Supplementary Material for Correlation-Guided Attention for Corner Detection Based Visual Tracking

Fei Du Peng Liu Wei Zhao Xianglong Tang
School of Computer Science and Technology, Harbin Institute of Technology, China
{feiaxyt, pengliu, zhaowei, tangxl}@hit.edu.cn

1. Network architecture details

Network backbone: We use ResNet-50 [3] as our backbone network and extract generic feature representations from the last layer of the *conv4* block. Table 1 illustrates the details of our backbone. ResNet-50 is modified to have a total stride of 8 pixels, and the respective field is increased by dilated convolutions [8]. An extra 1×1 convolutional layer is added to reduce the feature channel to 256. For both training and tracking, the template image and the test image are cropped to sizes 127×127 and 255×255, respectively. The resolutions of the template features and test image features are 15×15 and 31×31, respectively. As in [4], we crop the center 7×7 regions as the template features for the Siamese tracking module. A PrPool layer with spatial output size 7×7 is used to extract features for these RoIs. Features for the target template are extracted by a 5×5 PrPool layer.

Siamese tracking module: The architecture of the Siamese tracking module in our method is shown in figure 1. The similarity map and the distance offset map are respectively obtained in the similarity learning branch and the regression branch.

Spatial attention network: An hourglass network is used in the pixel-correlation guided spatial attention module to obtain spatial attention. The architecture is illustrated in table 2.

Channel attention network: A shared MLP is used to capture channel-wise dependencies in the channel-wise correlation guided channel attention module. The architecture is illustrated in table 3.

Corner detection network: Two upsample networks are used to learn heatmaps for the two corners. The architecture of this network is illustrated in table 4. Two nearest interpolation layers are employed to increase the resolution. In order to balance efficiency and accuracy, we increase the resolution to 31×31.

layer	template output size	test output size	structure
conv1	61×61	125×125	7×7, 64, stride 2
conv2_x	31×31	63×63	3×3 max pool, stride 2 $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	15×15	31×31	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
conv4_x	15×15	31×31	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
adjust	15×15	31×31	$1 \times 1, 256$

Table 1: Architecture of the backbone network.

2. More Ablation Experiments

Table 5 shows more results and complexity comparison of different integration methods. These methods are evaluated on the combined OTB2015 and UAV123 datasets. In addition to the integration methods described in the paper, we add two ablative methods named as *OursMaxPool* and *OursAvgPool*. They respectively denote that we only use a global max-pooling layer and a global average-pooling layer to gather channel descriptors in the channel attention module. The results show that we can achieve the best performance by jointly using max-pooling and average-pooling layers.

In our method, the second stage can be viewed as a bounding box refinement process. Thus, we compare our method with traditional bounding box regression model (denoted as *BBoxReg*). As in [7], *BBoxReg* concatenates the features of the template and the RoI and uses the fully connected layers to predict the refined bounding boxes. Note that *BBoxReg* does not perform similarity learning in the second stage, which is the same as our method. Table 6 shows that *BBoxReg* improves the baseline method, while our method achieves better performance. We attribute this to the proposed correlation-guided attention module, which effectively exploits the relationship between the template

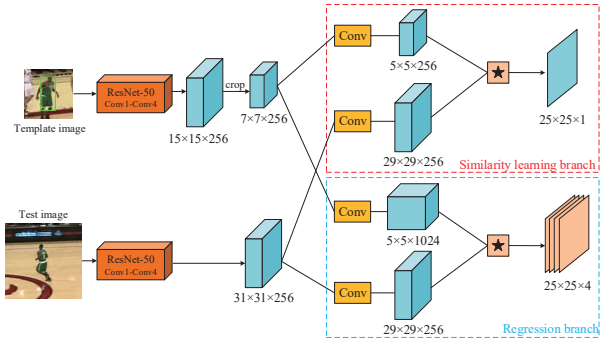

Figure 1: Architecture of the Siamese tracking module.

layer	output size	structure
	25 × 7 × 7	
conv1	32 × 5 × 5	3 × 3, 32 BatchNorm ReLU
conv2	32 × 3 × 3	3 × 3, 32 BatchNorm ReLU
dconv1	32 × 5 × 5	3 × 3, 32 BatchNorm ReLU
dconv2	1 × 7 × 7	3 × 3, 1 Sigmoid

Table 2: Architecture of the spatial attention network.

layer	output size	structure
	256 × 1 × 1	
fc1	64 × 1 × 1	1 × 1, 64 ReLU
fc2	256 × 1 × 1	1 × 1, 256 Sigmoid

Table 3: Architecture of the channel-wise attention network.

and the RoI to improve the performance of corner detection.

