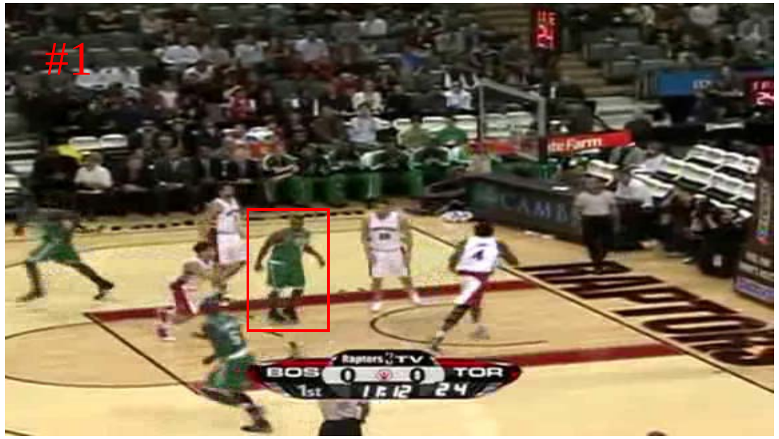
Recent Advances in Visual Tracking

吴毅

Nanjing Audit University

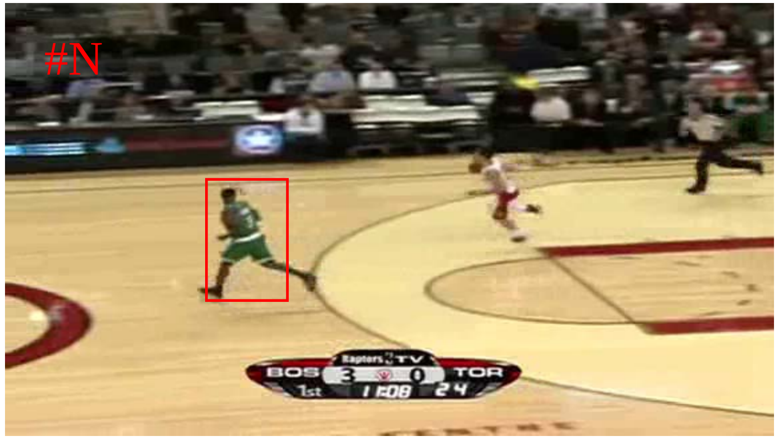
南京审计大学
NANJING AUDIT UNIVERSITY

Visual Tracking in Computer Vision



Initialization in the 1st frame

....



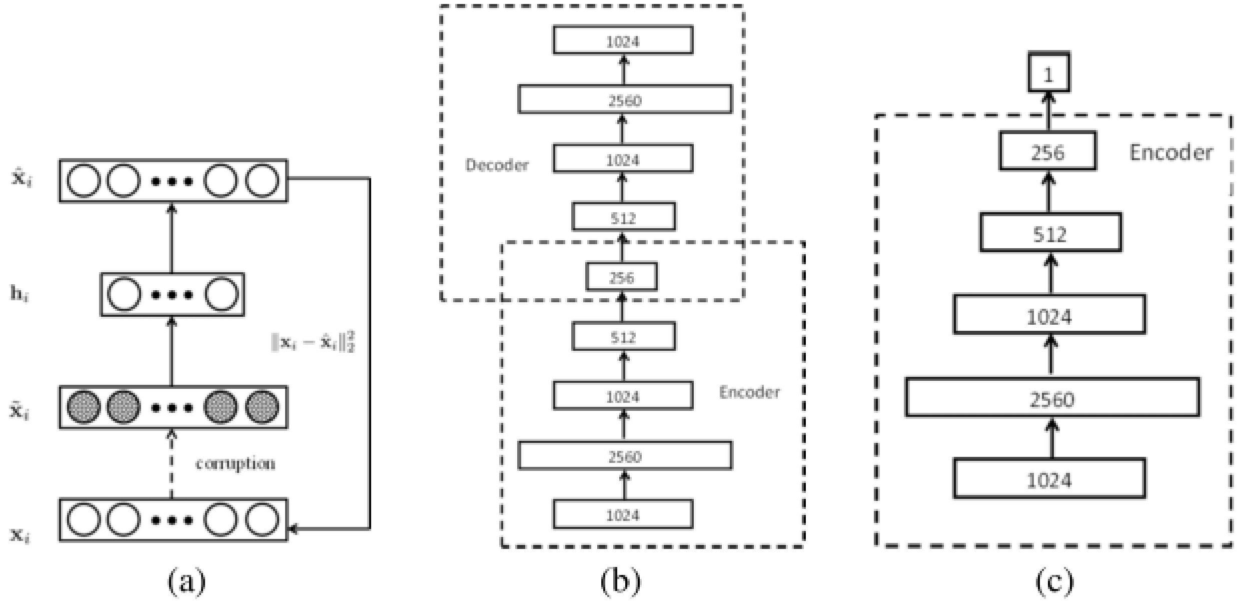
Estimated states in the N-th frame

- A fundamental problem in computer vision
- A challenging and difficult task

Outline

- Tracking based on deep learning
- Tracking based on correlation filters
- Other interesting work

Learning a Deep Compact Image Representation for Visual Tracking

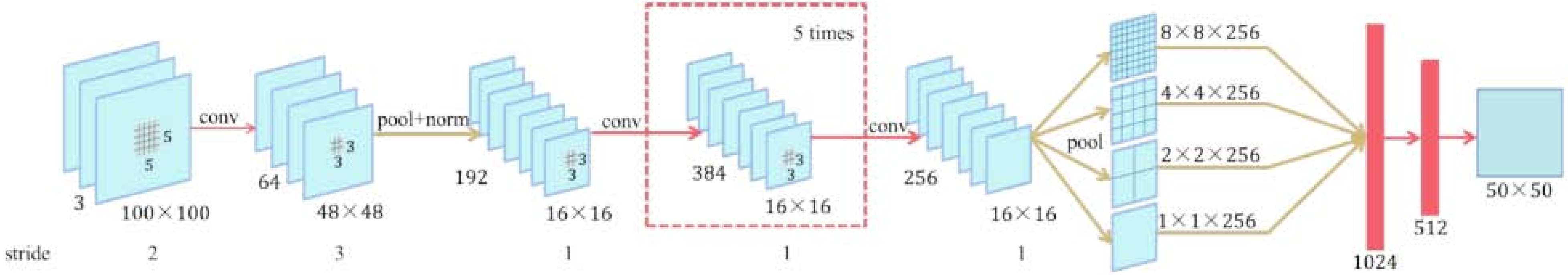


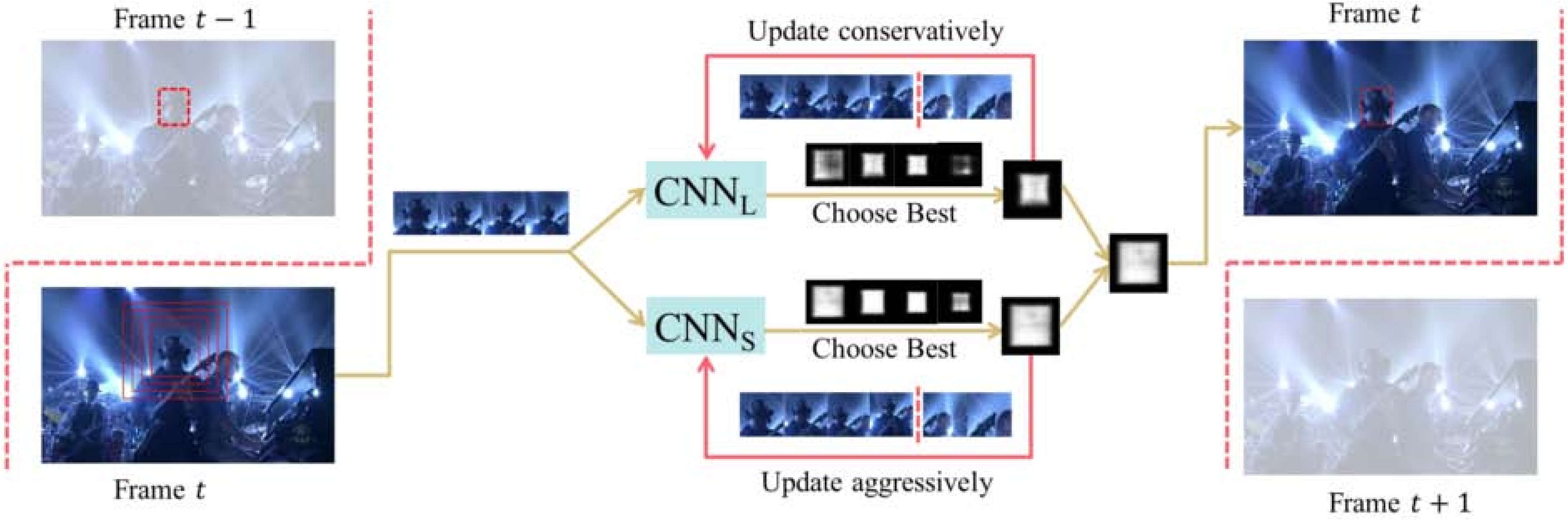
Stacked denoising autoencoder (SDAE)

Some filters in the first layer of the learned SDAE

Naiyan Wang and Dit-Yan Yeung, “Learning a Deep Compact Image Representation for Visual Tracking,” in *NIPS*, 2013.

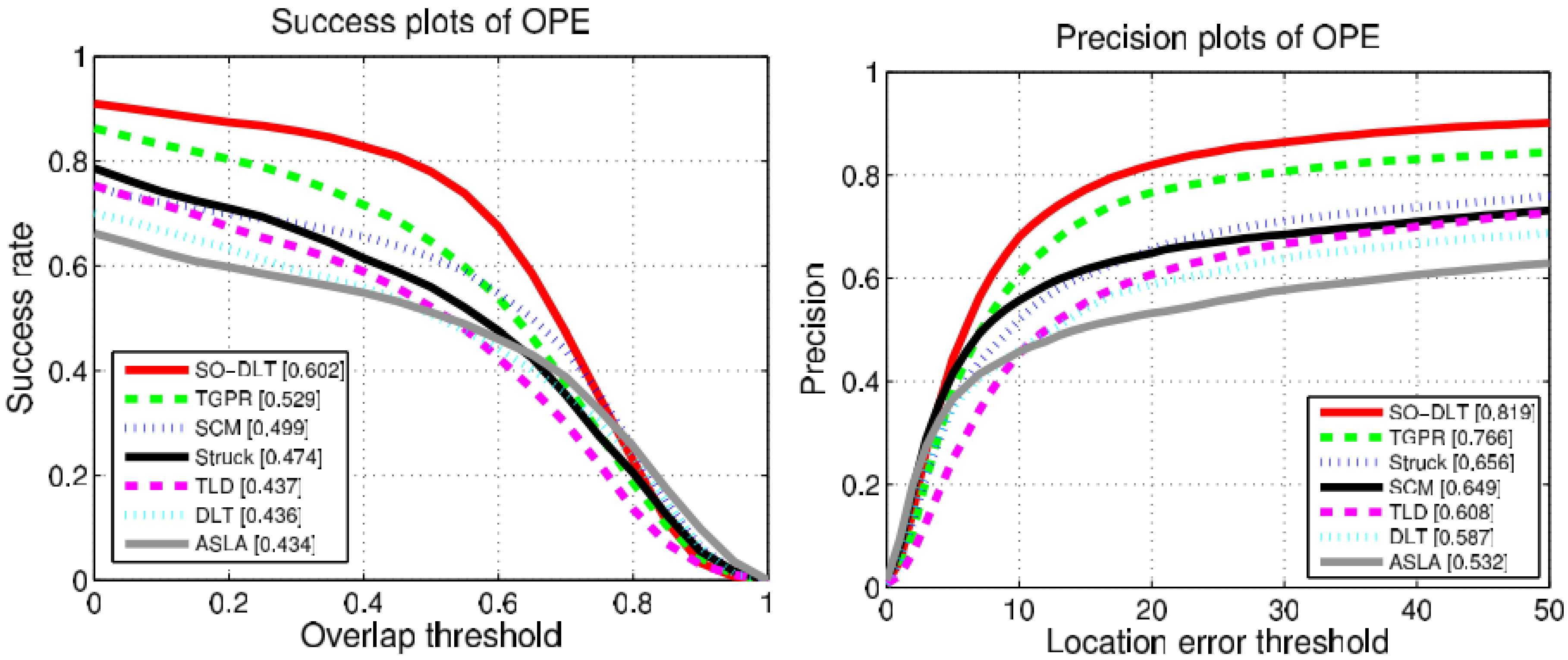
Transferring Rich Feature Hierarchies for Robust Visual Tracking





N. Wang, S. Li, A. Gupta, and D.-Y. Yeung, "Transferring Rich Feature Hierarchies for Robust Visual Tracking," *arXiv*, 2015.

