Supplementary Material

ATOM: Accurate Tracking by Overlap Maximization

Martin Danelljan[*,1,2] Goutam Bhat[*,1,2] Fahad Shahbaz Khan[1,3] Michael Felsberg[1]

[1] CVL, Linköping University, Sweden [2] CVL, ETH Zürich, Switzerland [3] Inception Institute of Artificial Intelligence, UAE

In this supplementary material we provide additional details and results. Section 1 provides details about the other network architectures evaluated for IoU prediction in section 4.1 of the main paper. Section 2 performs an empirical convergence analysis of the employed optimization procedure and the gradient descent. Detailed results on the LaSOT [4] dataset are provided in section 3. Section 4 provides results on the OTB-100 [12] dataset. The impact of the training data on performance of our tracker is analyzed in section 5. Section 6 provides detailed results on the UAV123 dataset. A video showing qualitative results of our tracker can be found at https://youtu.be/T8x8i1KkYGk.

1. Network Architectures for IoU Prediction

Here we describe the different network architectures for integrating the target appearance, investigated in section 4.1 of the main paper. Figure 1 visualizes the **Concatenation** architecture. In this architecture, both the reference and test branches have the same network structure. ResNet-18 `Block3` and `Block4` features that are extracted from the reference and test images are passed through two `Conv` layers, followed by `PrPool` and an `FC` layer. The processed features from both the ResNet blocks and both the images are concatenated and passed through a final `FC` layer which predicts the IoU. Note that due to the symmetric structure of the network, the weights for the `Conv` layers before `PrPool` are shared between the reference branch and the test branch. However the `FC` layers do not share the weights.

Figure 2 visualizes the **Siamese** architecture. Similar to **Concatenation**, both the reference and test branches have the same network structure. ResNet-18 `Block3` and `Block4` features that are extracted from the reference and test images are passed through two `Conv` layers, followed by `PrPool` and an `FC` layer. The processed features from both the ResNet blocks are then concatenated. The IoU prediction is obtained as the dot product of the features from the reference and the test branches. The `Conv` layers before `PrPool` have shared weights. In the final FC layer however, we found it beneficial not to share the weights between the branches.

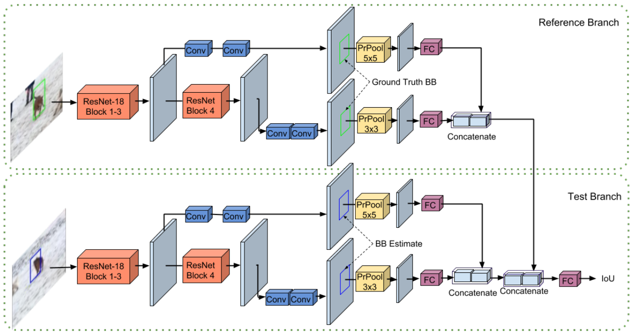


Figure 1. Architecture of the **Concatenation** network for IoU prediction evaluated in section 4.1 in the paper.

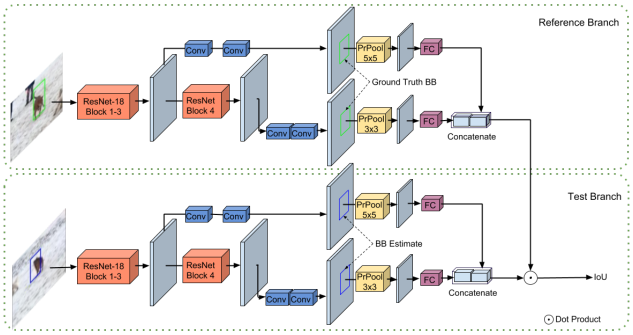


Figure 2. Architecture of the **Siamese** network for IoU prediction evaluated in section 4.1 in the paper.

2. Convergence Analysis

We empirically compare of the convergence speed of the employed optimization method (algorithm 1 in the paper) and Gradient Descent (GD). This is performed by comparing the loss for the online learning problem eq. (3), which is minimized in the first frame w.r.t. the filter weights w_1 and w_2. For our method, we use the settings described in the paper. In case of Gradient Descent we employ the same

