Supplementary Material of "Learning Attentions: Residual Attentional Siamese Network for High Performance Online Visual Tracking"

Qiang Wang[1,3*], Zhu Teng[2*], Junliang Xing[3†], Jin Gao[3], Weiming Hu[1,3], Stephen Maybank[4]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[3]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4]Department of Computer Science and Information Systems, Birkbeck College, University of London, UK.
{qiang.wang, jlxing, jin.gao, wmhu}@nlpr.ia.ac.cn zteng@bjtu.edu.cn sjmaybank@dcs.bbk.ac.uk

In this section, we provide some additional information of the RASNet tracker and extra experimental results as a supplement for the paper. To facilitate further studies, our source code and trained models are available at: https://github.com/foolwood/RASNet.

1. RASNet Architecture

The detailed architecture of the networks proposed in the paper is further elaborated in this section. We first define some abbreviations for the basic components used in the networks.

- *conv(w, c, s)*: convolution with $w \times w$ window size, c output channels and s stride.
- *dconv(w, c, s)* deconvolution (transpose of the convolution operation) with $w \times w$ window size, c output channels and s stride.
- *fc(s)*: a fully-connected layer with the output size s.
- *pool(w,s)*: a max pooling layer with $w \times w$ window size and s stride.
- *global pool*: a global average pooling layer.
- *sigmoid*: a sigmoid activation layer.

Feature extractor. We follow the same feature extractor in SiamFC [1] for fair comparisons. Table. 1 shows the details.

Residual Attention Network We construct 3 convolutional layers and 3 deconvolutional layers (Convolutional Encoder-Decoder) as delineated in Table. 2.

Dual Attention Network A superposition of *general attention* and *residual attention* is proposed in RASNet as shown in Fig. 1. We observe that the general attention in the tracking application is Gaussian-like distribution and the expected value of activations in the residual attention approximates towards 0, which is different from AttentionNet [5] with uniformly distribution in classification area. We call this method residual attention learning for visual tracking. The receptive field of the network is shown in Fig. 1(b). It is clear that the reception field in the residual attention map is the entire exemplar image, which facilitates the net with global information.

Channel Attention Network A compact network is detailed in Table. 3.

Table 1: Feature extractor Network

Layer	for examplar	for search	chans.
	127×127	255×255	$\times 3$
$conv(11, 96, 2)$	59×59	123×123	$\times 96$
$pool(3, 2)$	29×29	61×61	$\times 96$
$conv(5, 256, 1)$	25×25	57×57	$\times 256$
$pool(3, 2)$	12×12	28×28	$\times 256$
$conv(3, 384, 1)$	10×10	26×26	$\times 384$
$conv(3, 384, 1)$	8×8	24×24	$\times 384$
$conv(3, 256, 1)$	6×6	22×22	$\times 256$