Transferring Rich Feature Hierarchies for Robust Visual Tracking



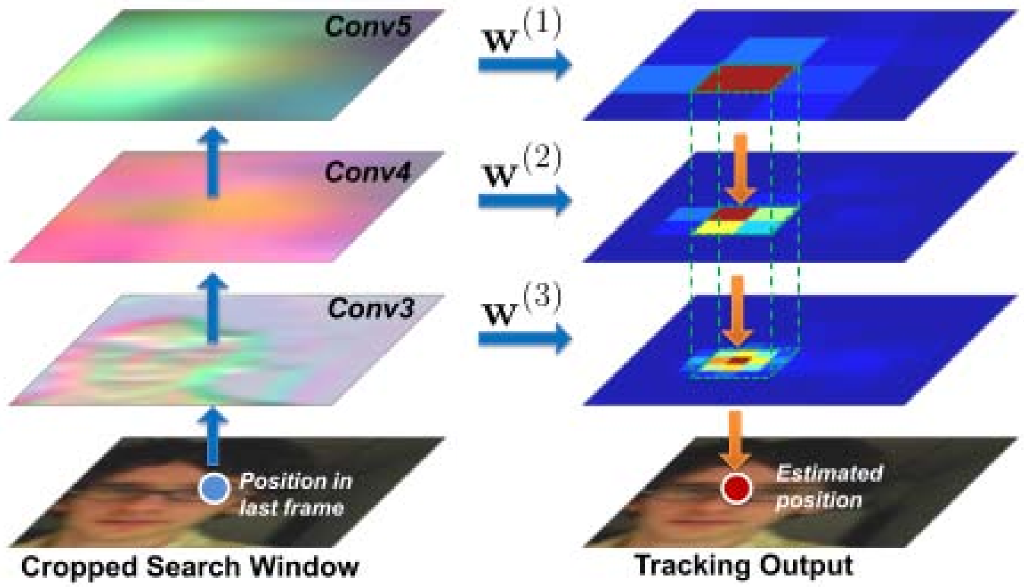
Results on OTB50

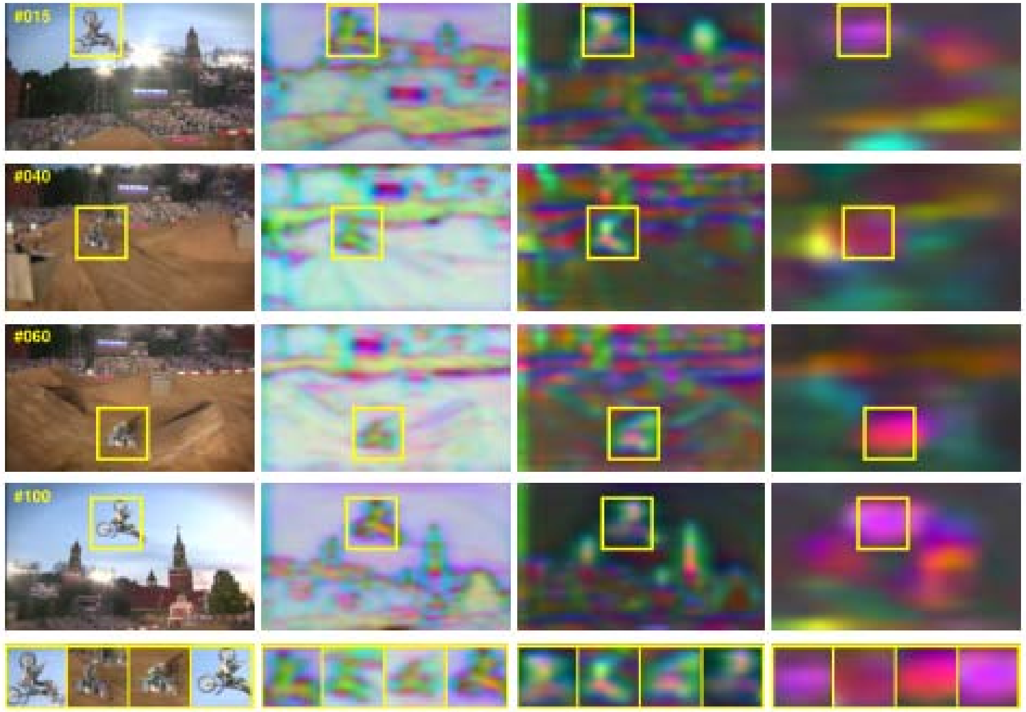
OTB50: Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, “Online Object Tracking: A Benchmark,” in *CVPR*, 2013.

Tracking based on deep learning in 2015

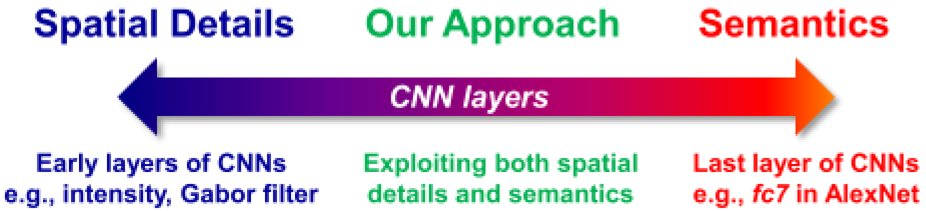
- C. Ma, J. Huang, X. Yang, and M. Yang, “Hierarchical Convolutional Features for Visual Tracking,” in *ICCV*, 2015.
- L. Wang, W. Ouyang, X. Wang, and H. Lu, “Visual Tracking with Fully Convolutional Networks,” in *ICCV*, 2015.
- S. Hong, T. You, S. Kwak, and B. Han, “Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network,” in *ICML*, 2015.
- H. Li, Y. Li, and F. Porikli, “DeepTrack: Learning Discriminative Feature Representations by Convolutional Neural Networks for Visual Tracking,” in *BMVC*, 2015.

Hierarchical Convolutional Features for Visual Tracking





(a) Input (b) *conv3-4* (c) *conv4-4* (d) *conv5-4*



Spatial Details Our Approach Semantics

CNN layers

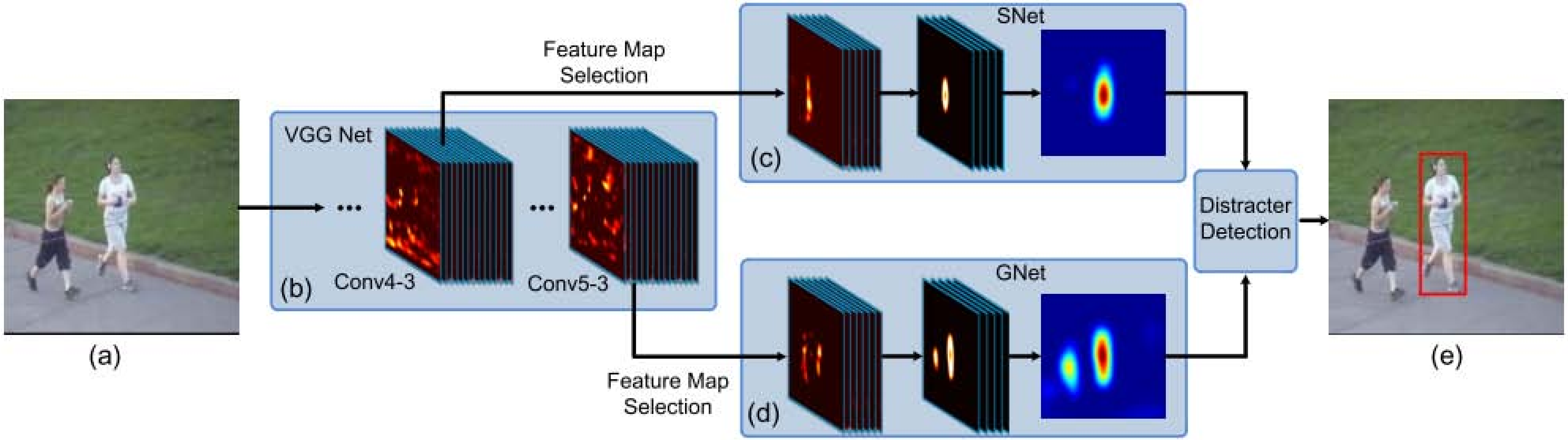
Early layers of CNNs e.g., intensity, Gabor filter

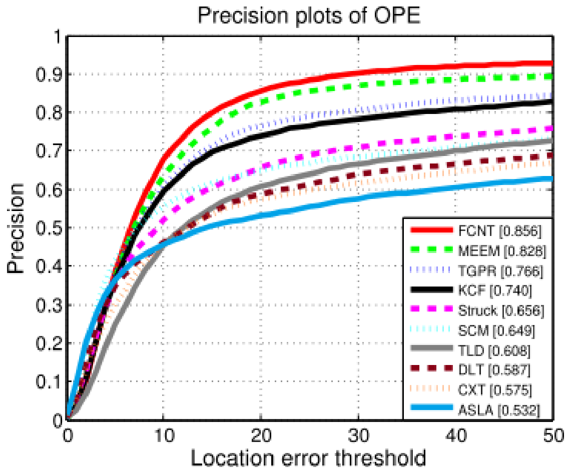
Exploiting both spatial details and semantics

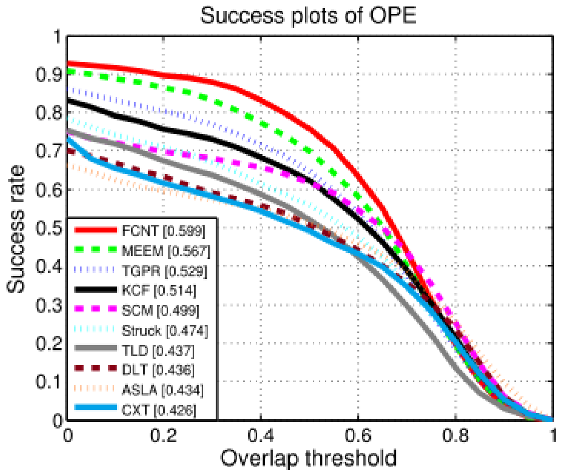
Last layer of CNNs e.g., *fc7* in AlexNet

C. Ma, J. Huang, X. Yang, and M. Yang, “Hierarchical Convolutional Features for Visual Tracking,” in *ICCV*, 2015.

Visual Tracking with Fully Convolutional Networks



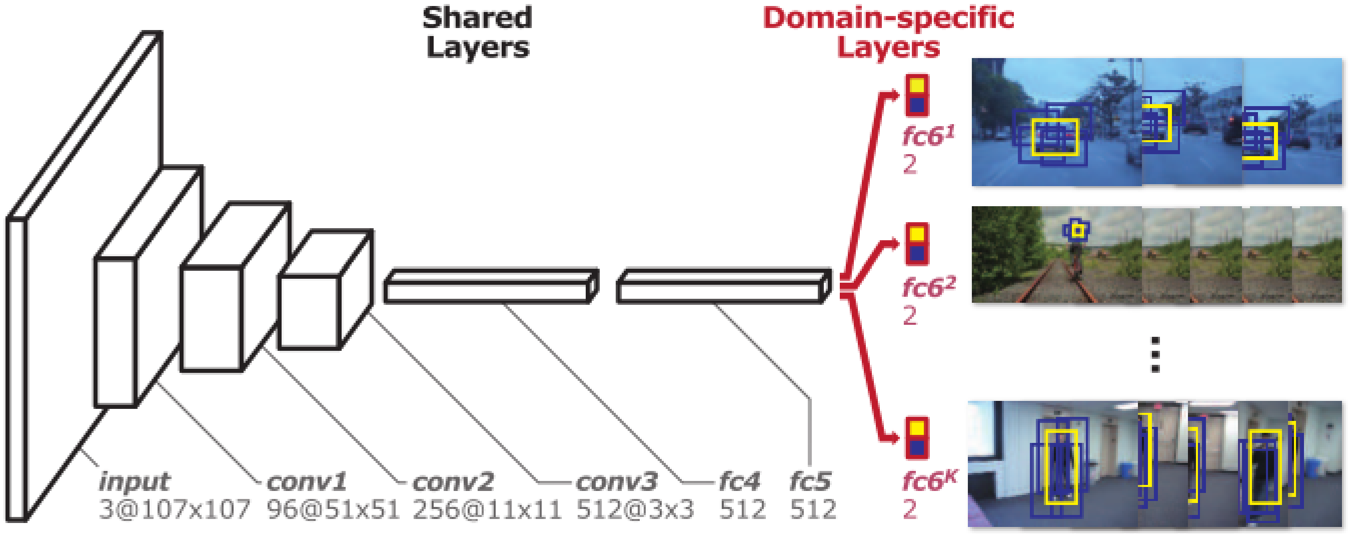


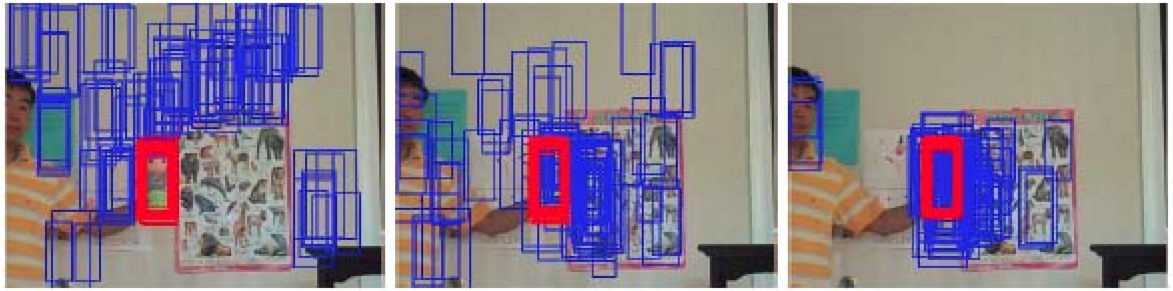


L. Wang, W. Ouyang, X. Wang, and H. Lu, “Visual Tracking with Fully Convolutional Networks,” in *ICCV*, 2015.

Learning Multi-Domain Convolutional Neural Networks for Visual Tracking

- VGG-M network
  - Five shared layers
  - K branches of domain-specific layers
- Hard minibatch mining
- Bounding box regression
  - Trained only at the 1st frame
- N(= 256) target candidates
- 1 FPS using NVIDIA Tesla K20m GPU