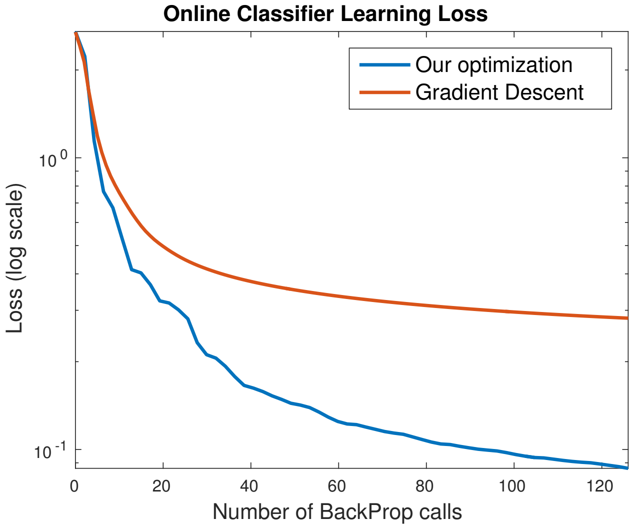


Figure 3. Comparison of convergence speed between our employed online optimization procedure and Gradient Descent. We plot the loss of the online classifier learning (eq. (3) in the paper) w.r.t. the number of performed `BackProp` iterations. The loss is averaged over five independent runs of the complete NFS dataset. The employed method achieves much faster convergence.

settings used in the ablation study (section 4.2 in the paper).

In figure 3 we plot the loss (eq. (3) in the paper) for each method. For a fair comparison, the loss is plotted w.r.t. the number of `BackProp` calls performed by each method. The loss in figure 3 is computed as an average of five complete runs over the full NFS dataset [5]. Our CG-based optimization algorithm exhibits superior convergence speed compared to Gradient Descent. Moreover, the employed optimization methods does not require tuning of the step length and momentum parameters.

3. Detailed results on LaSOT dataset

In table 4 in the main paper, we provide a state-of-the-art comparison on the large-scale LaSOT dataset in terms of normalized precision and success. Here, we provide the success plot for the same. The success plots are obtained using the overlap precision (OP) score, which is computed as the percentage of frames in the dataset for which the intersection-over-union (IoU) overlap between the tracker prediction and the ground truth bounding box is higher than a certain threshold. The OP scores are plotted for a range of thresholds in $[0, 1]$ to obtain the success plot. The area under this plot gives the AUC (success) score, which is reported in the legend. Figure 4 shows the success plot over the 280 test videos. Our approach ATOM significantly outperforms the previous best approach DaSiamRPN [14] with an absolute gain of 10.0% in AUC score.

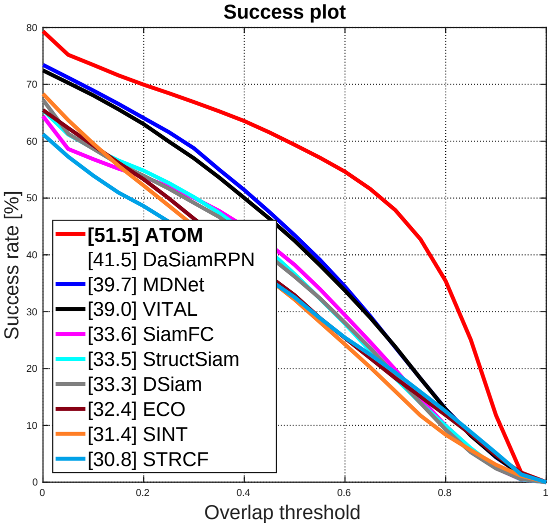


Figure 4. Success plot on the LaSOT dataset. Note that due to the unavailability of raw results for DaSiamRPN, we only report the final AUC score in the legend. Our approach ATOM outperforms all previous methods by a large margin.

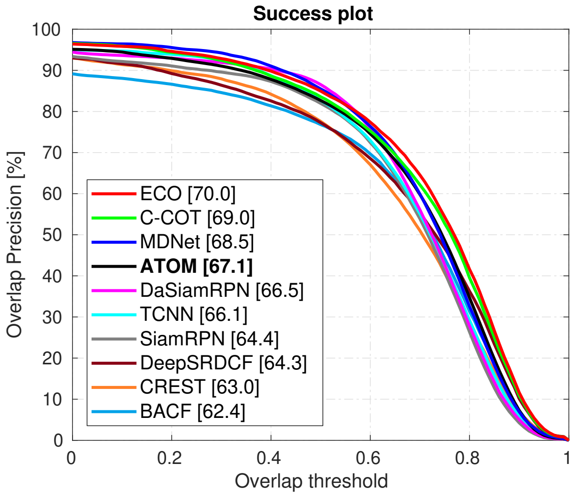


Figure 5. State-of-the-art comparison on the OTB-100 dataset. Our approach obtains results competitive with the state-of-the-art approaches.

4. Results on OTB-100 dataset

Here, we compare our approach with the state-of-the-art trackers on the OTB-100 [12] dataset. The success plot over all the 100 videos are shown in figure 5. Our approach achieves results competitive with the state-of-the-art approaches, with an AUC score of 67.1%. Note that the best results are obtained by the correlation filter based methods, ECO [2] and CCOT [3]. These methods employ brute-force multi-scale search for target estimation. Since OTB-100 has limited changes in aspect ratio (see figure 2 in [7]), the fixed aspect ratio constraint in multi-scale search strategy helps these methods to obtain a better accuracy.