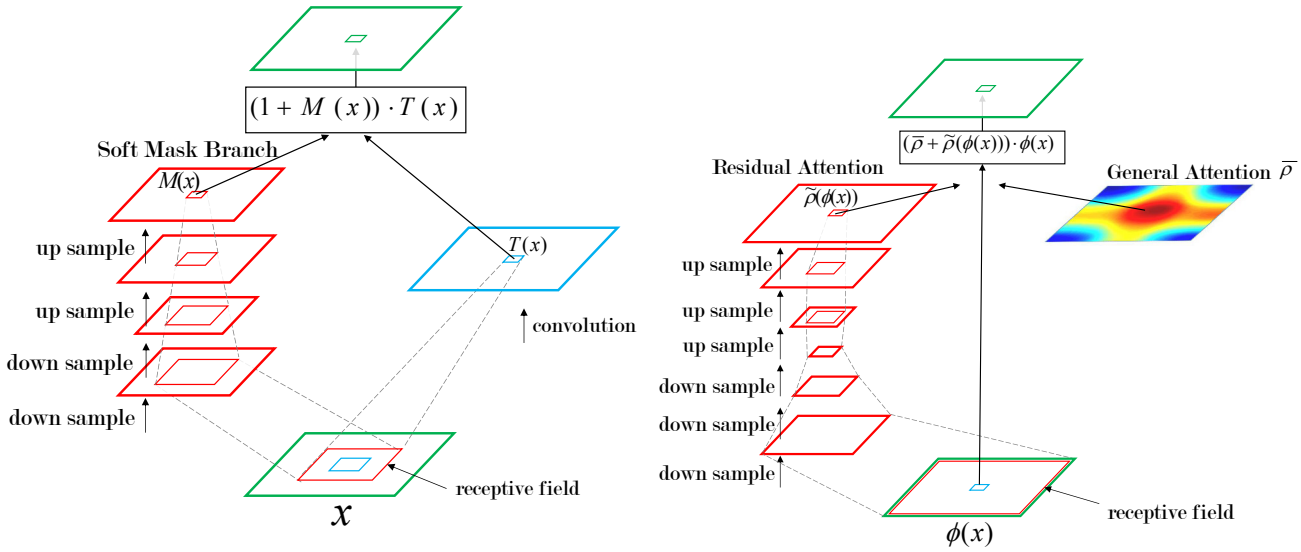
Table 2: Residual Attention Network

Layer	$\phi(\mathbf{x})$	chans.
	6×6	$\times 256$
$conv(3, 256, 1)$	4×4	$\times 256$
$conv(3, 356, 1)$	2×2	$\times 256$
$conv(2, 256, 1)$	1×1	$\times 256$
$dconv(2, 256, 1)$	2×2	$\times 256$
$dconv(3, 256, 1)$	4×4	$\times 256$
$dconv(3, 1, 1)$	6×6	$\times 1$

Table 3: Channel Attention Network

Layer	$\phi(\mathbf{x})$	chans.
	6×6	$\times 256$
$globalpool$	1×1	$\times 256$
$fc(64)$	1×1	$\times 64$
$fc(256)$	1×1	$\times 256$
$sigmoid$	1×1	$\times 256$

*Equal contribution.
†Corresponding author.



(a) AttentionNet (b) Dual Attention

Figure 1: Illustration for Attention architectures. (a) residual attention network for image classification (b) residual attention module used in our RASNet for visual tracking.

2. Attribute evaluation on OTB

The success plots evaluated on 11 attributes separately (illumination variation (IV), scale variation (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), background clutters (BC), low resolution (LR)) are shown in Fig. 2 for OTB-2013 and presented in Fig. 3 for OTB-2015.

3. Qualitative evaluation on OTB

Fig. 4 reports part of tracking results on ten challenging video sequences, and the comparative trackers include CREST [2], SINT [3], SiamFC [1], and CFNet [4].

4. Limitations

The proposed tracker may get a risk in long-term visual tracking since there is no online training in the tracker. To pursue both tracking speed and tracking performance, we will conduct further studies on combining a fast on-line learning strategy with RASNet.