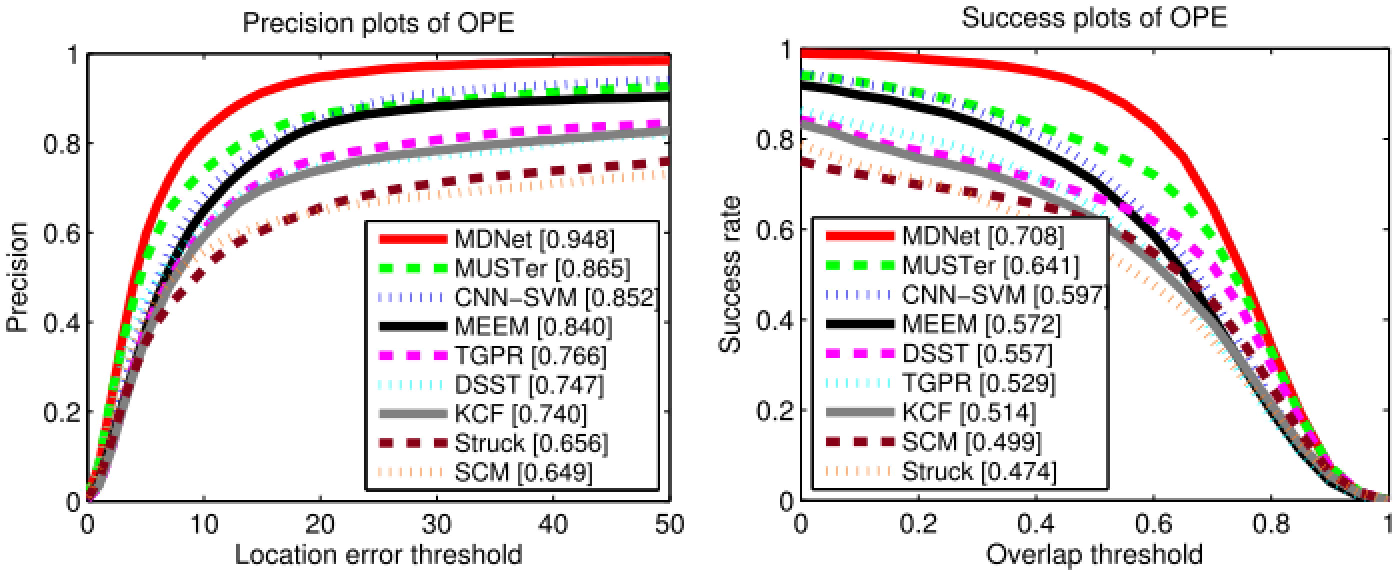




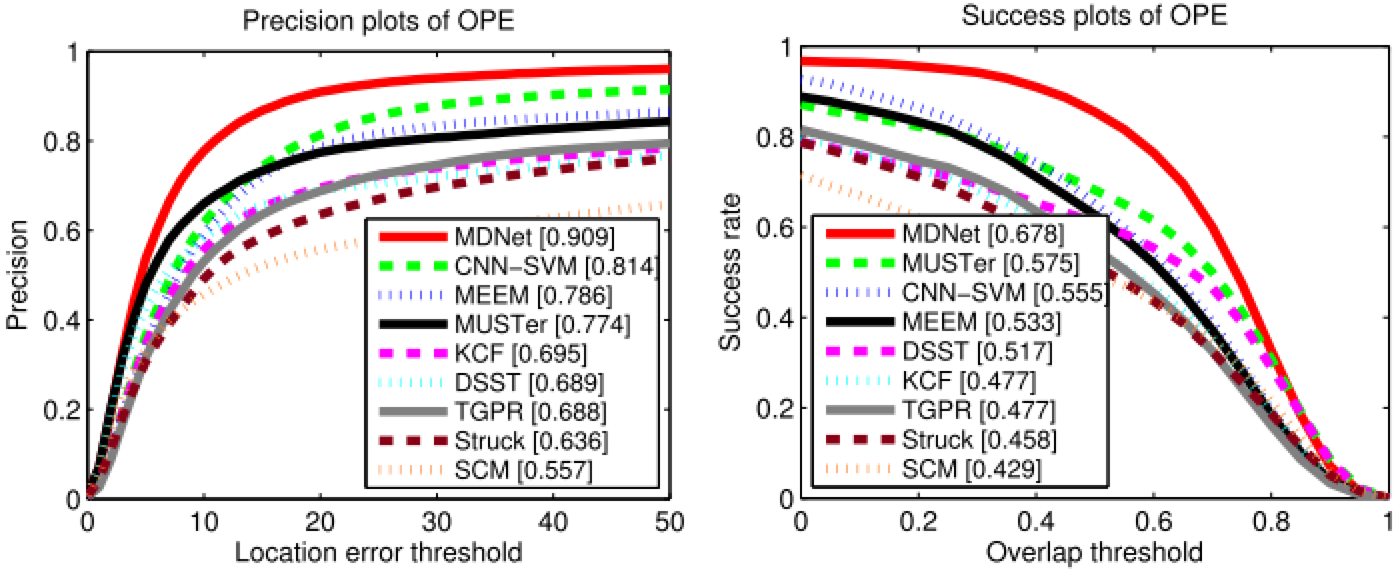
(a) 1st minibatch (b) 5th minibatch (c) 30th minibatch

http://cvlab.postech.ac.kr/research/mdnet/

H. Nam and B. Han, "Learning Multi-Domain Convolutional Neural Networks for Visual Tracking," *CVPR*, 2016.



(a) OTB50 result



(b) OTB100 result

OTB50: Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, “Online Object Tracking: A Benchmark,” in *CVPR*, 2013.

OTB100: Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, “Object Tracking Benchmark,” *PAMI*, 37:9, pp. 1834–1848, 2015.

SANet: Structure-Aware Network for Visual Tracking

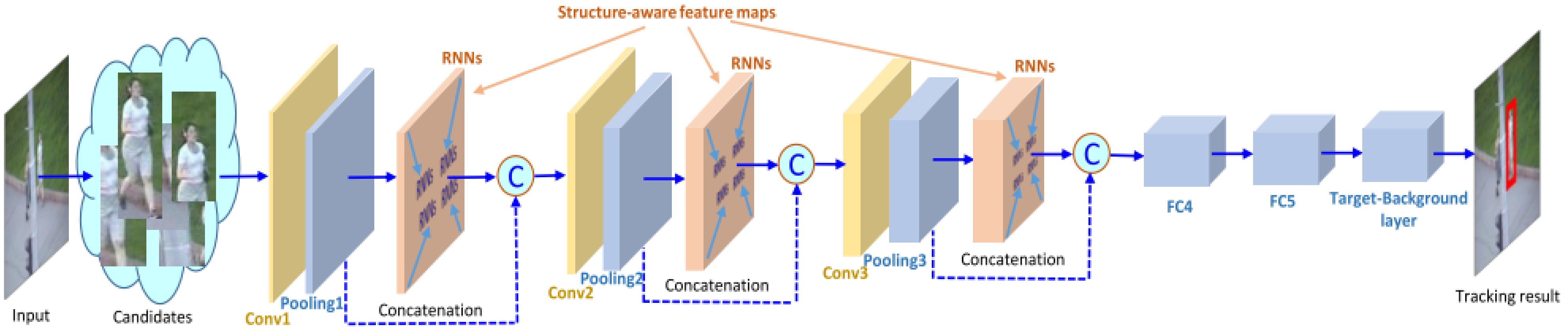


Figure 2. Illustration of the proposed SANet for visual tracking.

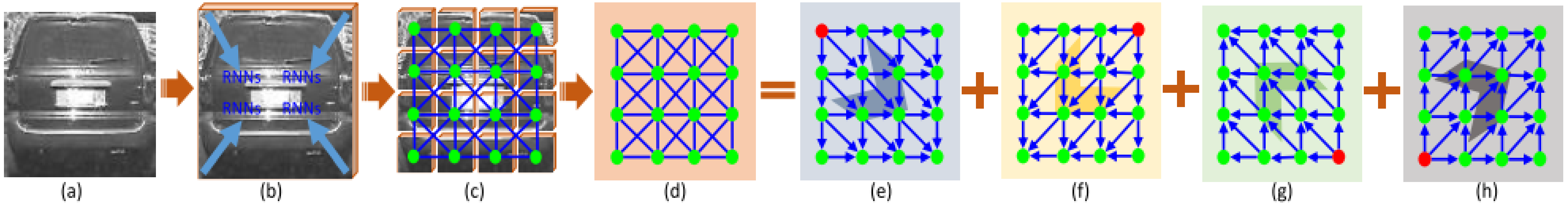
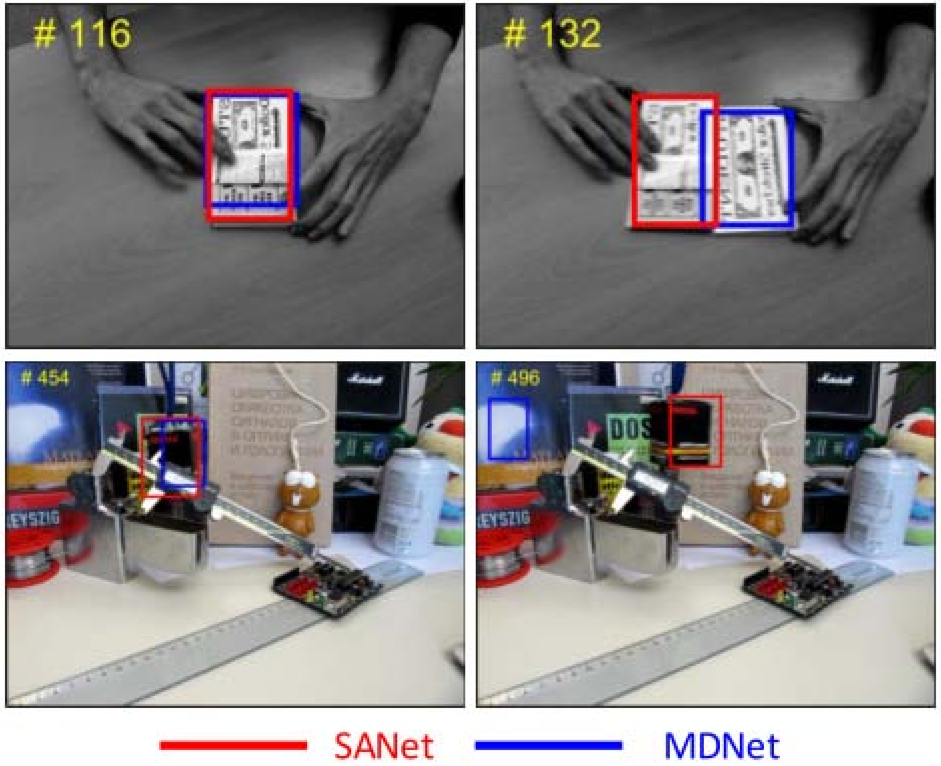
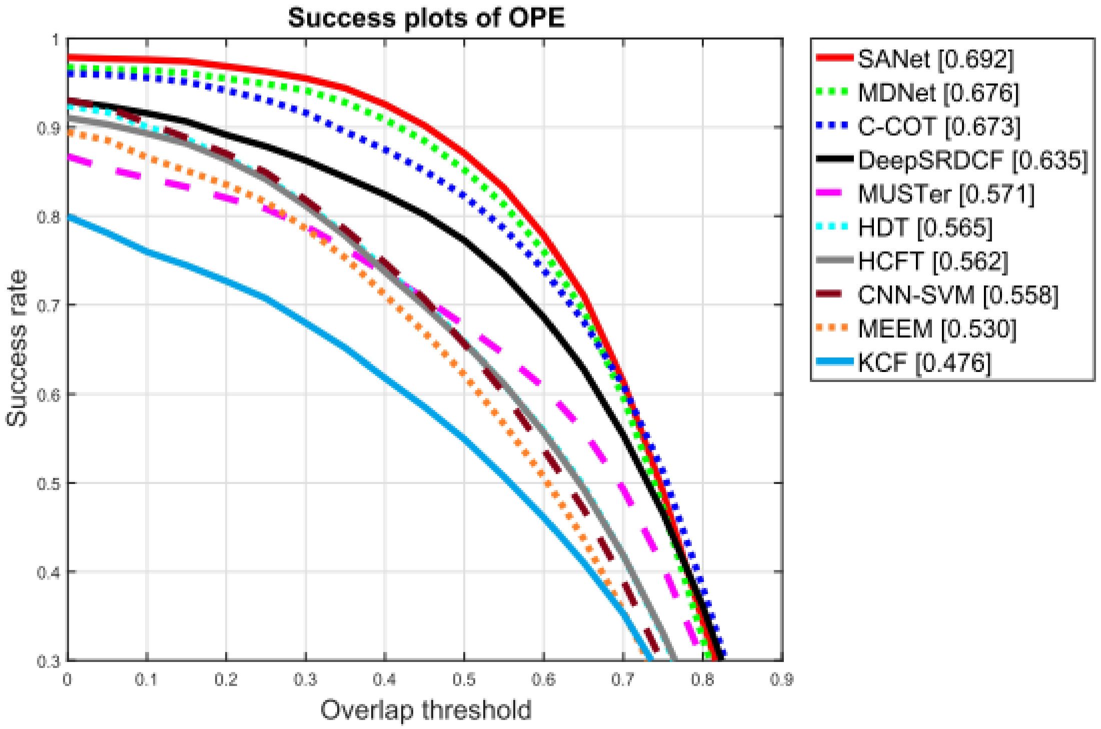


Figure 3. Decomposition of undirected cyclic graph into four directed acyclic graphs. Images (a) and (b) are inputs. Self-structure of object is encoded in an undirected cyclic graph in images (c) and (d). Images (e), (f), (g) and (h) are four directed acyclic graphs along southeast, southwest, northwest and northeast directions.