3. Impact of Training Data

The GOT-10k dataset is used in our method since corner detection can benefit from high-quality training image pairs. We compare our method with the state-of-the-art Siamese tracker (SiamRPN++) without using the GOT-10k dataset for training. Thus, our method has the same training datasets with SiamRPN++. Table 7 shows the comparison results on two challenging datasets, UAV123 and LaSOT. Without training with GOT-10k, our method achieves better performance compared to SiamRPN++. The results show that our method achieves state-of-the-art performance.

layer	output size	structure
	256 × 7 × 7	
conv1	256 × 7 × 7	3 × 3, 256, padding=1 BatchNorm ReLU
conv2	64 × 7 × 7	1 × 1, 64 BatchNorm ReLU
interp1	64 × 15 × 15	Nearest interpolation
conv3	64 × 15 × 15	3 × 3, 64, padding=1 BatchNorm ReLU
conv4	32 × 15 × 15	1 × 1, 32 BatchNorm ReLU
interp2	32 × 31 × 31	Nearest interpolation
conv5	32 × 31 × 31	3 × 3, 32, padding=1 BatchNorm ReLU
conv6	1 × 31 × 31	1 × 1, 1

Table 4: Architecture of the corner detection network.

4. Qualitative results

Figure 2 illustrates an example of the pixel-wise correlation similarity maps. As can be seen, different parts of the target are highlighted in different maps, and the outline of the target is encoded in the entire set of similarity maps.

Figure 3 shows two examples of the spatial attention maps and the detection results based on the heatmaps for the corners. As the resolution of the spatial attention maps is relatively small, we visualize the maps by resizing them to an appropriate size. As can be seen, the spatial attention maps highlight the top-left and bottom-right regions of the target, where there is rich information for detecting the top-left and bottom-right corners. The heatmaps for the top-left and bottom-right corners are shown in one image, and the detection results are obtained from the heatmaps using the soft-argmax function [5].

Figure 4 shows the tracking results of the first stage and the second stage of the proposed method. As can be seen, the corner detection module significantly improves the quality of bounding box estimation in the first stage. Note that the tracking result of the first stage in one frame is based on the result of the second stage in previous frame. Without the second stage, the Siamese tracker drifts easily due to the accumulating errors.

Figure 6 shows some tracking results on eight challenging sequences. Our CGACD method is compared with state-of-the-art trackers including SiamRPN++ [4], ATOM [2], MDNet [6], and ECO [1].

5. Attribute-Based Results on OTB2015

Figure 5 shows the success plots of 10 trackers on 11 attributes of OTB2015. These attributes include variation

Method	PS (%)	AUC (%)	#Params	MFLOPs
W/o template	81.8	62.0	1.31M	190
ConcatInte	83.1	62.5	2.49M	306
SiamInte	85.3	64.9	1.31M	190
ConcatAtt	84.0	63.7	1.69M	207
SiamAtt	85.2	65.1	1.51M	194
ChannelAtt	83.8	63.5	1.34M	191
SpatialAtt	85.4	65.3	1.36M	193
OursMaxPool	86.3	66.1	1.40M	193
OursAvgPool	85.8	65.6	1.40M	193
Ours	**86.7**	**66.3**	1.40M	193

Table 5: Comparison of different approaches to integrating the template and the RoI on the combined OTB2015 and UAV123 datasets.

	Baseline	BBoxReg	Ours
PS (%)	82.7	85.6	**87.2**
AUC (%)	61.5	64.4	**66.8**

Table 6: Comparison results of bounding box regression model and our method.

