Supplementary Material

PG-Net: Pixel to Global Matching Network for Visual Tracking

Bingyan Liao, Chenye Wang, Yayun Wang, Yaonong Wang, and Jun Yin

A. Network architecture

The detailed architecture of backbone network is shown in Table 1. Backbone network has the same structure with Resnet50 before Block2. In Block3 and Block4, we replace the down-sample operation with a series of dilated convolutions. The dilations are set to 2 and 4 respectively in Block3 and Block4. Table 2 presents the detailed architecture of target location network. It is a fully convolutional network for calculating similarity and locating the target.

	Layers			Number
Input	Search: 255	template: 127		
	Conv: k7 s2 n64 p0	BN	ReLU	× 1
	Maxpool: k2 s2 p1	——	——	× 1
Size	Search: 63	template: 31		
Block 1	Conv: k1 s1 n64 p0	BN	ReLU	× 3
	Conv: k3 s1 n64 p1	BN	ReLU	
	Conv: k1 s1 n256 p0	BN	——	
	Adjusted Sum	——	ReLU	
Size	Search: 63	template: 31		
Block 2	Conv: k1 s1 n128 p0	BN	ReLU	× 4
	Conv: k3 s2/1/1/1 n128 p1	BN	ReLU	
	Conv: k1 s1 n256 p0	BN	——	
	Adjusted Sum	——	ReLU	
Size	Search: 31	template: 15		
Block 3	Conv: k1 s1 n256 p0	BN	ReLU	× 6
	Conv: k3 s1 n256 d1/2/2/2/2/2 p1/2/2/2/2/2	BN	ReLU	
	Conv: k1 s1 n1024 p0	BN	——	
	Adjusted Sum	——	ReLU	
Size	Search:31	template: 15		
Block 4	Conv: k1 s1 n512 p0	BN	ReLU	× 3
	Conv: k3 s1 n512 d2/4/4 P2/4/4	BN	ReLU	
	Conv: k1 s1 n2048 p0	BN	——	
	Adjusted Sum	——	ReLU	
Size	Search:31	template： 15		

Table 1. Backbone network architecture. Where *k, s, n, d, p* denote kernel size, stride, convolutional number, dilation and padding respectively. The parameter *d* is omitted when it is 1 before Block2. *Adjusted Sum* denotes element-wise sum operation between input and output of each residual block. A convolution layer will be used to adjust the number of channel in element-wise sum process if necessary.

<table>
<tr><th></th><th colspan="4">Layers</th><th>Number</th></tr>
<tr><td rowspan="2">Adjusted layer</td><td rowspan="2">Search Feature1/2/3</td><td>——</td><td rowspan="2">Template Feature1/2/3</td><td>Crop to: 7x7</td><td rowspan="6">× 3</td></tr>
<tr><td>Conv: k3 s1 n256 p0
BN
ReLU</td><td>Conv: k3 s1 n256 p0
BN
ReLU</td></tr>
<tr><td></td><td colspan="4">PG-Corr</td></tr>
<tr><td></td><td colspan="4">Conv: k3 s1 n256 p0 + BN + ReLU</td></tr>
<tr><td></td><td colspan="4">Conv: k3 s1 n256 p0 + BN + ReLU</td></tr>
<tr><td rowspan="3">Detection head</td><td rowspan="3">Cls1/2/3</td><td>Conv: k3 s1 n256 p1
BN
ReLU</td><td rowspan="3">Reg1/2/3</td><td>Conv: k3 s1 n 256 p1
BN
ReLU</td></tr>
<tr><td>Conv: k1 s1 n10 p1</td><td>Conv: k1 s1 n20 p1</td></tr>
<tr><td>Cls1/2/3</td><td>Reg1/2/3</td></tr>
<tr><td rowspan="2">Enhanced detection</td><td colspan="2">Weighted Sum: Cls1/2/3</td><td colspan="2">Weighted Sum: Reg1/2/3</td><td rowspan="2">× 1</td></tr>
<tr><td colspan="2">Conv: k3 s1 n10 p1</td><td colspan="2">Conv: k3 s1 n20 p1</td></tr>
<tr><td>Output</td><td colspan="2">Cls</td><td colspan="2">Reg</td><td></td></tr>
</table>

Table 2. Target location network architecture, Where *k, s, n, p* denote kernel size, stride, convolutional number and padding respectively. The weights in Weighted sum operation are learned from training data.