H. Fan and H. Ling, “SANet: Structure-Aware Network for Visual Tracking,” *arXiv*, 2016.

SANet: Structure-Aware Network for Visual Tracking

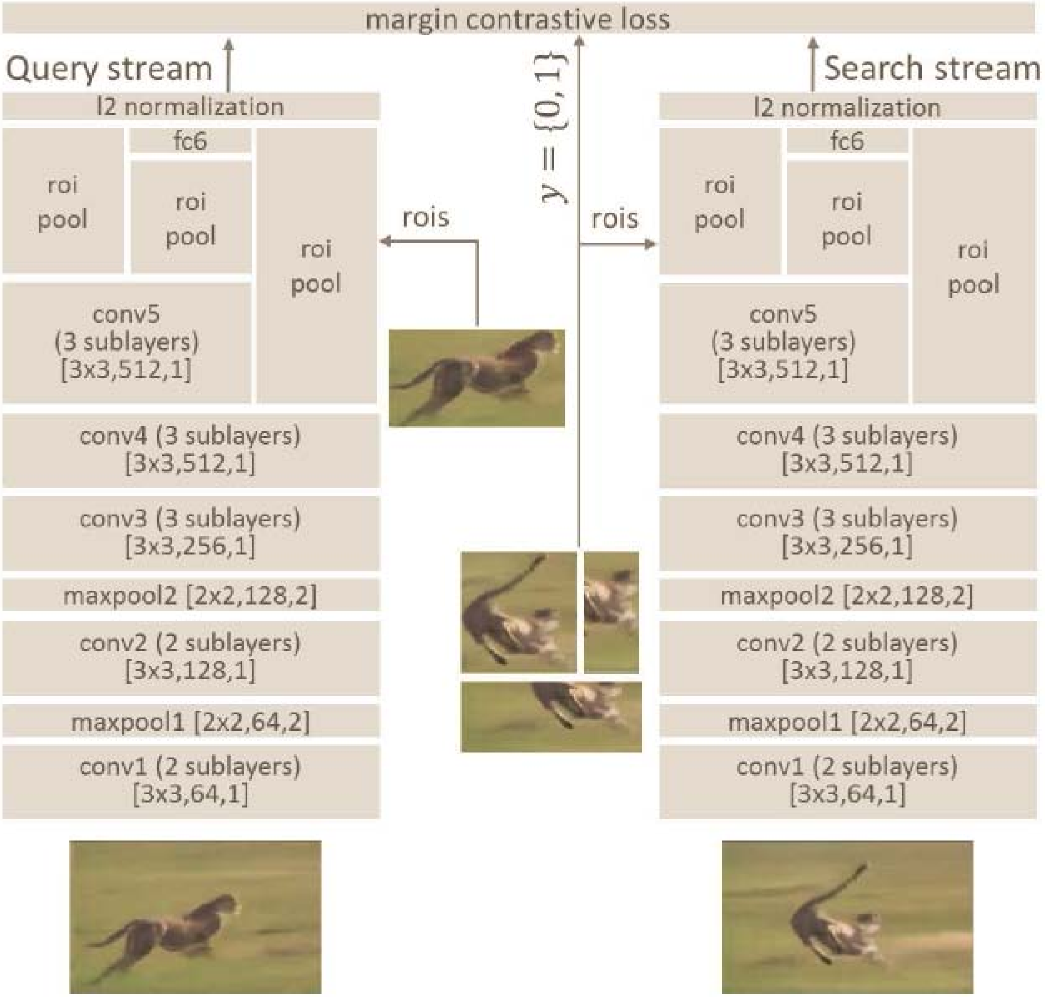




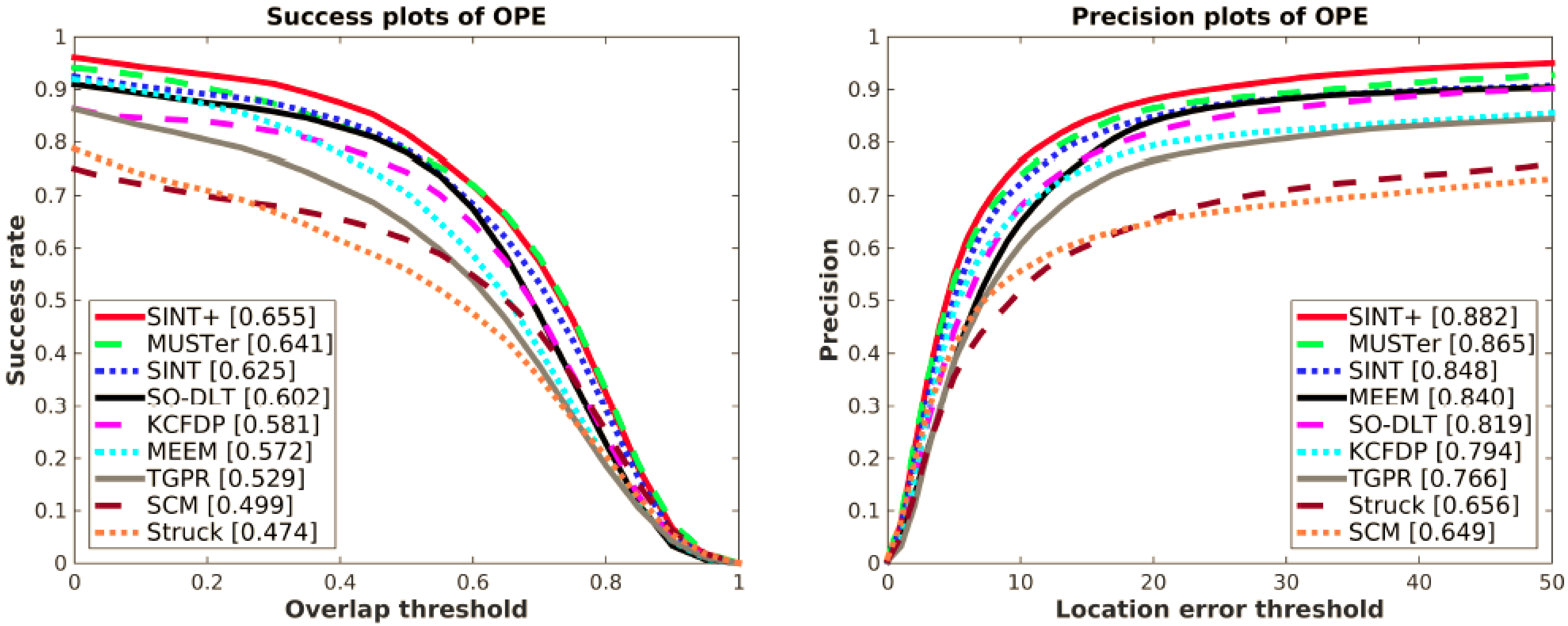
OTB100

Siamese Instance Search for Tracking

- Similarity learning
- No model updating
- ImageNet pretrained network
- Region-of-interest (ROI) pooling layer
- Fusion of multiple layers
- Bounding box regression
- Training set
    - ALOV dataset
    - 60,000 pairs of frames
    - 128 pairs of boxes for each pair of frame


margin contrastive loss
Query stream
Search stream
$y = \{0, 1\}$
l2 normalization
fc6
roi pool
roi pool
roi pool
rois
rois
conv5 (3 sublayers) [3x3,512,1]
conv4 (3 sublayers) [3x3,512,1]
conv3 (3 sublayers) [3x3,256,1]
maxpool2 [2x2,128,2]
conv2 (2 sublayers) [3x3,128,1]
maxpool1 [2x2,64,2]
conv1 (2 sublayers) [3x3,64,1]

R. Tao, E. Gavves, and A. W. M. Smeulders, “Siamese Instance Search for Tracking,” in *CVPR*, 2016.



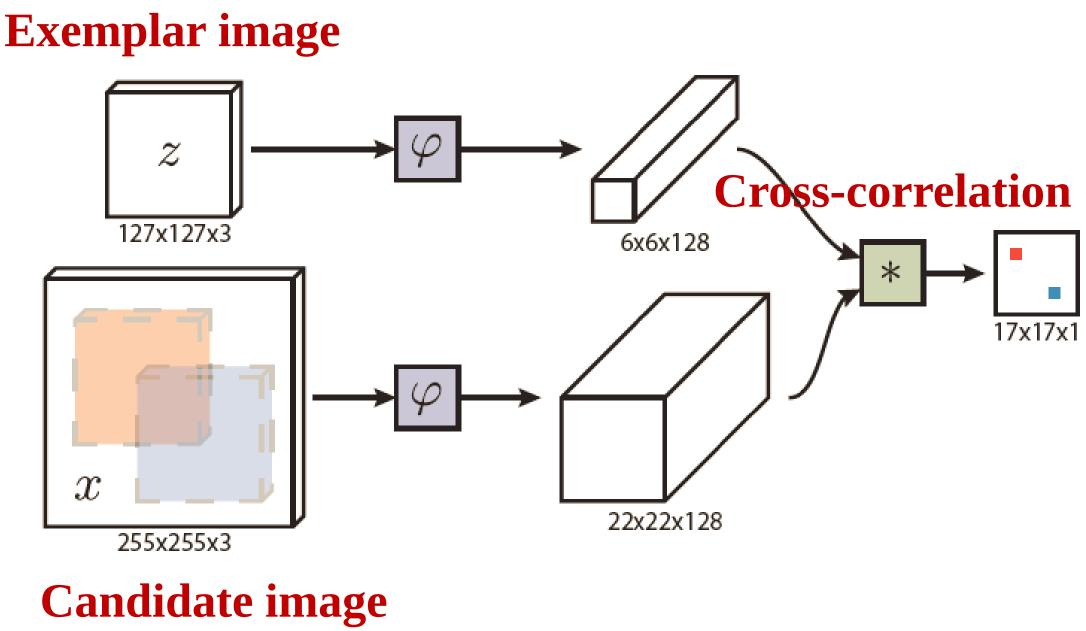
	OPE	TRE	SRE
MEEM	57.2 / 84.0	58.5 / 83.2	51.8 / 76.9
MUSTer	62.1 / 83.6	60.9 / 81.1	56.2 / 78.9
SINT	62.5 / 84.8	64.3 / 84.9	57.9 / 80.6

Robustness evaluation on OTB50

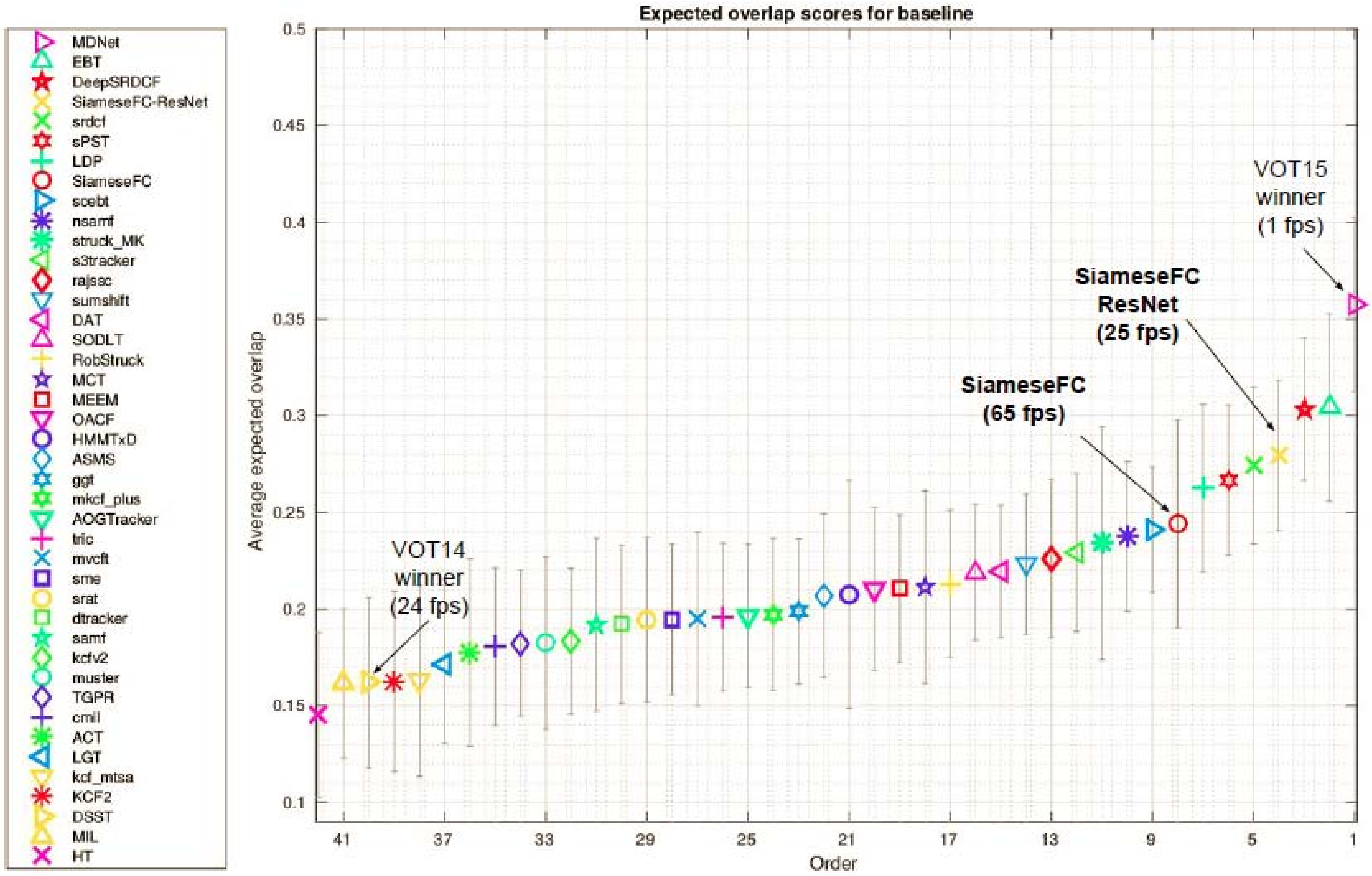
SINT+: a variant of SINT, using an adaptive candidate sampling and optical flow

Fully-Convolutional Siamese Networks for Object Tracking

- Cross-correlation layer
- Output: a scalar-valued score map
- ImageNet Video dataset for training
  - 30 different classes of animals and vehicles
  - Almost 4000 videos for training
  - More than one million annotated frames
  - 843,371 objects from 2820 videos



L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, “Fully-Convolutional Siamese Networks for Object Tracking,” *ECCV*, 2016.



VOT15 winner (1 fps)

SiameseFC ResNet (25 fps)

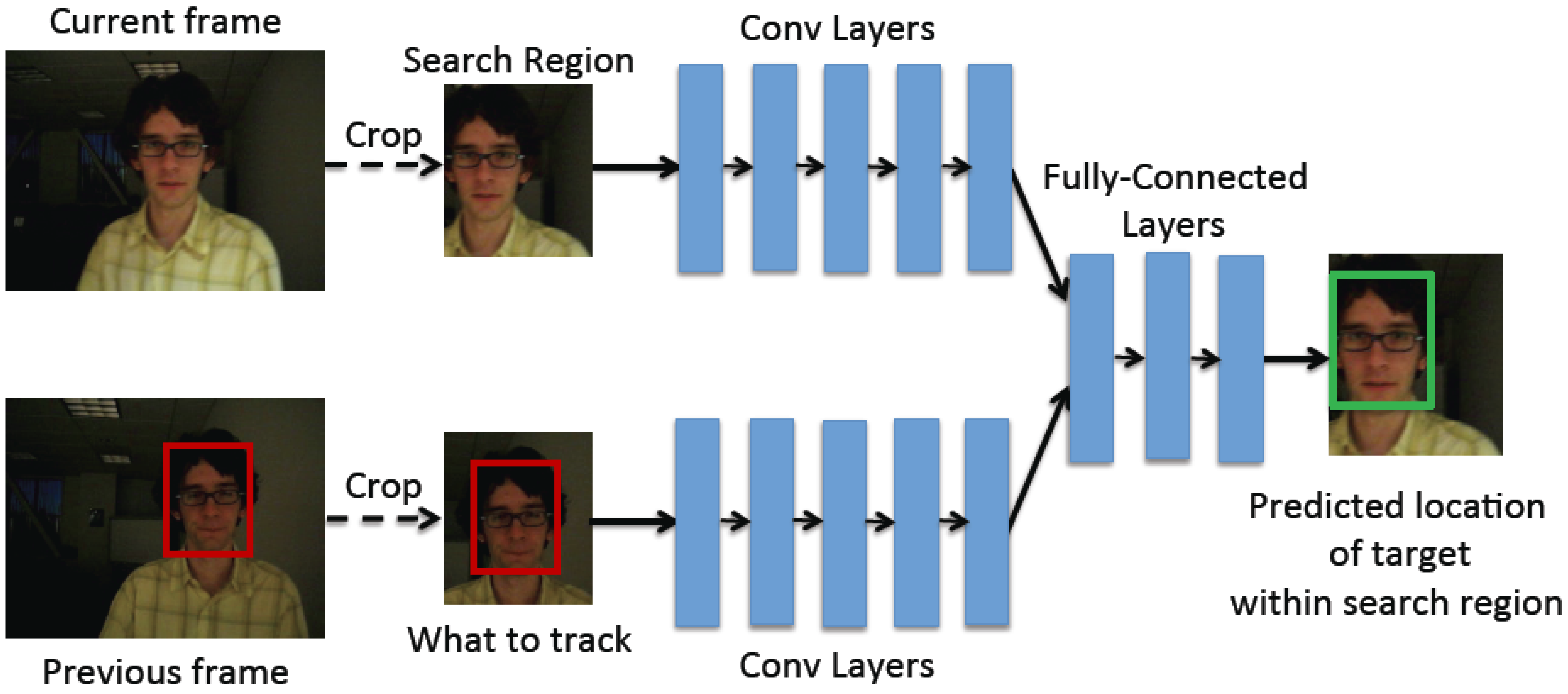
SiameseFC (65 fps)

VOT14 winner (24 fps)

