Ocean: Object-aware Anchor-free Tracking

Supplementary Material

The supplementary material presents the details of target localization in Sec. 4 and the additional experiments of Sec. 5:
1) We provide details of scale penalty in target localization.
1) We provide additional ablation experiments.
2) We provide qualitative comparisons of our tracker with state-of-the-arts.

Target localization. We impose a penalty on scale change to suppress the large variation of object size and aspect ratio.

$$\alpha = exp(k \cdot max(\frac{r}{r'}, \frac{r'}{r}) \cdot max(\frac{s}{s'}, \frac{s'}{s})), \tag{1}$$

where k is a hyper-parameter, r and r' represent the aspect ratio of the predicted bounding boxes in the previous and current frames respectively, while s and s' denote the size (*i.e.*, height and width) of the predicted boxes in the previous and current frames. The final target classification probability $\hat{p}_{cls}$ is calculated as $\hat{p}_{cls} = \alpha \cdot p_{cls}$. The maximum value in the classification map $\hat{p}_{cls}$ indicates the position of the foreground target. To keep the shape of predicted bounding boxes changing smoothly, a linear weight function is used to calculate the final scale as $\hat{s}_{reg} = \beta \cdot s' + (1 - \beta) \cdot s$, where β is a weight parameter. The hyperparameter k in Eq. (1) for the penalty of large scale change is set to 0.021, while the scale weight β is set to 0.7.

Feature combination. We further evaluate the impact of dilated convolutions in the feature combination module and report the results on VOT-2018 in Tab. 1. The baseline setting is a normal convolution with dilation stride of 1 along both the X-axis and Y-axis, *i.e.*, Φ_{11}. We observe that adding a standard convolution Φ_{11} brings an improvement of 0.8 points in terms of EAO (② *vs.* ①). This indicates that the proposed parallel convolutions in the feature combination module is effective. It is very interesting to see that if we modify the dilation strides along X and Y directions to be different, the performance can be further improved, *e.g.*, 1.3 points gains for ③ *vs.* ② while 1.0 points gains for ④ *vs.* ②. This verifies that the irregular dilations is effective to enhance feature representability. A combination of the three dilation kernels with different strides obtains the best results in our experiment.

Feature visualization. We visualize the features extracted before and after the alignment, *i.e.*, the regular-region feature and object-aware feature, in Fig. 1. We observe that the object-aware feature focuses on the entire object, while the regular-region feature concentrates on the center part of the target. The former improves the reliability of the classification since it provides a global view of the target. The latter contributes more to localize the object centerness since the features are more sensitive to local changes.

#Num	Dilated Kernels	EAO
①	Φ_{11}	0.425
②	$\Phi_{11}\Phi_{11}$	0.433
③	$\Phi_{11}\Phi_{12}$	0.446
④	$\Phi_{11}\Phi_{21}$	0.443
⑤	$\Phi_{11}\Phi_{12}\Phi_{21}$	0.467

Table 1. Analysis of the impact of different strides over dilated convolution in the feature combination module.

Fig. 1. Visualization of (a) the regular-region feature and (b) the object-aware feature over the video "ants1".

Impact of the convolution layers in the anchor-free networks. For both of the regression network and regular-region classification network (*i.e.*, "Conv" of the anchor-free networks in the Fig. 3 of the main paper, we use four convolution layers to predict the fine-grained regression and classification score maps. To study the impact of the number of convolution layers in these two networks, we perform an ablation experiment on OTB-100 [5], as shown in Tab. 2. We can see that as the number of convolution layers increases, the performance (*i.e.*, AUC) becomes saturated. The balanced choice between performance and speed is 3 or 4 in our model.

The number of layers	0	1	2	3	4	5
Performance (AUC ↑)	0.645	0.657	0.665	0.672	0.673	0.670

Table 2. Impact of the number of layers in the anchor-free networks.

Analysis of rectification capacity. The rectification capacity indicates the prediction accuracy ($mIoU$) of the model when the tracker drifts from the target. To evaluate the rectification capacity, we first sample exemplar and search image pairs from adjacent frames in the VOT-2018 dataset [3]. We sample locations on the score maps away from the target to simulate weak predictions, and the shifting magnitude is illustrated in the first row of Tab. 3. Then we compute the overlap between the predicted bounding box and the groundtruth, *i.e.*, $mIoU$ in Tab. 3. Larger $mIoU$ means that the regression network can better rectify the inaccurate prediction. We can see that the performance ($mIoU$) of the proposed model outperforms SiamRPN++ [4] when the tracker drifts away from the target's center. This demonstrates the superior robustness of our tracker compared to the anchor-based method.