References

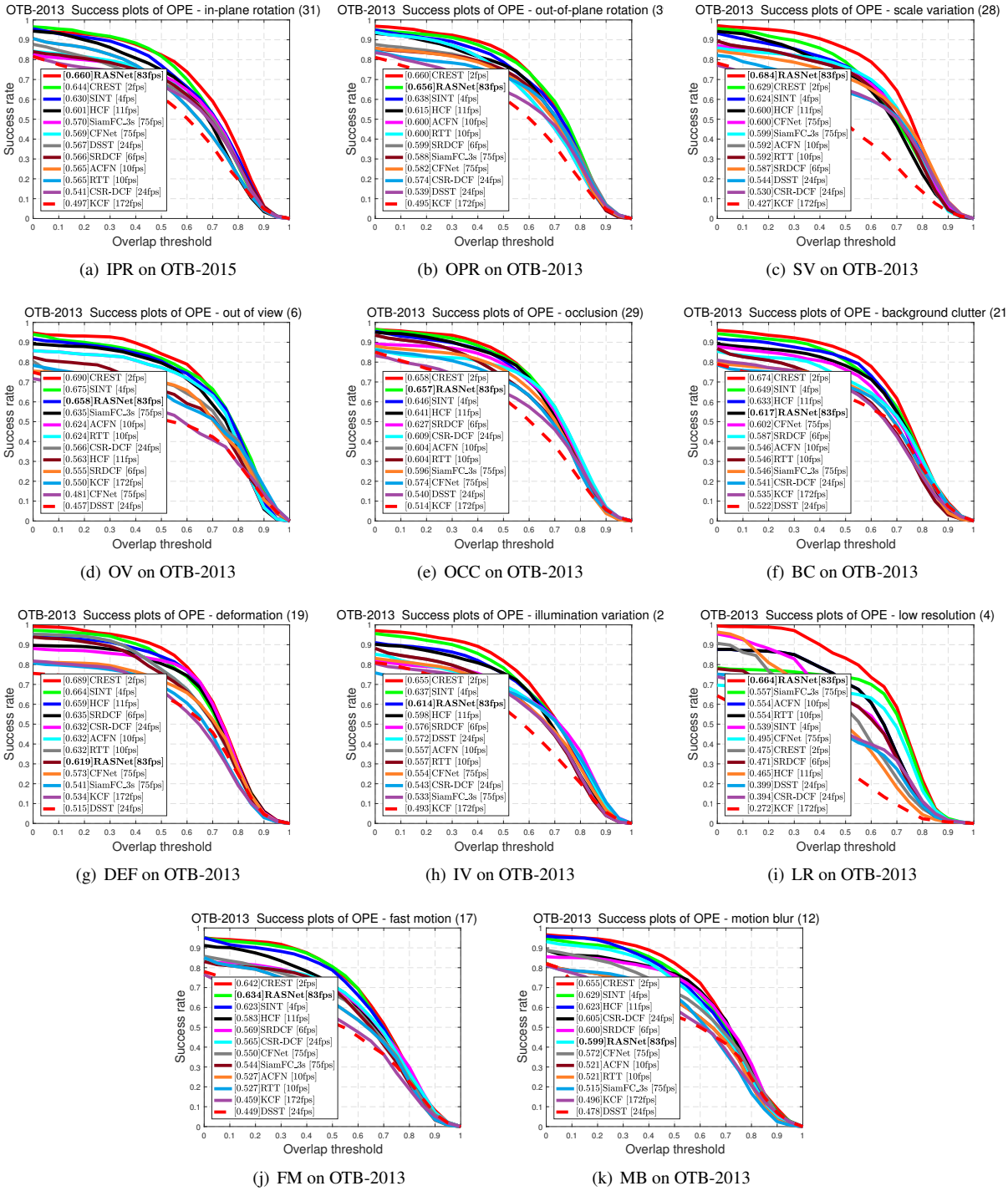
[1] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*, pages 850–865, 2016. 1, 2, 5

[2] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M.-H. Yang. Crest: Convolutional residual learning for visual tracking. In *IEEE International Conference on Computer Vision*, Oct 2017. 2, 5

[3] R. Tao, E. Gavves, and A. W. M. Smeulders. Siamese instance search for tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 5

[4] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. Torr. End-to-end representation learning for correlation filter based tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 5