Results on VOT2015

VOT: https://github.com/votchallenge/vot-toolkit

Learning To Track at 100 FPS With Deep Regression Networks

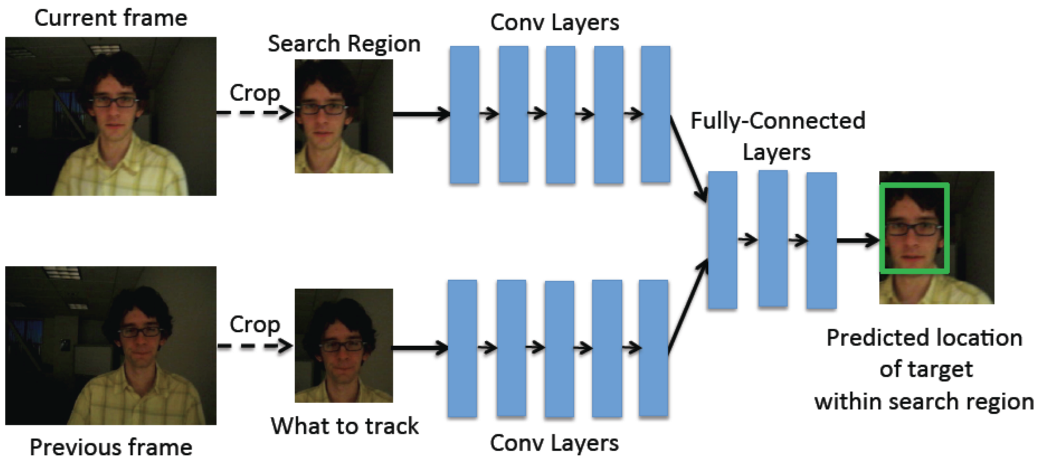


http://davheld.github.io/GOTURN/GOTURN.html

D. Held, S. Thrun, and S. Savarese, “Learning To Track at 100 FPS With Deep Regression Networks,” in *ECCV*, 2016.

Learning To Track at 100 FPS With Deep Regression Networks

- Training set
  - 307 videos from ALOV300
    - 251 for training: 13082 images of 251 objects
    - 56 for validation: 2795 images of 56 objects
  - Training set of imagenet detection challenge
    - 239,283 annotations from 134,821 images
- Test set
  - 25 videos from VOT 2014
- Tracking speed
  - **165 FPS** on GTX Titan X GPU
  - 2.7 FPS using CPU



Current frame

Crop

Search Region

Conv Layers

Fully-Connected Layers

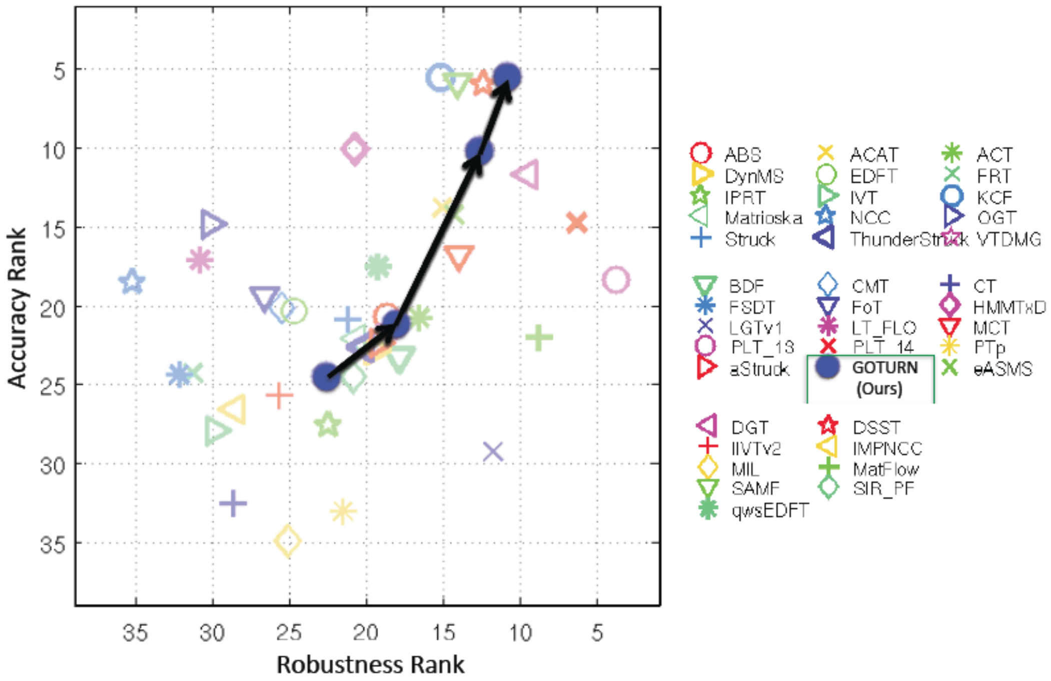
Predicted location of target within search region

Previous frame

Crop

What to track

Conv Layers



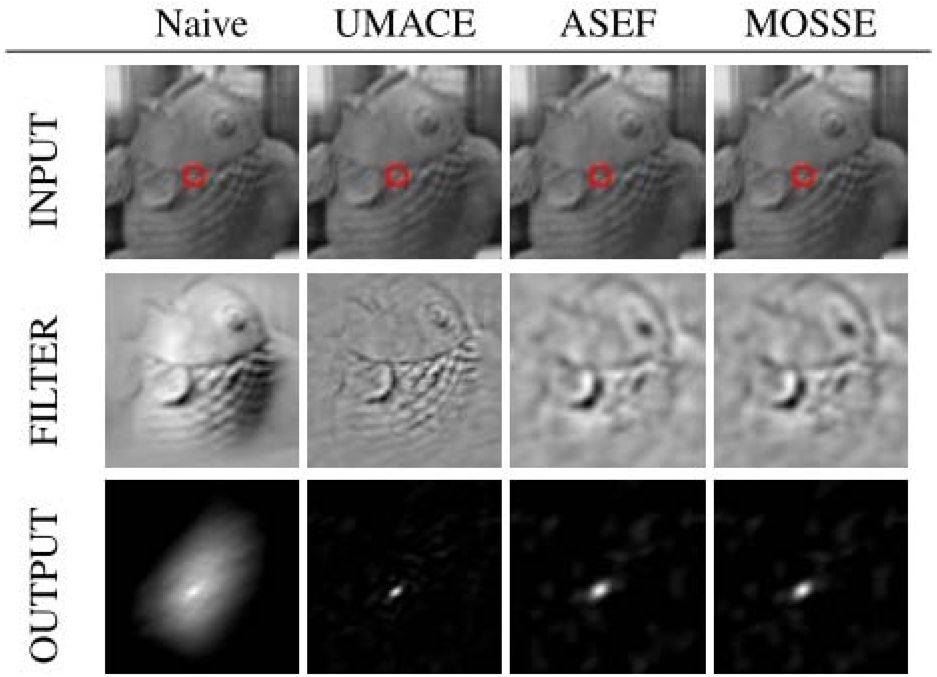
Tracking based on deep learning

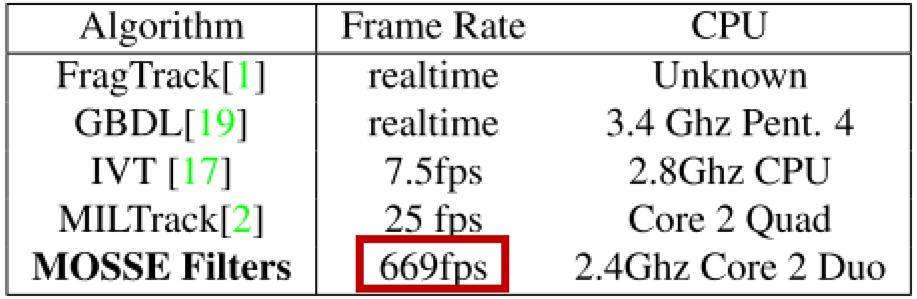
- Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, “Hedged Deep Tracking,” *CVPR*, 2016.
- L. Wang, W. Ouyang, X. Wang, and H. Lu, “STCT: Sequentially Training Convolutional Networks for Visual Tracking,” in *CVPR*, 2016.
- Z. Chi, H. Li, H. Lu, and M.-H. Yang, “Dual Deep Network for Visual Tracking,” *TIP*, vol. 26, no. 4, pp. 2005–2015, 2017.
- G. Ning, Z. Zhang, C. Huang, Z. He, X. Ren, and H. Wang, “Spatially Supervised Recurrent Convolutional Neural Networks for Visual Object Tracking,” *arXiv:1607.05781*, 2016.

Outline

- Tracking based on deep learning
- Tracking based on correlation filters
- Other interesting work

Visual object tracking using adaptive correlation filters



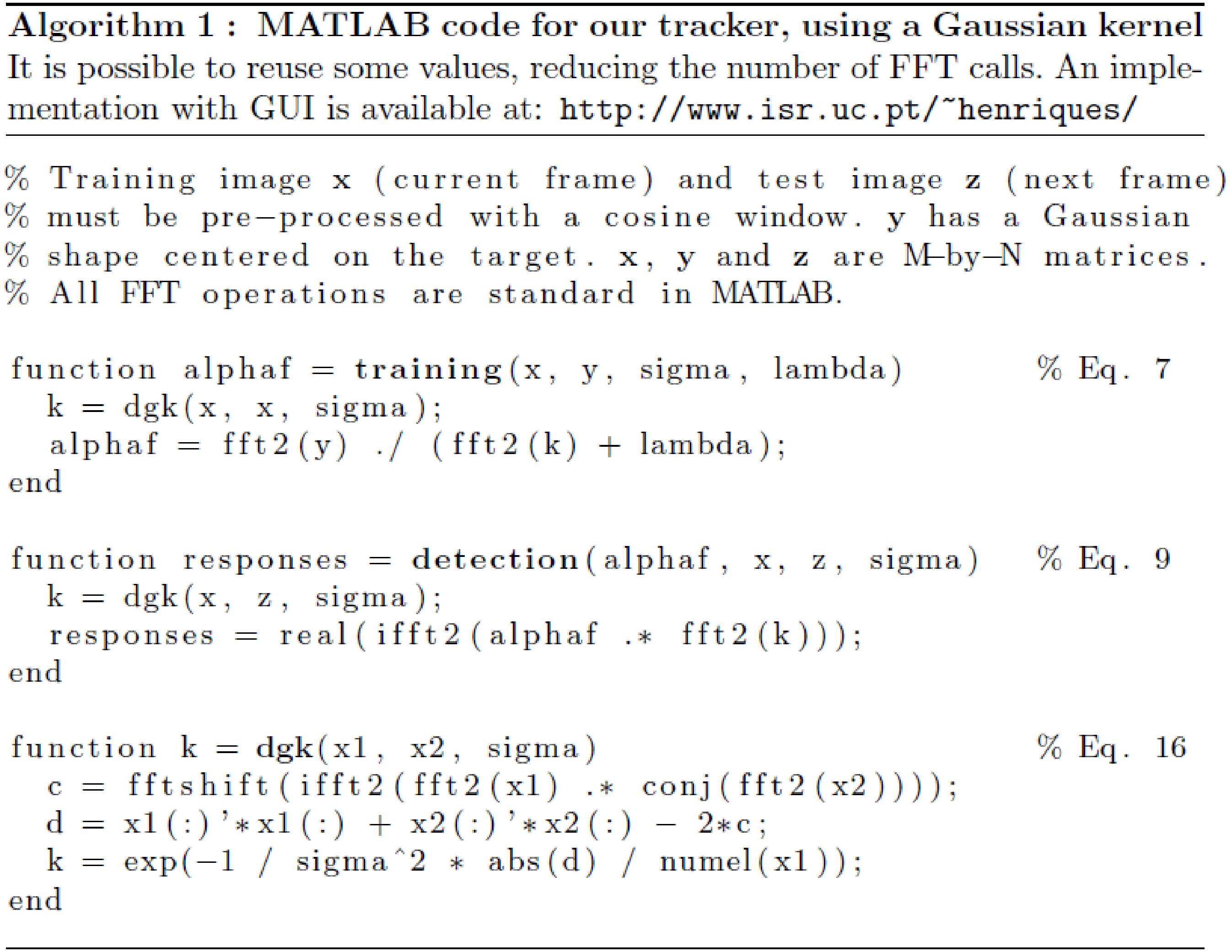


Algorithm	Frame Rate	CPU
FragTrack[1]	realtime	Unknown
GBDL[19]	realtime	3.4 Ghz Pent. 4
IVT [17]	7.5fps	2.8Ghz CPU
MILTrack[2]	25 fps	Core 2 Quad
MOSSE Filters	669fps	2.4Ghz Core 2 Duo

MOSSE $$H^* = \frac{\sum_i G_i \odot F_i^*}{\sum_i F_i \odot F_i^*}$$

D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, “Visual object tracking using adaptive correlation filters,” in *CVPR* 2010

Exploiting the Circulant Structure of Tracking-by-Detection with Kernels



Algorithm 1 : MATLAB code for our tracker, using a Gaussian kernel
It is possible to reuse some values, reducing the number of FFT calls. An implementation with GUI is available at: `http://www.isr.uc.pt/˜henriques/`

```
% Training image x (current frame) and test image z (next frame)
% must be pre-processed with a cosine window. y has a Gaussian
% shape centered on the target. x, y and z are M-by-N matrices.
% All FFT operations are standard in MATLAB.

function alphaf = training(x, y, sigma, lambda)       % Eq. 7
  k = dgk(x, x, sigma);
  alphaf = fft2(y) ./ (fft2(k) + lambda);
end

function responses = detection(alphaf, x, z, sigma)   % Eq. 9
  k = dgk(x, z, sigma);
  responses = real(ifft2(alphaf .* fft2(k)));
end

function k = dgk(x1, x2, sigma)                       % Eq. 16
  c = fftshift(ifft2(fft2(x1) .* conj(fft2(x2))));
  d = x1(:)'*x1(:) + x2(:)'*x2(:) - 2*c;
  k = exp(-1 / sigma^2 * abs(d) / numel(x1));
end
```