	SINT	ECO	DSiam	StructSiam	SiamFC	VITAL	MDNet	DaSiamRPN	**ATOM-VID**	**ATOM**
Norm. Prec. (%)	35.4	33.8	40.5	41.8	42.0	45.3	46.0	49.6	**55.0**	**57.6**
Success (%)	31.4	32.4	33.3	33.5	33.6	39.0	39.7	41.5	**49.5**	**51.5**

Table 1. Comparision of our approach trained using only ImageNet-VID (denoted ATOM-VID) on the LaSOT dataset. Our approach, trained using considerably less data as compared to the previous best approach DaSiamRPN, significantly outperforms it with an absolute gain of 8.0% in AUC score.

	Staple	SAMF	CSRDCF	ECO	SiamFC	CFNet	MDNet	UPDT	DaSiamRPN	**ATOM-VID**	**ATOM**
Precision (%)	47.0	47.7	48.0	49.2	53.3	53.3	56.5	55.7	59.1	**61.8**	**64.8**
Norm. Prec. (%)	60.3	59.8	62.2	61.8	66.6	65.4	70.5	70.2	73.3	**74.6**	**77.1**
Success (%)	52.8	50.4	53.4	55.4	57.1	57.8	60.6	61.1	63.8	**69.8**	**70.3**

Table 2. Comparision of our approach trained using only ImageNet-VID (denoted ATOM-VID) on the TrackingNet dataset.

5. Impact of training data

In this section, we investigate the impact of using recent large-scale tracking datasets for offline training of our IoU predictor network. We train our network using only the ImageNet-VID [10] dataset, that has been commonly used to train trackers [1, 11, 13] in recent years. We compare this version, denoted ATOM-VID, with the state-of-the-art approaches on two recent datasets, namely LaSOT [4] and TrackingNet [8]. For comparision, we also include our final version ATOM, trained using the train splits of LaSOT, TrackingNet and COCO [6]. Results are shown in table 1 for LaSOT and table 2 for TrackingNet, respectively. Among previous approaches, DaSiamRPN [14] uses bounding box regression strategy and achieves the best results on both datasets. Note that DaSiamRPN is trained using the large-scale YoutubeBB [9], ImageNet-VID, COCO and ImageNet DET [10] datasets. Our approach ATOM-VID, trained using only ImageNet-VID, significantly outperforms DaSiamRPN with an absolute gain of 8.0% in AUC score on LaSOT, and 6.0% in AUC score on TrackingNet. Using the recent tracking datasets for training further improves the results, providing an absolute gain of 2.0% on LaSOT and 0.5% on TrackingNet. While using a larger training set improves the tracking performance as expected, our approach still achieves state-of-the-art results when using less data compared to recent methods.

6. Additional Results on UAV123

Here, we provide detailed results on the UAV123 dataset [7]. In UAV123, each video is annotated with 12 different attributes: aspect ratio change, background clutter, camera motion, fast motion, full occlusion, illumination variation, low resolution, out-of-view, partial occlusion, scale variation, similar objects, and viewpoint change. Figure 6 shows the success plots for all the attributes. Our approach obtains the best results on all 12 attributes. Thanks to our target estimation module, our approach excels in case of aspect ratio change, scale variation, and viewpoint change. Furthermore, due to our robust online-learned classifier, our tracker also outperforms previous methods in case of similar objects, illumination variation, partial occlusion, and low resolution.

References

[1] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *ECCV workshop*, 2016. 3

[2] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg. ECO: efficient convolution operators for tracking. In *CVPR*, 2017. 2

[3] M. Danelljan, A. Robinson, F. Khan, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *ECCV*, 2016. 2

[4] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019. 1, 3

[5] H. K. Galoogahi, A. Fagg, C. Huang, D. Ramanan, and S. Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *ICCV*, 2017. 2

[6] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 3

[7] M. Mueller, N. Smith, and B. Ghanem. A benchmark and simulator for uav tracking. In *ECCV*, 2016. 2, 3

[8] M. Müller, A. Bibi, S. Giancola, S. Al-Subaihi, and B. Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, 2018. 3