Distance (*pixels*)	8	16	24	32	40	48	56
$mIoU$ of SiamRPN++ [4]	0.65	0.48	0.36	0.30	0.28	0.25	0.21
$mIoU$ of our tracker	0.73	0.73	0.72	0.71	0.71	0.72	0.54

Table 3. Comparisons of rectification capacity between SiamRPN++ and our model. $mIoU$ indicates mean IoU. *Distance* indicates shifting magnitude (ℓ_1 distance) for generating search images.

Qualitative comparisons. Fig. 2 qualitatively compares the results of recent top-performing trackers: DiMP [1], ATOM [2], SiamRPN++ [4] and our Ocean tracker on 6 challenging sequences. SiamRPN++ drifts from the target when fast motion occurs (*soccer*). The underlying reason is that fast motion results in imprecise prediction, which is difficult to be rectified by the anchor-based method. By contrast, our model performs better in this case, since it is capable of rectifying inaccurate bounding boxes and robust to noisy predictions. This verifies the advancement of our anchor-free regression mechanism compared to anchor-based methods. When the target undergoes large deformation or rotation, *e.g.*, *dinasour* and *motocross1*, the predicted locations of ATOM and DiMP are not accurate enough. One of the reasons is that the regular-region sampling strategy in these approaches may lack global information to generate discriminative appearance features. Benefiting from the object-aware feature, our model can predict better results in this case.

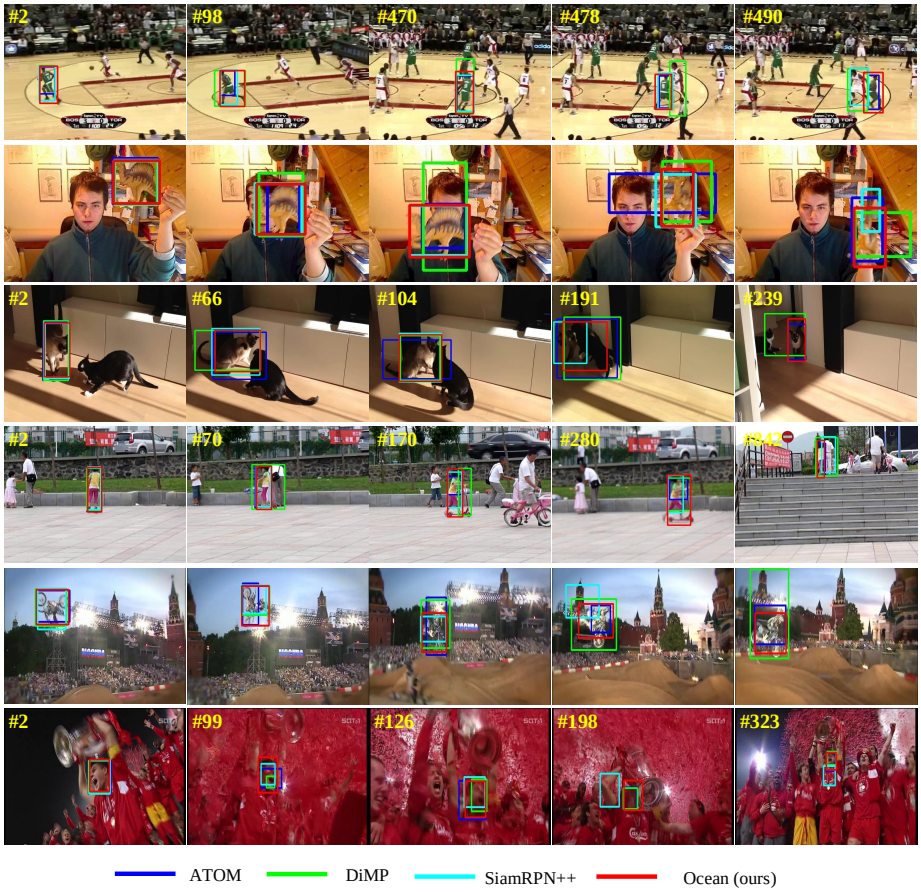

Fig. 2. Visual comparisons of our tracker with statr-of-the-art trackers on 6 video sequences: *basketball, dinasour, fernando, girl, motocross1 and soccer.*

References

1. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: ICCV. pp. 6182–6191 (2019)
2. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: CVPR. pp. 4660–4669 (2019)
3. Kristan, M., Leonardis, A., Matas, et al.: The sixth visual object tracking vot2018 challenge results. In: ECCVW. pp. 1–52 (2018)
4. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: SiamRPN++: Evolution of siamese visual tracking with very deep networks. In: CVPR. pp. 4282–4291 (2019)
5. Wu, Y., Lim, et al.: Object tracking benchmark. TPAMI **37**(9), 1834–1848 (2015)