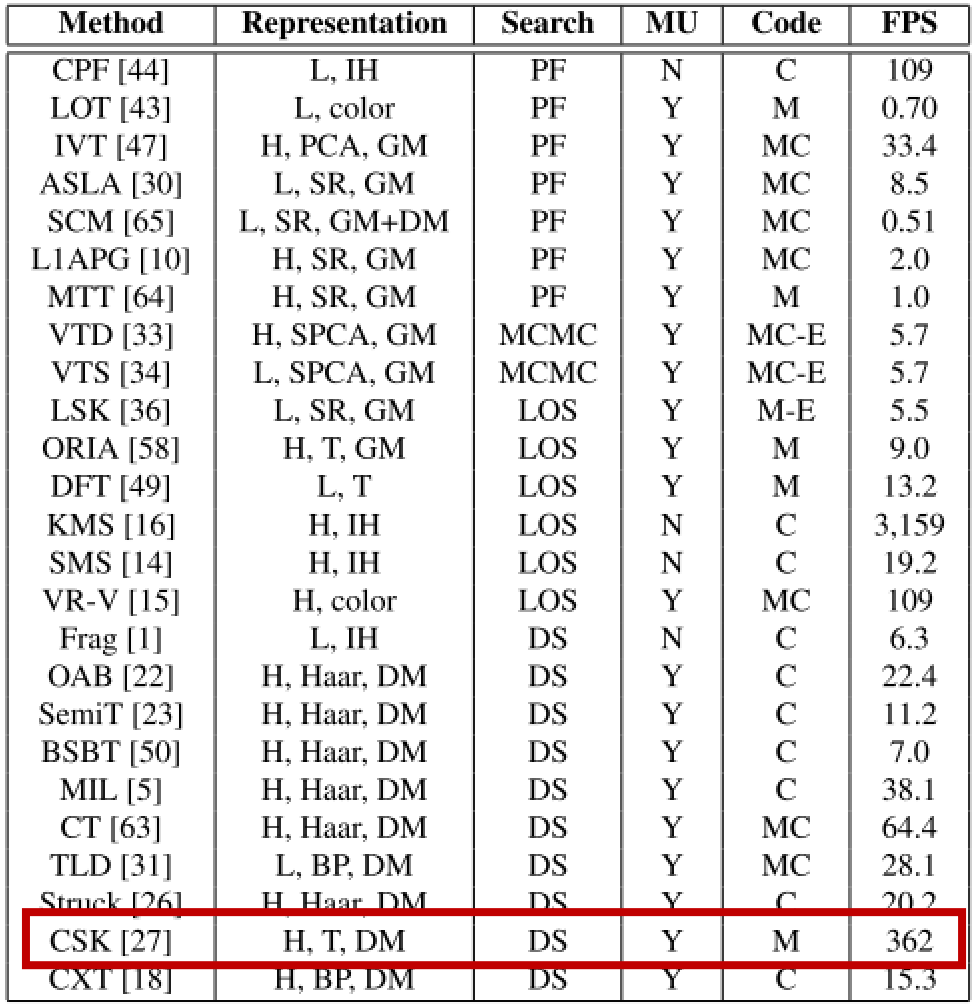
$$\boldsymbol{\alpha} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\mathbf{y})}{\mathcal{F}(\mathbf{k}) + \lambda}\right)$$

$$\hat{\mathbf{y}} = \mathcal{F}^{-1}\left(\mathcal{F}(\bar{\mathbf{k}}) \odot \mathcal{F}(\boldsymbol{\alpha})\right)$$

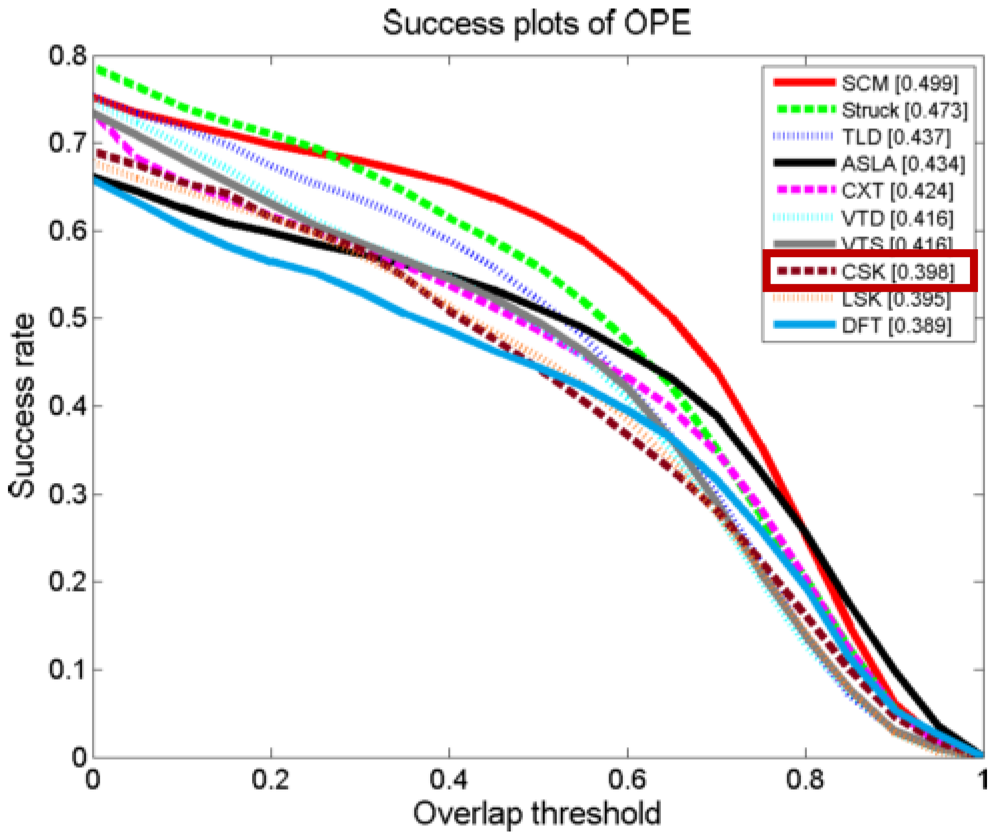
$$\mathbf{k}^{\text{gauss}} = \exp\left(-\frac{1}{\sigma^2}\left(\|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2 - 2\mathcal{F}^{-1}\left(\mathcal{F}(\mathbf{x}) \odot \mathcal{F}^*(\mathbf{x}')\right)\right)\right)$$

J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “Exploiting the Circulant Structure of Tracking-by-Detection with Kernels,” in *ECCV*, 2012.

Exploiting the Circulant Structure of Tracking-by-Detection with Kernels



Method	Representation	Search	MU	Code	FPS
CPF [44]	L, IH	PF	N	C	109
LOT [43]	L, color	PF	Y	M	0.70
IVT [47]	H, PCA, GM	PF	Y	MC	33.4
ASLA [30]	L, SR, GM	PF	Y	MC	8.5
SCM [65]	L, SR, GM+DM	PF	Y	MC	0.51
L1APG [10]	H, SR, GM	PF	Y	MC	2.0
MTT [64]	H, SR, GM	PF	Y	M	1.0
VTD [33]	H, SPCA, GM	MCMC	Y	MC-E	5.7
VTS [34]	L, SPCA, GM	MCMC	Y	MC-E	5.7
LSK [36]	L, SR, GM	LOS	Y	M-E	5.5
ORIA [58]	H, T, GM	LOS	Y	M	9.0
DFT [49]	L, T	LOS	Y	M	13.2
KMS [16]	H, IH	LOS	N	C	3,159
SMS [14]	H, IH	LOS	N	C	19.2
VR-V [15]	H, color	LOS	Y	MC	109
Frag [1]	L, IH	DS	N	C	6.3
OAB [22]	H, Haar, DM	DS	Y	C	22.4
SemiT [23]	H, Haar, DM	DS	Y	C	11.2
BSBT [50]	H, Haar, DM	DS	Y	C	7.0
MIL [5]	H, Haar, DM	DS	Y	C	38.1
CT [63]	H, Haar, DM	DS	Y	MC	64.4
TLD [31]	L, BP, DM	DS	Y	MC	28.1
Struck [26]	H, Haar, DM	DS	Y	C	20.2
CSK [27]	H, T, DM	DS	Y	M	362
CXT [18]	H, BP, DM	DS	Y	C	15.3

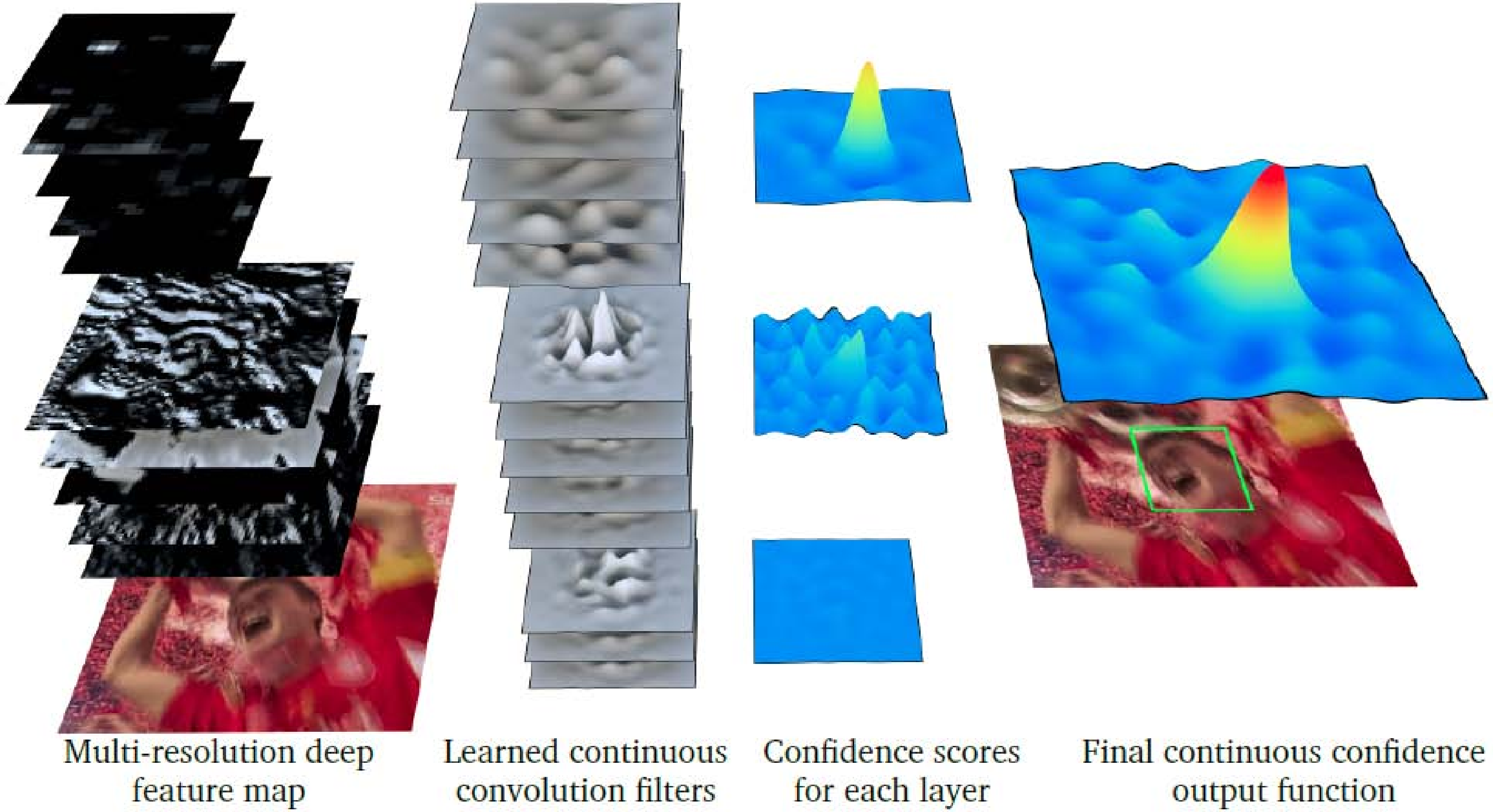


OTB50: Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, “Online Object Tracking: A Benchmark,” in *CVPR*, 2013.

Tracking based on correlation filters

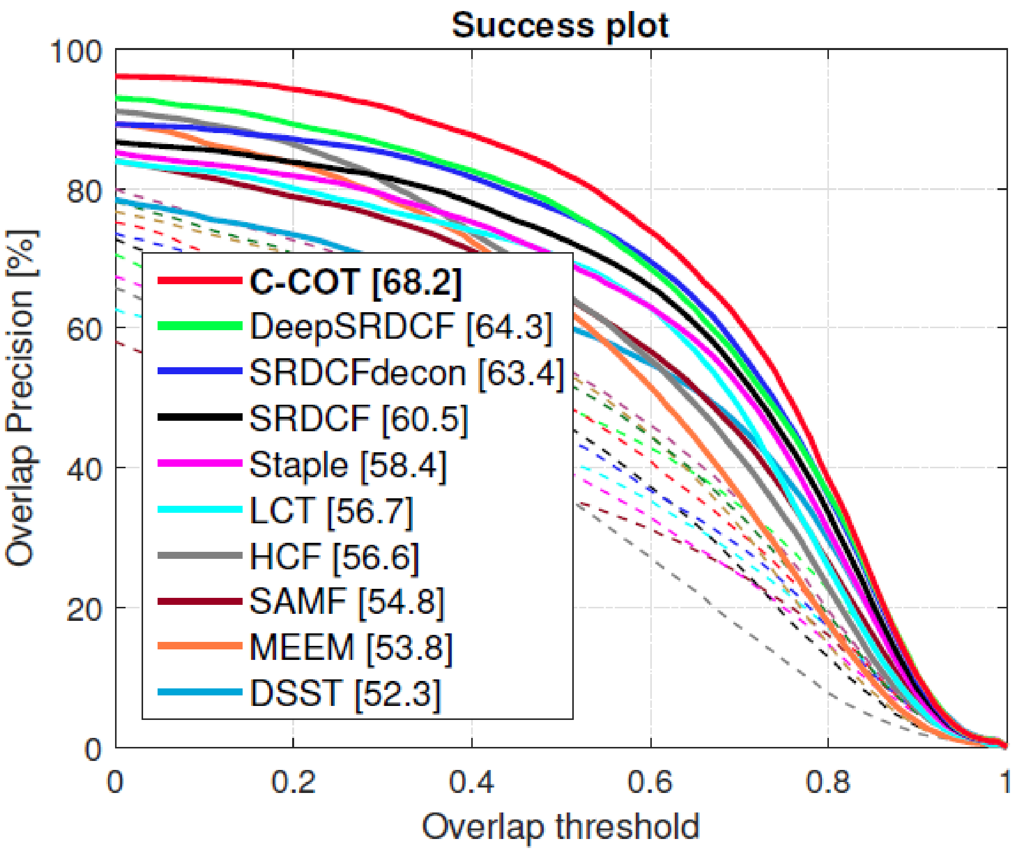
- **Martin Danelljan**, G. Häger, Fahad Shahbaz Khan, and M. Felsberg, “Accurate Scale Estimation for Robust Visual Tracking,” in BMVC, 2014.
- **Martin Danelljan**, F. S. Khan, M. Felsberg, and J. Van De Weijer, “Adaptive Color Attributes for Real-Time Visual Tracking,” in CVPR, 2014.
- T. Liu, G. Wang, and Q. Yang, “Real-time part-based visual tracking via adaptive correlation filters,” in CVPR, 2015.
- Y. Li, J. Zhu, and S. C. H. Hoi, “Reliable Patch Trackers: Robust Visual Tracking by Exploiting Reliable Patches,” in CVPR, 2015.
- C. Ma, X. Yang, C. Zhang, and M.-H. Yang, “Long-term Correlation Tracking,” in CVPR, 2015.
- **Martin Danelljan**, G. Hager, F. S. Khan, and M. Felsberg, “Convolutional Features for Correlation Filter Based Visual Tracking,” in ICCV Workshops, 2015.
- **Martin Danelljan**, H. Gustav, F. S. Khan, and M. Felsberg, “Learning Spatially Regularized Correlation Filters for Visual Tracking,” in *ICCV*, 2015.
- M. Tang and J. Feng, “Multi-kernel correlation filter for visual tracking,” in *ICCV*, 2015.

Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking

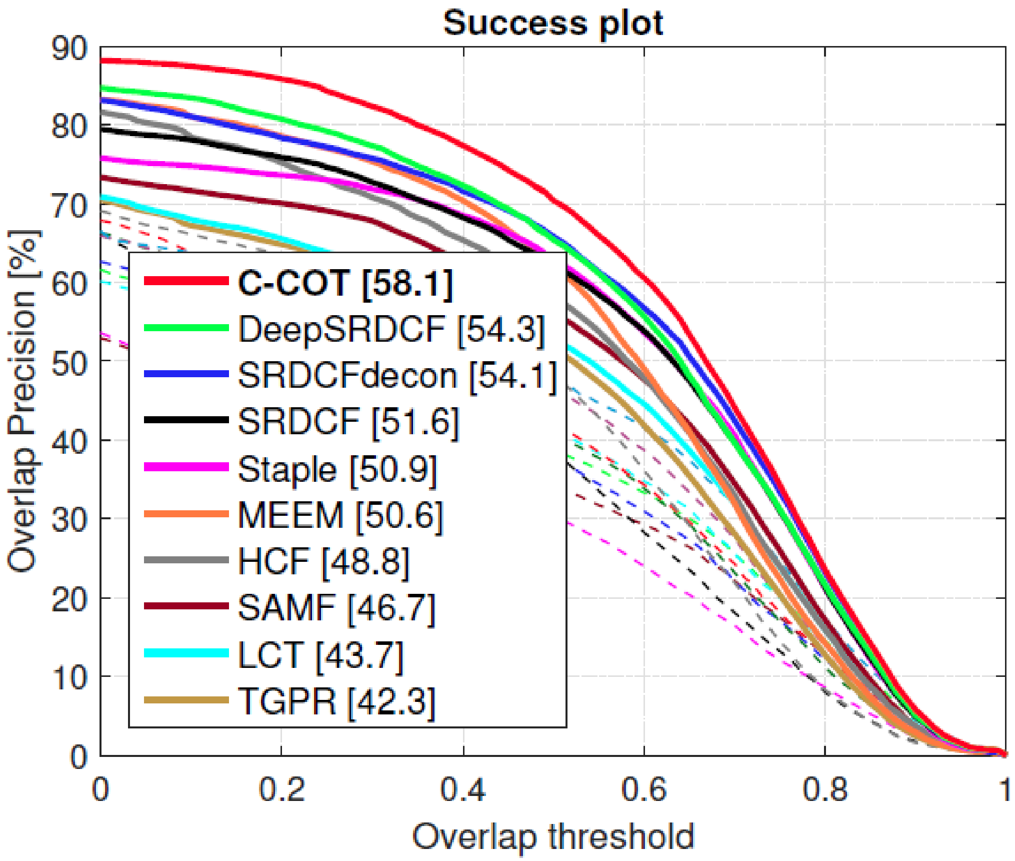

Multi-resolution deep feature map
Learned continuous convolution filters
Confidence scores for each layer
Final continuous confidence output function

M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking," in *ECCV*, 2016.

Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking

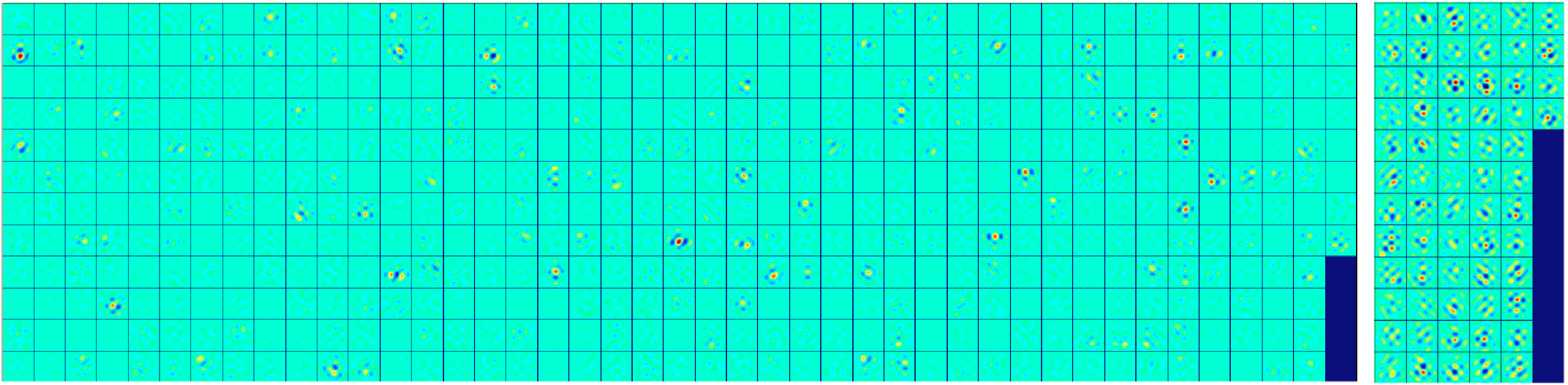


OTB100



Temple-Color

ECO: Efficient Convolution Operators for Tracking



(a) 512 filters learned by C-COT

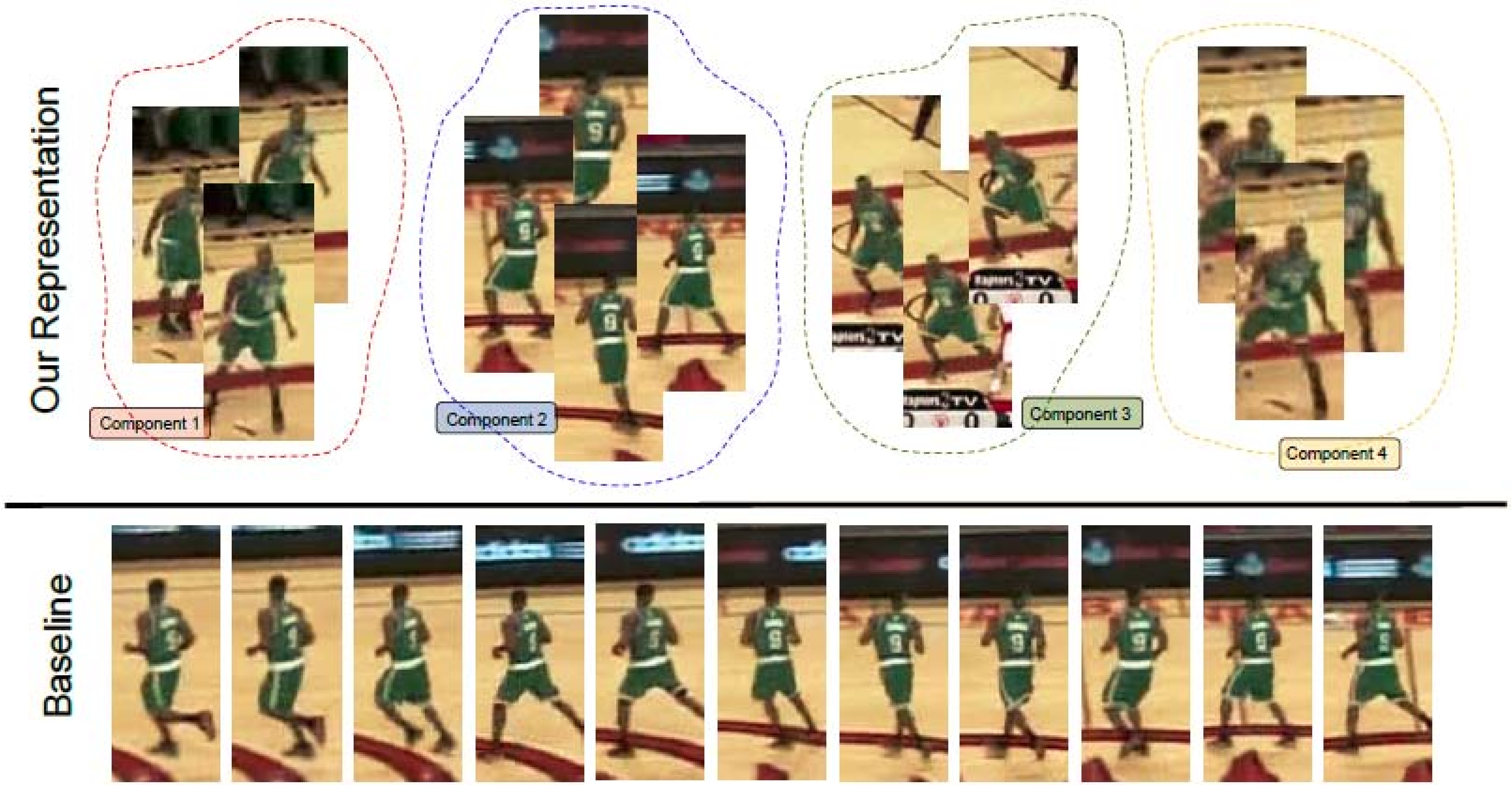
(b) 64 filters of ECO

Visualization of the learned filters corresponding to the last convolutional layer in the deep network

	Conv-1	Conv-5	HOG	CN
Feature dimension, D	96	512	31	11
Filter dimension, C	16	64	10	3

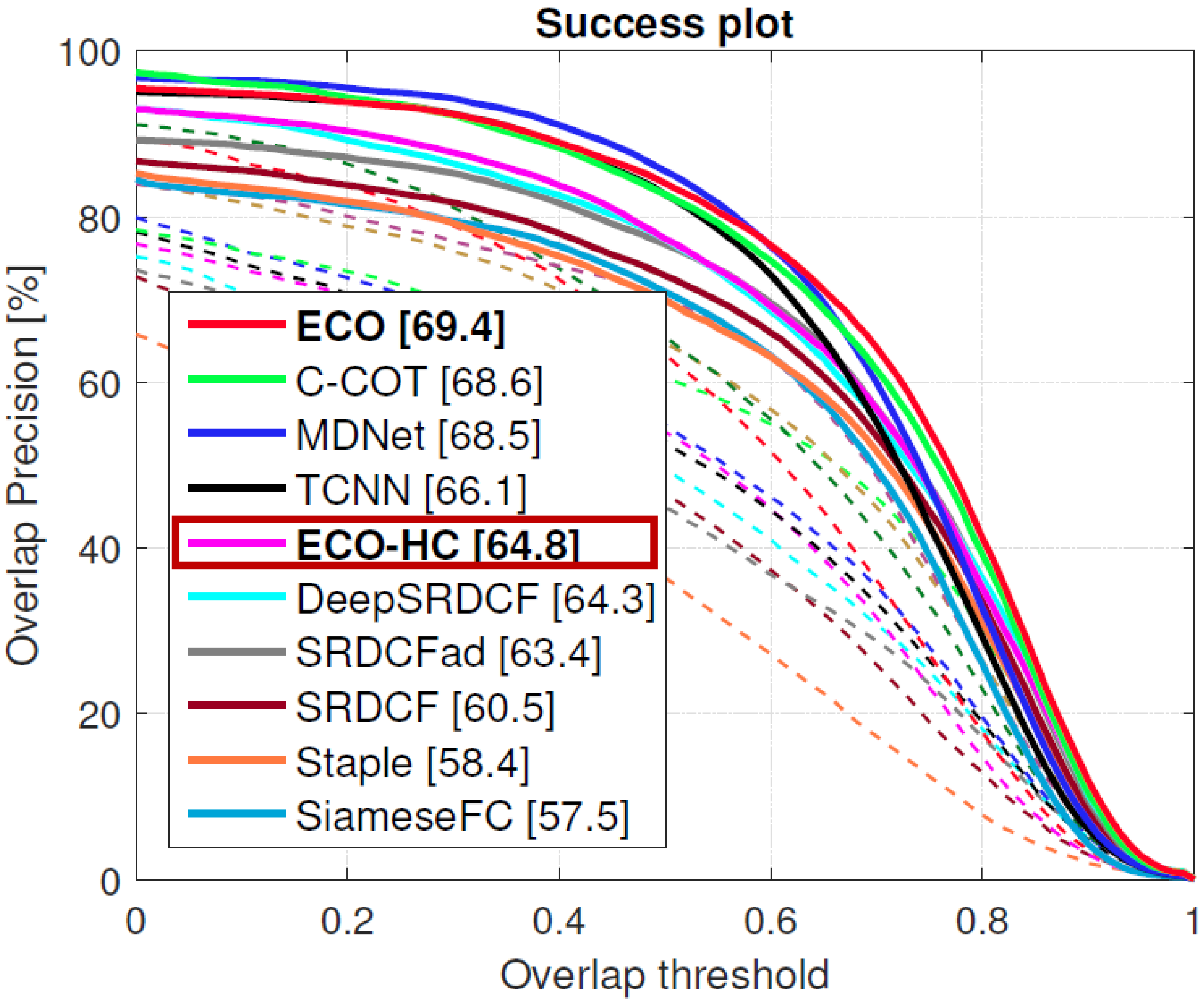
M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, “ECO: Efficient Convolution Operators for Tracking,” in *CVPR*, 2017.

ECO: Efficient Convolution Operators for Tracking



ECO: Efficient Convolution Operators for Tracking

- ECO
  - Conv-1, Conv-5 in the VGG-m network, HOG and Color-Names (CN)
  - 8 fps using Tesla K40 GPU
- ECO-HC
  - HOG and Color-Names (CN)
  - 60 fps using Intel i7-6700 CPU



Results on OTB100

Tracking based on correlation filters

- G. Zhu, J. Wang, Y. Wu, X. Zhan, and H. Lu, “MC-HOG Correlation Tracking with Saliency Proposal,” in *AAAI*, 2016.
- Y. Sui, Z. Zhang, G. Wang, Y. Tang, and L. Zhang, “Real-Time Visual Tracking: Promoting the Robustness of Correlation Filter Learning,” in *ECCV*, 2016.
- S. Liu, T. Zhang, X. Cao, and C. Xu, “Structural Correlation Filter for Robust Visual Tracking,” in *CVPR*, 2016.
- Z. Cui, S. Xiao, J. Feng, and S. Yan, “Recurrently Target-Attending Tracking,” in *CVPR*, 2016.
- A. Bibi, M. Mueller, and B. Ghanem, “Target response adaptation for correlation filter tracking,” in *ECCV*, 2016.
- H. K. Galoogahi, A. Fagg, and S. Lucey, “Learning Background-Aware Correlation Filters for Visual Tracking,” *arXiv: 1703.04590*, 2017.