B. Tracking results on LaSOT

We present some qualitative results of the proposed method and compare them with other state-of-the-art trackers on LaSOT dataset, as shown in Figure 1. It can be observed that our method has outstanding performance.

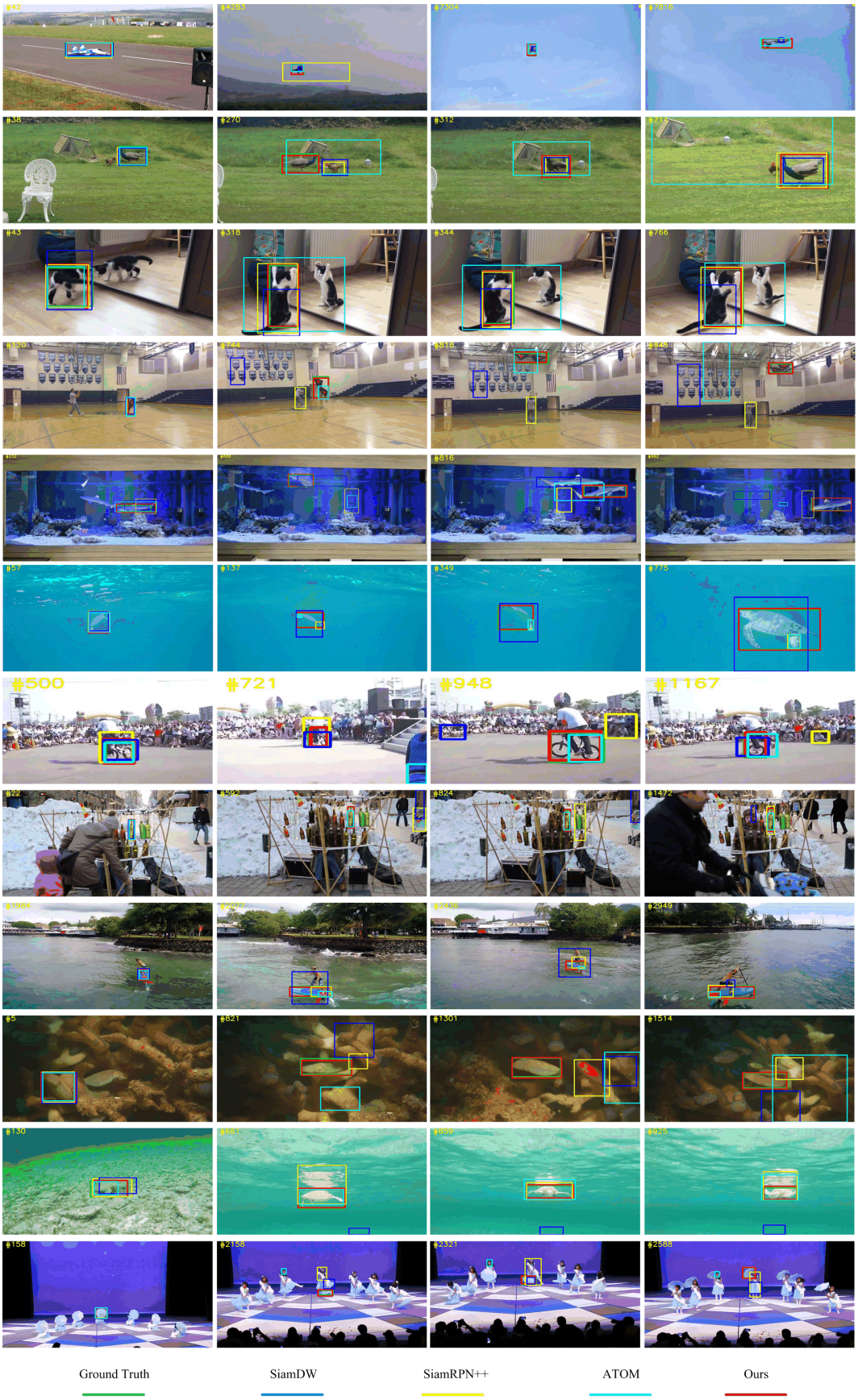

Figure 1. Qualitative comparison on ten challenging sequences (from top to bottom: *airplane-9*, *bird-3*, *cat-20*, *kite-6*, *shark-2*, *turtle-16*, *bicycle-2*, *bottle-12*, *surfboard-5*, *sepia-13*, *turtle-9*, *umbrella-19*) of LaSOT.