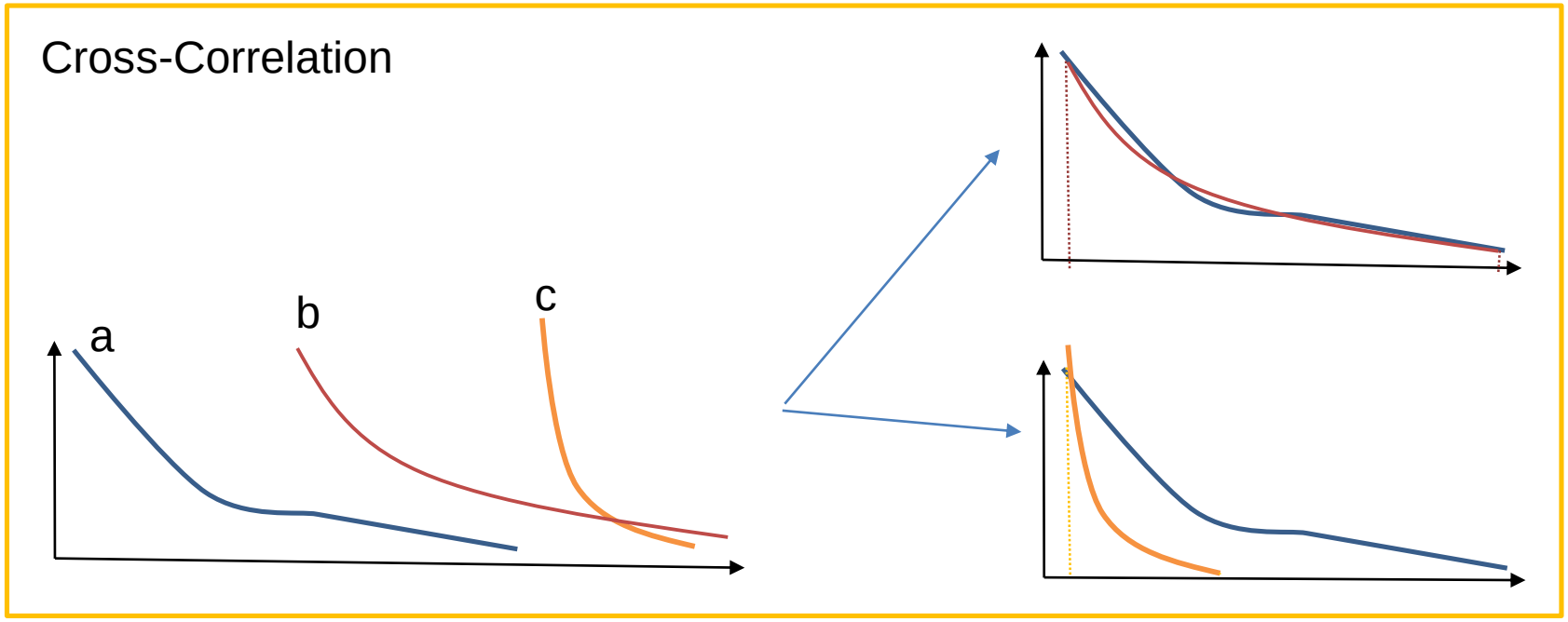
Model Architecture – Get Similarity score & score map

Similarity score : $f(z, x) = \varphi(z) * \varphi(x) + b\,\mathbb{1},$

Score map : similarity map, the highest score is the next image to be tracked

z : Exampler
x : Search image
*** : Cross-Correlation**
φ : Embeeding (Siamese Net)
$b\,\mathbb{1}$: b x indicator function