Trackers	UAV123		LaSOT	
	PS (%)	AUC (%)	PS_{norm} (%)	AUC (%)
SiamRPN++	80.3	61.0	56.9	49.6
Ours	82.6	61.5	60.9	50.3

Table 7: Comparison results of SiamRPN++ and our method using the same training datasets.

(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 clutter (BC), and low resolution (LR). Our C-GACD achieves high performance on all the attitudes.

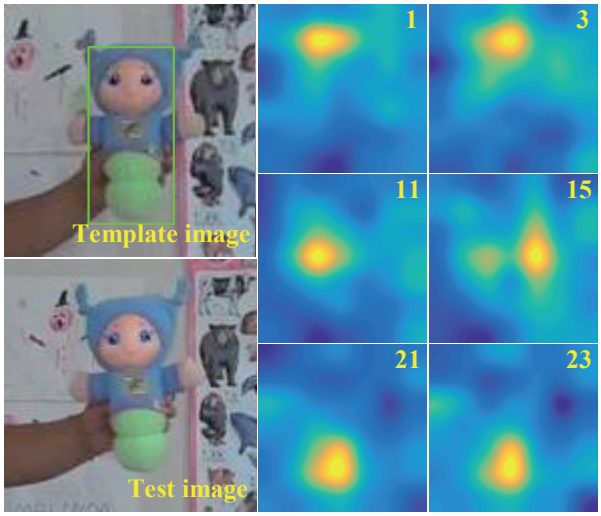

Figure 2: Visualization of the pixel-wise correlation similarity maps. 25 similarity maps are obtained by calculating the pixel-wise correlation between the template and the RoI feature maps. Each map represents the similarity between the corresponding pixel in the template feature maps and all pixels in the RoI feature maps. Different parts of the target are highlighted in different maps.

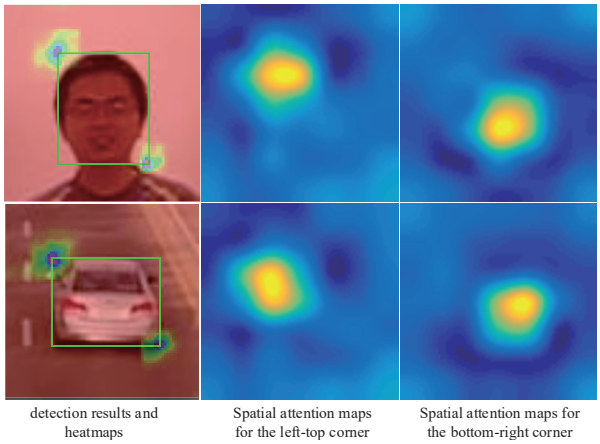

Figure 3: Visualization of spatial attention maps and the detection results based on the heatmaps for the corners.

References

[1] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: efficient convolution operators for tracking. In *CVPR*, 2017. 2

[2] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: accurate tracking by overlap maximization. In *CVPR*, 2019. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[4] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. *arXiv preprint arXiv: 1812.11703*, 2018. 1, 2

[5] Diogo C. Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *arXiv preprint arXiv: 1710.02322*, 2017. 2

[6] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, 2016. 2

[7] Guangting Wang, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. SPM-Tracker: series-parallel matching for real-time visual object tracking. In *CVPR*, 2019. 1