[5] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017. 1



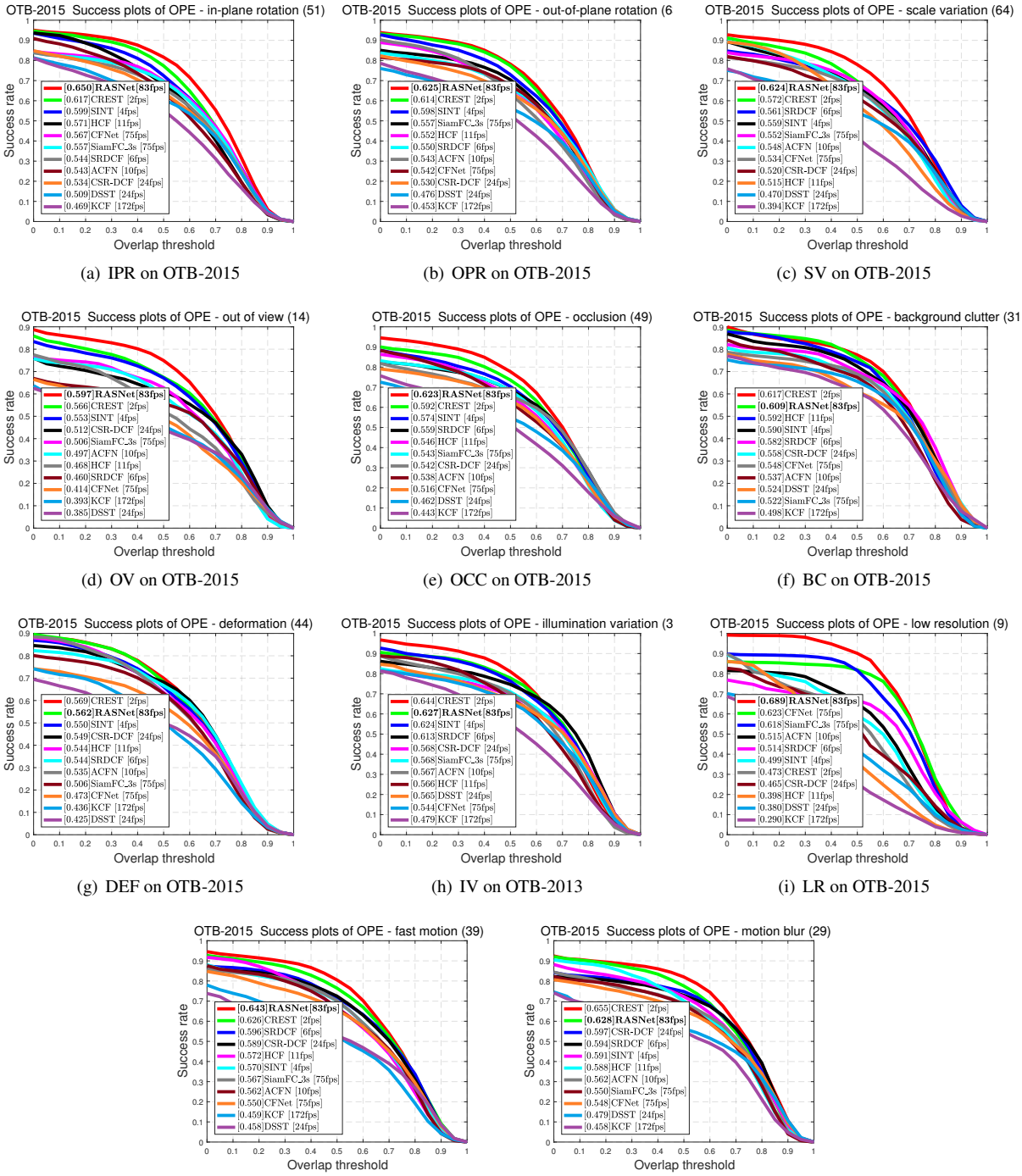
(a) IPR on OTB-2015 (b) OPR on OTB-2013 (c) SV on OTB-2013

(d) OV on OTB-2013 (e) OCC on OTB-2013 (f) BC on OTB-2013

(g) DEF on OTB-2013 (h) IV on OTB-2013 (i) LR on OTB-2013

(j) FM on OTB-2013 (k) MB on OTB-2013

Figure 2: The success plots on OTB-2013 for eleven challenge attributes: in-plain rotation, out-of-plane rotation, scale variation, out of view, occlusion, background clutter, deformation, illumination variation, low resolution, fast motion and motion blur.



(a) IPR on OTB-2015 (b) OPR on OTB-2015 (c) SV on OTB-2015

(d) OV on OTB-2015 (e) OCC on OTB-2015 (f) BC on OTB-2015

(g) DEF on OTB-2015 (h) IV on OTB-2013 (i) LR on OTB-2015

(j) FM on OTB-2013 (k) MB on OTB-2015

Figure 3: The success plots on OTB-2015 for eleven challenge attributes: in-plain rotation, out-of-plane rotation, scale variation, out of view, occlusion, background clutter, deformation, illumination variation, low resolution, fast motion and motion blur.

Figure 4: Qualitative evaluation of our RASNet tracker, CREST [2], SINT [3], SiamFC [1] and CFNet [4] on ten challenging sequences (from top to down: *clifBar*, *freeman3*, *car1*, *jump*, *dragonBaby*, *bird1*, *motorRolling*, *carScale*, *ironman*, and *matrix*, respectively). Our RASNet tracker performs favorably against the state-of-the-art trackers.