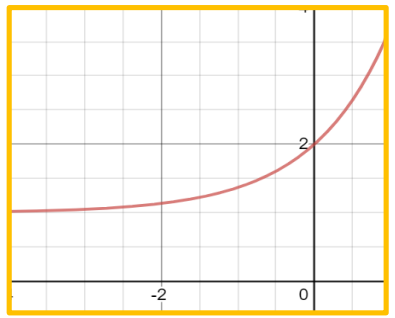




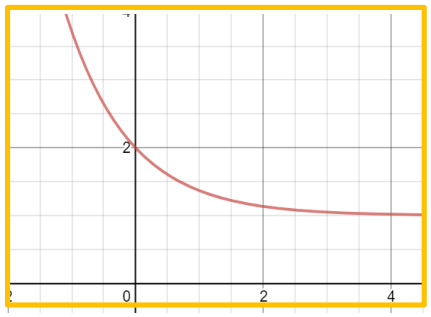
Model Architecture – Caculating loss

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

y :label (-1 or 1)
v : similarity



< y = -1 >

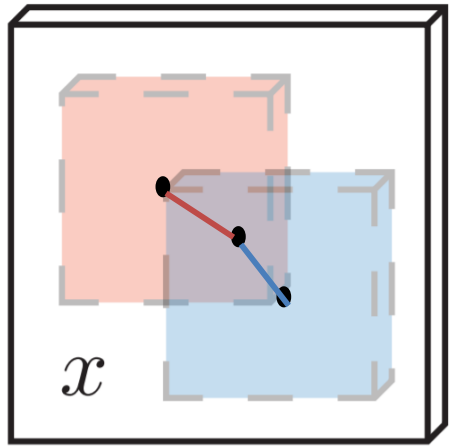


< y = 1 >

< Logistic loss >

Determine pos/neg : $y[u] = \begin{cases} +1 & \text{if } k\|u - c\| \leq R \\ -1 & \text{otherwise} \end{cases}$.

k : strides in calculating score map
u : index of center of candidate image
c : center of search image
R : threshold