[8] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 1

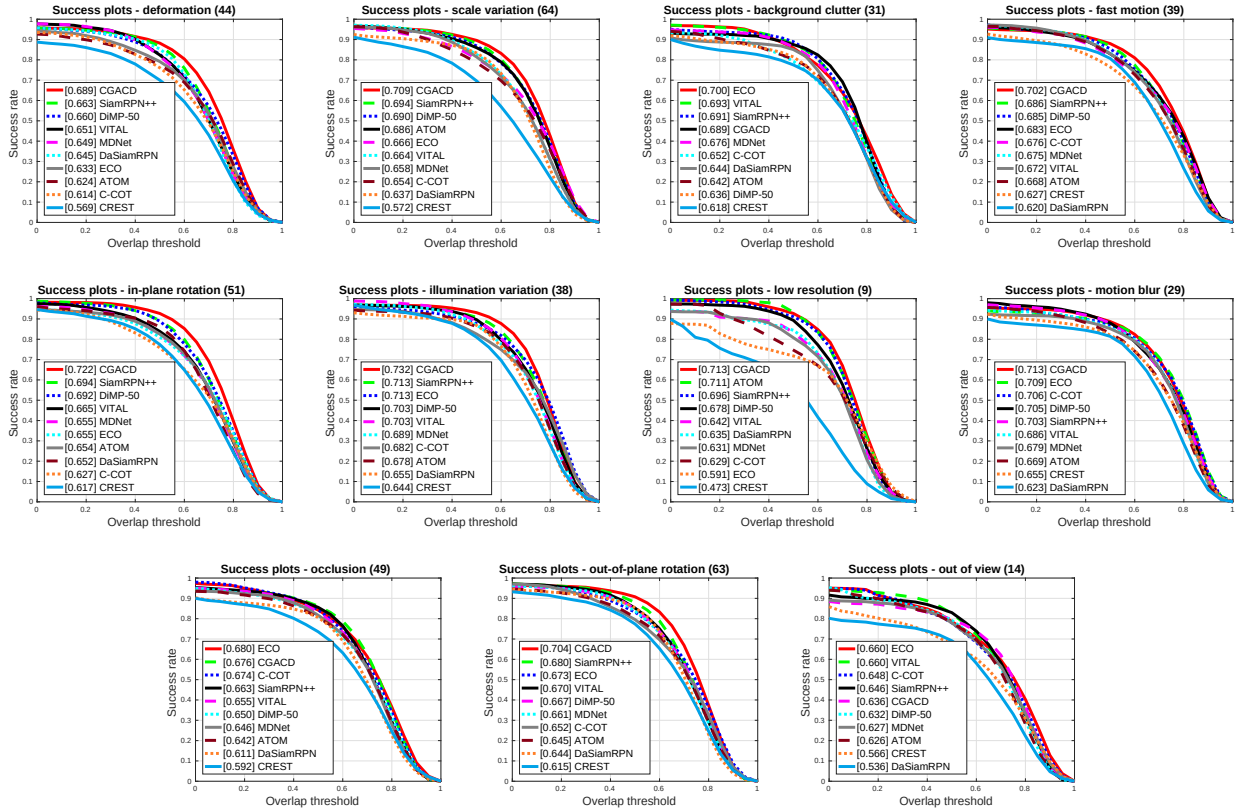

Figure 5: The success plots on 11 attributes of OTB2015

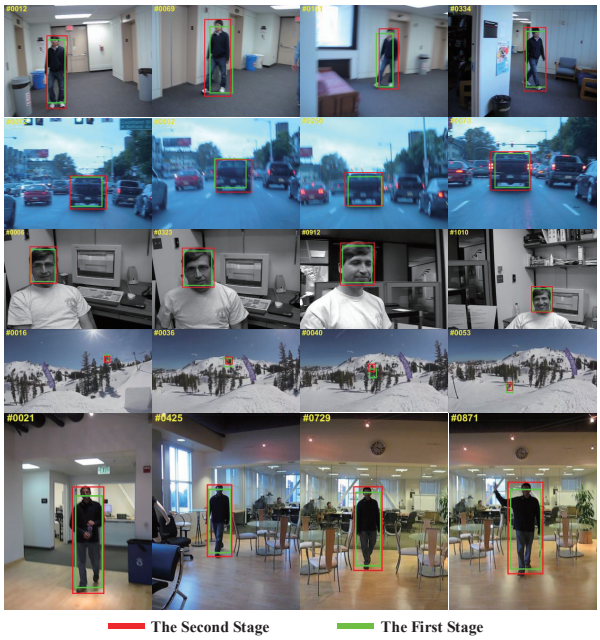

Figure 4: Qualitative results of the first stage and the second stage of the proposed method.

Figure 6: Qualitative comparison of state-of-the-art trackers on OTB2015.