[9] 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. *CVPR*, 2017. 3

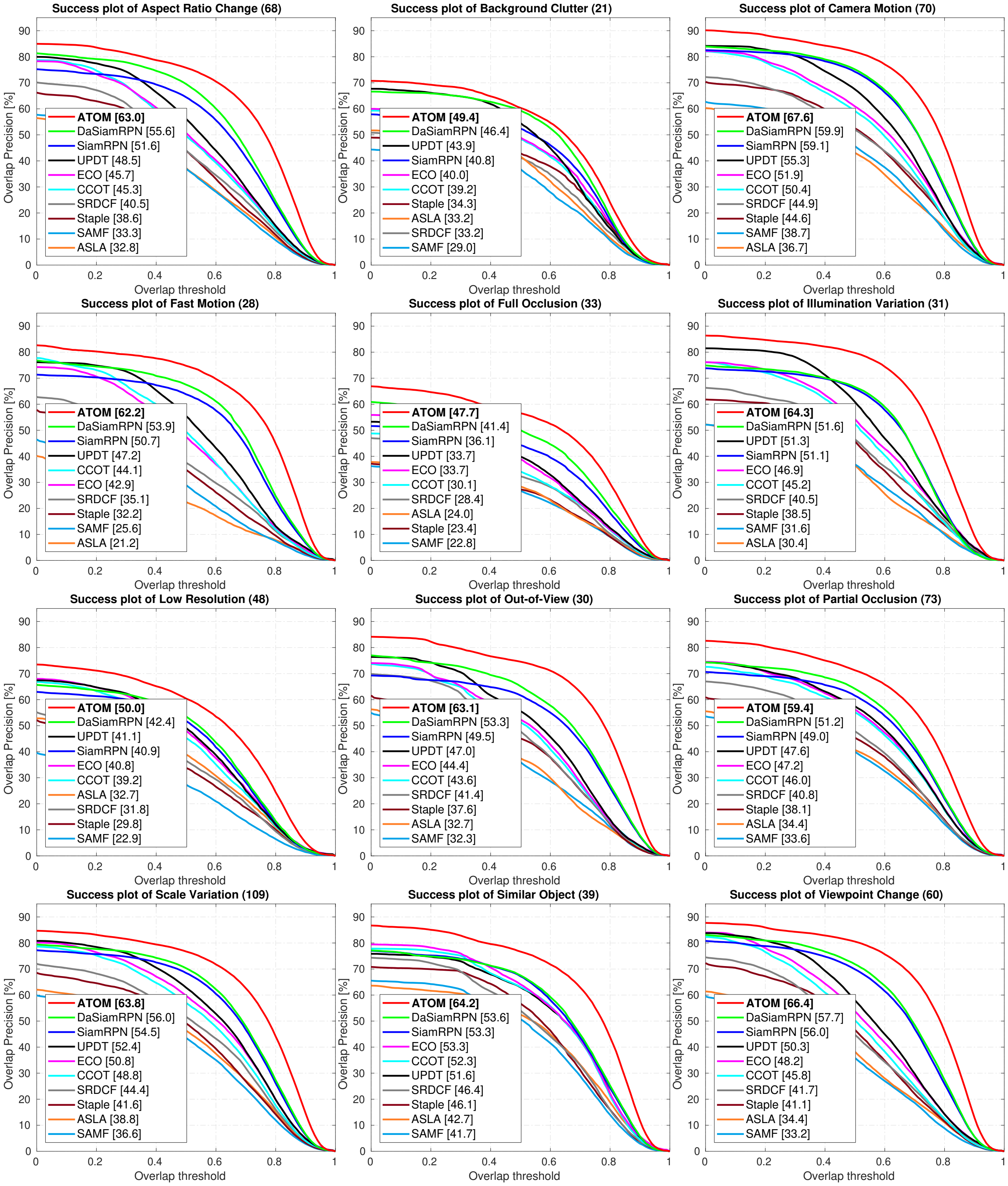
[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, pages 1–42, April 2015. 3

[11] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. S. Torr. End-to-end representation learning for correlation filter based tracking. In *CVPR*, 2017. 3

[12] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. *TPAMI*, 37(9):1834–1848, 2015. 1, 2

[13] Y. Yao, X. Wu, S. Shan, and W. Zuo. Joint representation and truncated inference learning for correlation filter based tracking. In *ECCV*, 2018. 3

[14] Z. Zhu, Q. Wang, L. Bo, W. Wu, J. Yan, and W. Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, 2018. 2, 3


Success plot of Aspect Ratio Change (68)

Success plot of Background Clutter (21)

Success plot of Camera Motion (70)

Success plot of Fast Motion (28)

Success plot of Full Occlusion (33)

Success plot of Illumination Variation (31)

Success plot of Low Resolution (48)

Success plot of Out-of-View (30)

Success plot of Partial Occlusion (73)

Success plot of Scale Variation (109)

Success plot of Similar Object (39)

Success plot of Viewpoint Change (60)

Figure 6. Attribute analysis on the UAV123 dataset. Our approach **ATOM** obtains the best performance on all 12 attributes.