< |u-c| : uclidean distance >

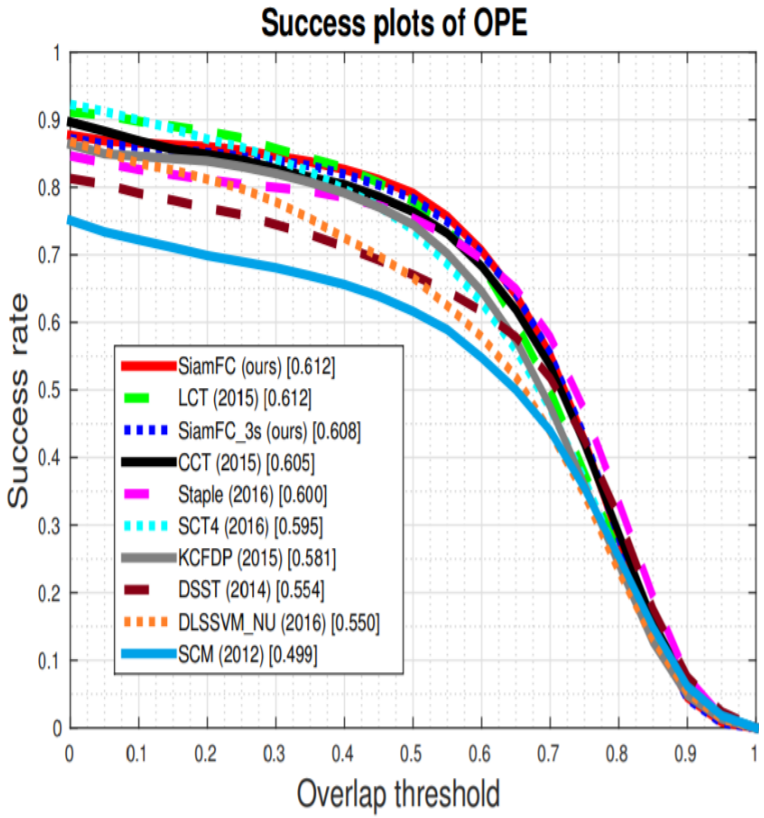
Model Architecture – Caculating loss

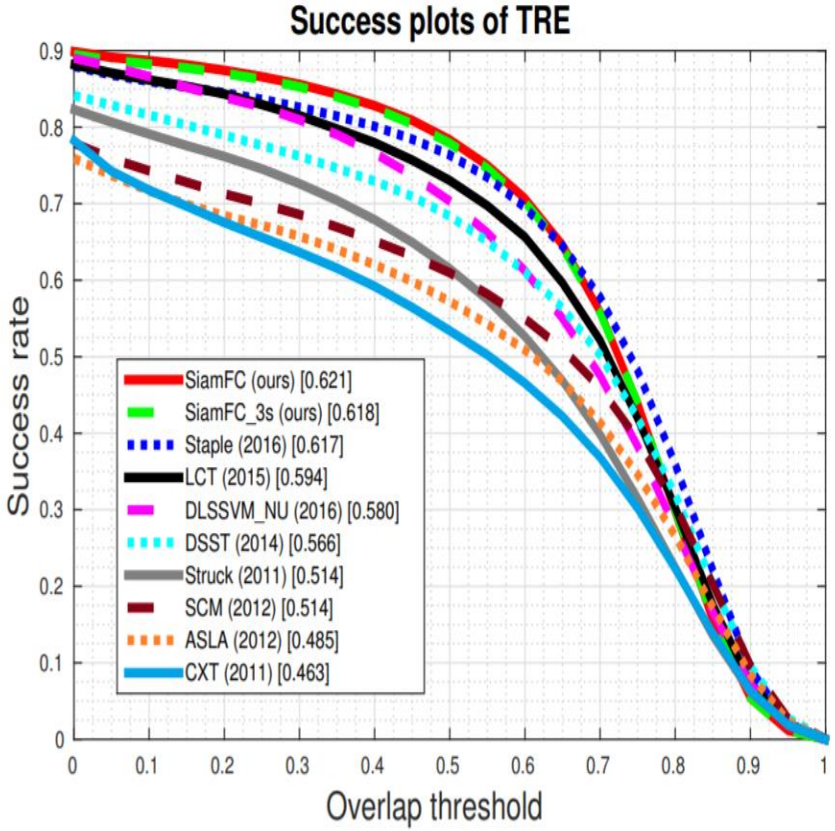
Loss : $L(y, v) = \frac{1}{|\mathcal{D}|} \sum_{u \in \mathcal{D}} \ell(y[u], v[u])$

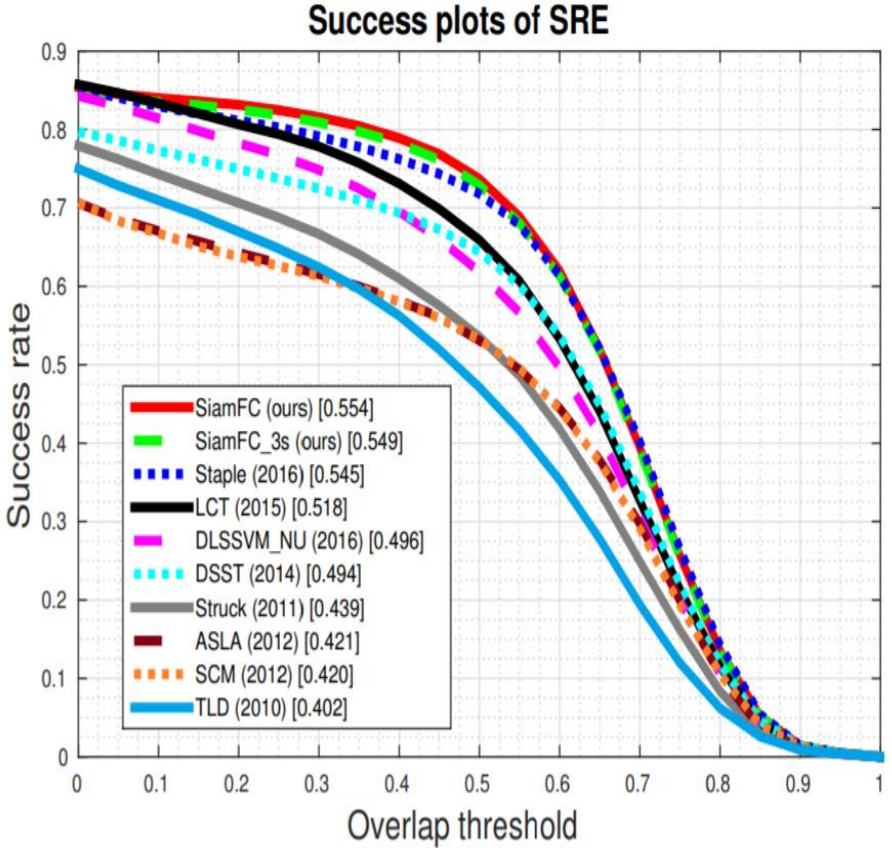
Learning to : $\arg\min_{\theta} \mathop{\mathbb{E}}_{(z,x,y)} L(y, f(z, x; \theta))$