Outline

- Tracking based on deep learning
- Tracking based on correlation filters
- Other interesting work

Other interesting work - dataset

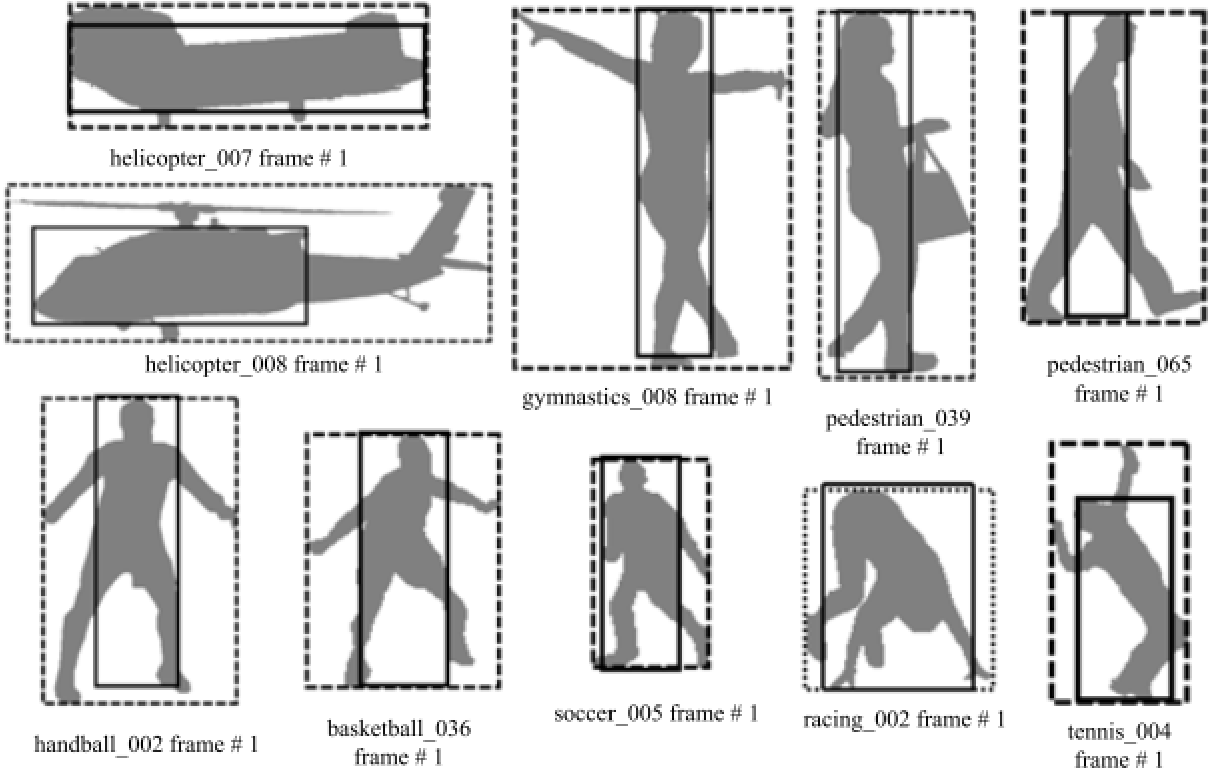
- M. Mueller, N. Smith, and B. Ghanem, “A Benchmark and Simulator for UAV Tracking,” in *ECCV*, 2016.
- A. Li, M. Lin, Y. Wu, M.-H. Yang, and S. Yan, “NUS-PRO: A New Visual Tracking Challenge,” *PAMI*, vol. 38, no. 2, pp. 335–349, 2016.
- E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, “YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video,” *arXiv:1702.00824*, 2017.

A Benchmark and Simulator for UAV Tracking

M. Mueller, N. Smith, and B. Ghanem, “A Benchmark and Simulator for UAV Tracking,” in *ECCV*, 2016.

NUS-PRO: A New Visual Tracking Challenge

- 12 categories
- 365 videos



helicopter_007 frame # 1

helicopter_008 frame # 1

gymnastics_008 frame # 1

pedestrian_039 frame # 1

pedestrian_065 frame # 1

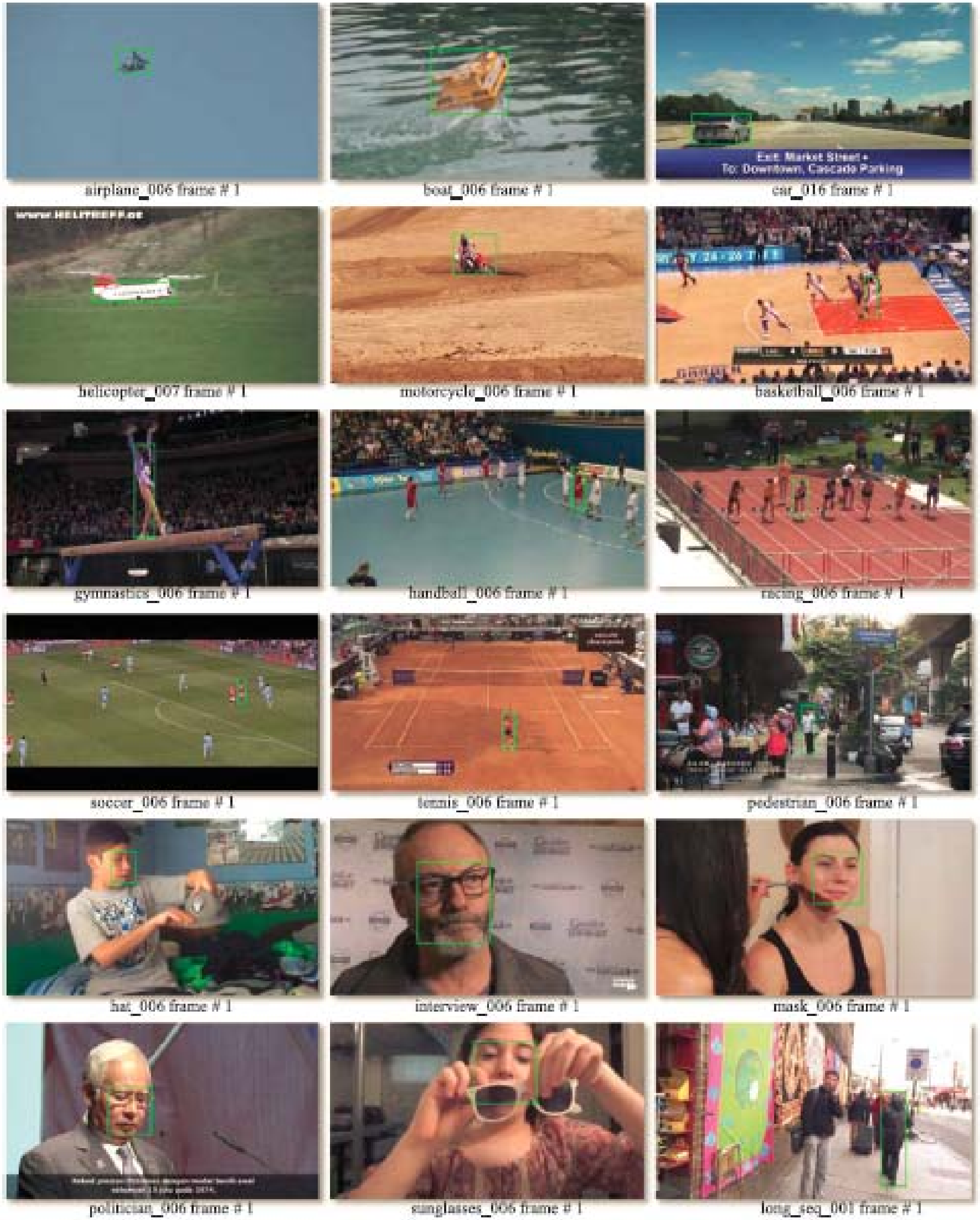
handball_002 frame # 1

basketball_036 frame # 1

soccer_005 frame # 1

racing_002 frame # 1

tennis_004 frame # 1



airplane_006 frame # 1

boat_006 frame # 1

car_016 frame # 1

helicopter_007 frame # 1

motorcycle_006 frame # 1

basketball_006 frame # 1

gymnastics_006 frame # 1

handball_006 frame # 1

racing_006 frame # 1

soccer_006 frame # 1

tennis_006 frame # 1

pedestrian_006 frame # 1

hat_006 frame # 1

interview_006 frame # 1

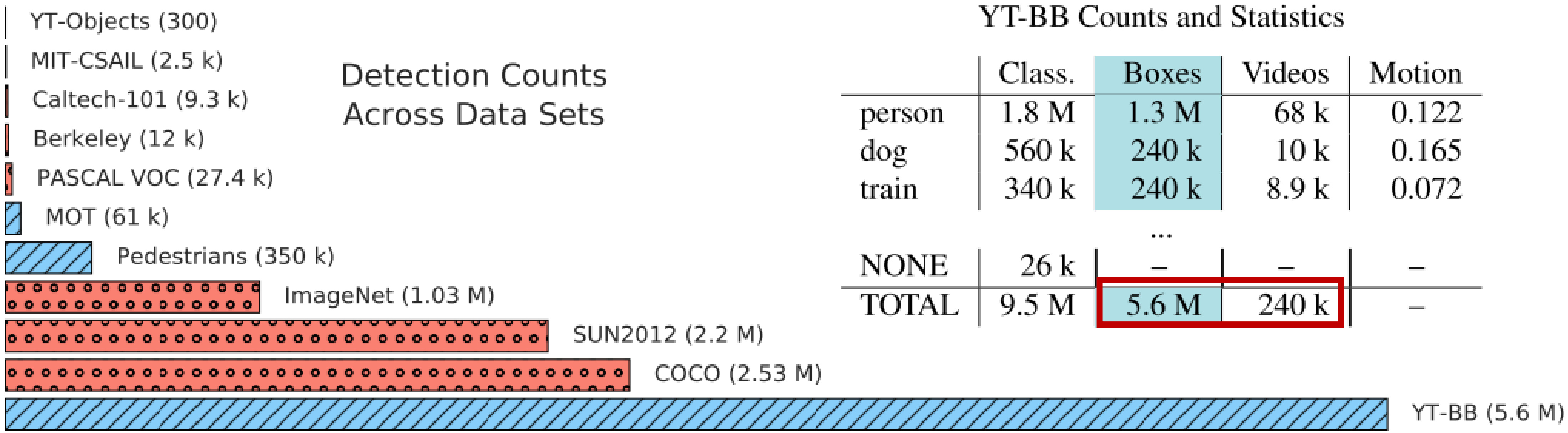
mask_006 frame # 1

politician_006 frame # 1

sunglasses_006 frame # 1

long_seq_001 frame # 1

YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video



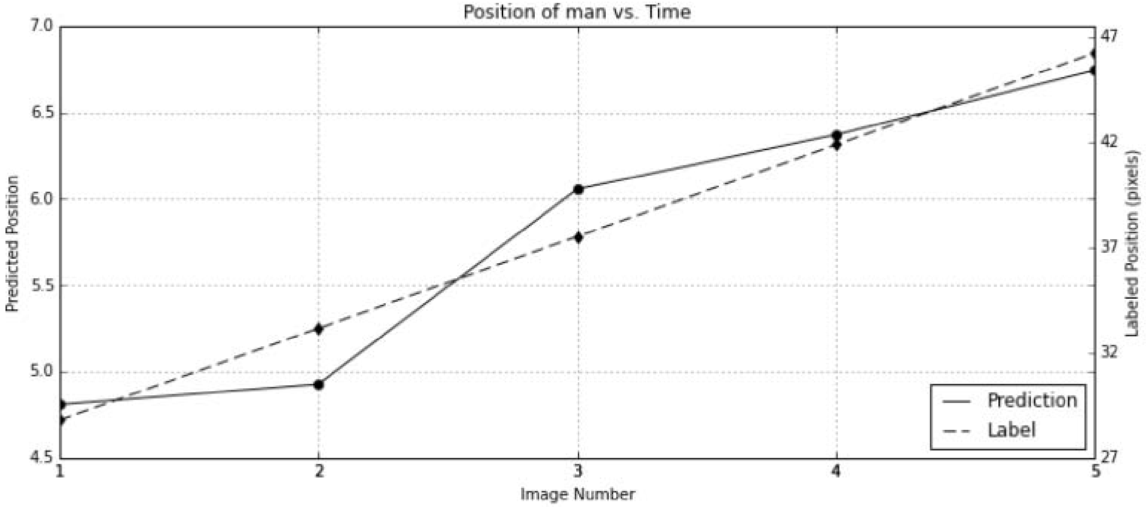
YT-BB Counts and Statistics

	Class.	Boxes	Videos	Motion
person	1.8 M	1.3 M	68 k	0.122
dog	560 k	240 k	10 k	0.165
train	340 k	240 k	8.9 k	0.072
		...		
NONE	26 k	–	–	–
TOTAL	9.5 M	5.6 M	240 k	–

E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video," *arXiv:1702.00824*, 2017.

Other interesting work – AAAI best paper

- R. Stewart and S. Ermon, “Label-Free Supervision of Neural Networks with Physics and Domain Knowledge,” in *AAAI*, 2017.



Other interesting work

- L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. Torr, “Staple: Complementary Learners for Real-Time Tracking,” in *CVPR*, 2016.
- K. Krafka, A. Khosla, P. Kellnhofer, and H. Kannan, “Eye Tracking for Everyone,” in *CVPR*, 2016.
- J. Xiao, L. Qiao, R. Stolkin, and A. Leonardis, “Distractor-supported single target tracking in extremely cluttered scenes,” in *ECCV*, 2016.
- Y. Sui, G. Wang, Y. Tang, and L. Zhang, “Tracking Completion,” *ECCV*, 2016.
- S. Zhang, Y. Gong, J. Bin Huang, J. Lim, J. Wang, N. Ahuja, and M. H. Yang, “Tracking persons-of-interest via adaptive discriminative features,” in *ECCV*, 2016.
- E. Sánchez-Lozano, B. Martinez, G. Tzimiropoulos, and M. Valstar, “Cascaded Continuous Regression for Real-time Incremental Face Tracking,” in *ECCV*, 2016.
- J. Choi, H. Jin, C. Jiyeoup, J. Yiannis, D. Jin, and Y. Choi, “Visual Tracking Using Attention-Modulated Disintegration and Integration,” in *CVPR*, 2016.
- M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, “Adaptive Decontamination of the Training Set: A Unified Formulation for Discriminative Visual Tracking,” in *CVPR*, 2016.
- G. Zhu, F. Porikli, and H. Li, “Beyond Local Search: Tracking Objects Everywhere with Instance-Specific Proposals,” in *CVPR*, 2016.

Visual-tracking.net

Thank you!