Experiment

- OPE (One Path Evaluation) : Evaluate one tracker on the entire sequence with initialization from the ground truth position in the first frame
- TRE (Temporal robustness evaluation) : Change the start at different frames of the video and then evaluate
- SRE (Spatial robustness evaluation) : Sample the initial bonding box in the first frame by shifting or scaling the ground truth

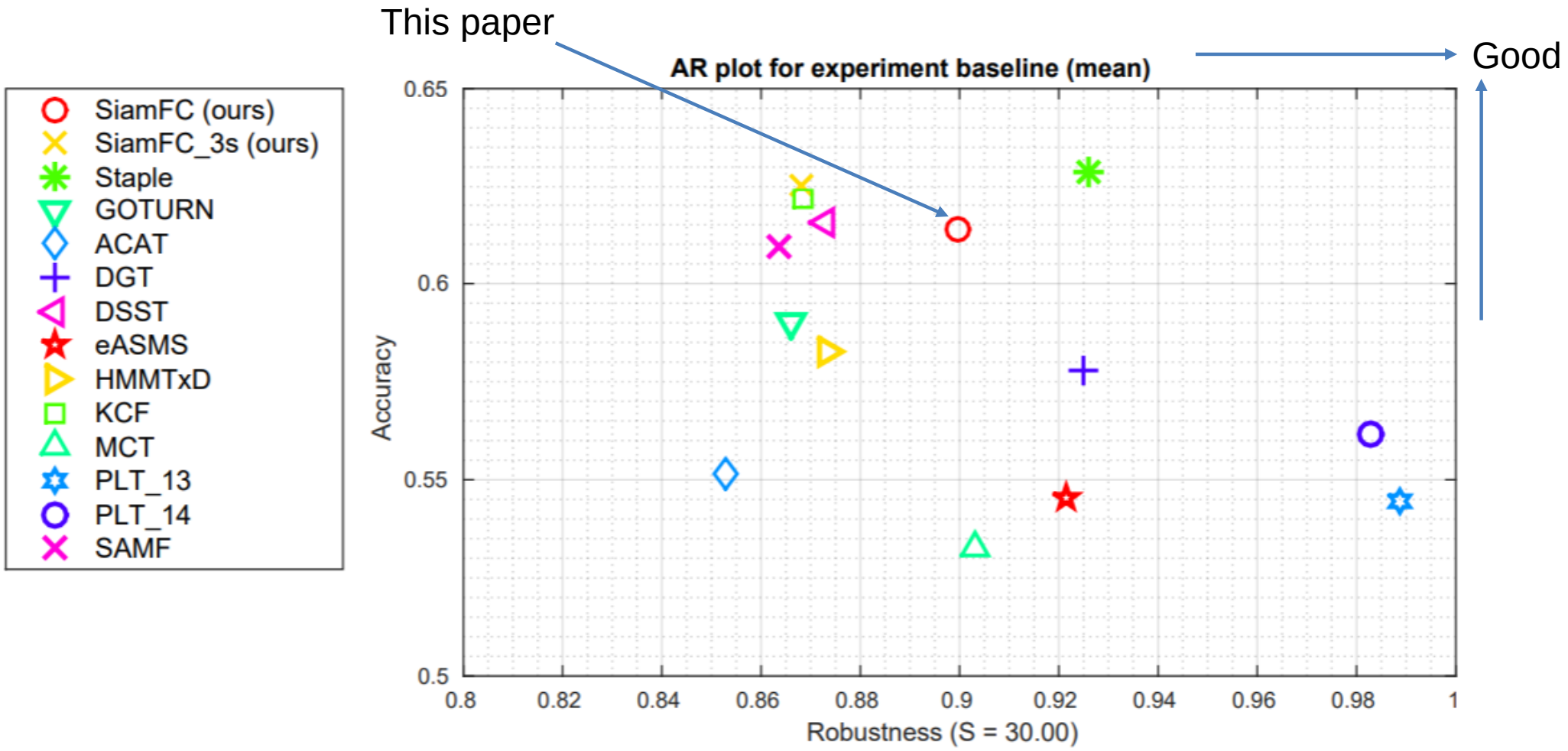






dongguk UNIVERSITY

What is Biometry?



dongguk UNIVERSITY

Thank you!