Transforming Model Prediction for Tracking - Supplementary

Christoph Mayer Martin Danelljan Goutam Bhat Matthieu Paul Danda Pani Paudel
Fisher Yu Luc Van Gool
Computer Vision Lab, D-ITET, ETH Zürich, Switzerland

Appendices

In this supplementary material, we first provide details about training, model architecture and inference in Sec. A. Further, we report visual results such as a comparison to state-of-the-art trackers, a comparison of different model predictors and failure cases of our tracker in Sec. B. Afterwards, we provide more detailed results of the experiments shown in the main paper in Sec. C.

A. Training, Architecture and Inference

First, we provide additional details about the training followed by a detailed description of the architectures employed and finally we provide further inference details.

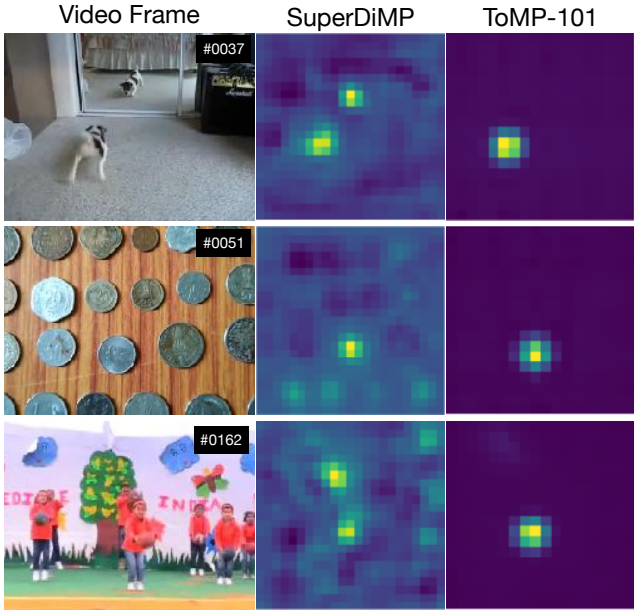


Figure 1. Visual comparison of the target score maps resulting from different model predictors.

A.1. Training and Architecture Details

For training we produce the target states y by using a Gaussian with standard deviation $1/4$ relative to the base target size and by settting $\tau = 0.05$ to differentiate between foreground and background regions in the corresponding classification loss l_{cls} adopted from DiMP [1]. For the model predictor we extract features with a stride of 16 from the third block of the ResNet that we use as backbone. We initialize the backbone with the official weights obtained by training the backbone on ImageNet [16] and freeze the batch norm statistics during training. Since we use a channel dimension of 256 for the Transformer and the ResNet features have 1024 channels we employ an single convolutional layer to decrease the number of channels before feeding the features into the Transformer Encoder. The Transformer Encoder consists of layers containing multi-headed self attention and a feed-forward network.

Two Stage Model Prediction	Previous Tracking Results	Confidence Threshold η	LaSOT	NFS	OTB
✓	✓	0.85	67.3	66.9	**70.3**
✓	✓	0.90	**67.6**	**66.9**	70.1
✓	✓	0.95	67.4	66.0	69.8

Table 1. Analysis of different inference settings an of their impact on the tracking performance in terms of AUC of the success curve.

training frames	NFS	OTB	UAV	LaSOT	LaSOTExtSub	Speed [FPS]
1 initial	65.3	67.8	68.7	65.7	43.7	**26.2**
1 initial + 1 recent	66.9	70.1	69.0	**67.6**	**45.4**	**24.8**
2 initial + 1 recent	**67.6**	**70.5**	67.2	**68.0**	**45.4**	20.5
1 initial + 2 recent	66.7	**70.8**	**69.4**	**67.6**	44.4	21.8
1 initial + 3 recent	66.8	**70.5**	**69.2**	**67.6**	44.2	17.6
1 initial + 4 recent	**67.2**	70.1	68.2	67.3	**44.7**	13.2
1 initial + 5 recent	66.8	70.1	69.1	67.2	43.9	11.3

Table 2. Comparison of different number of training samples in success AUC.

	$L_{\text{centerness}}$	NFS	OTB	UAV	LaSOT	LaSOTExtSub
Classification	✗	**66.9**	**70.1**	**69.0**	**67.6**	**45.4**
Classification	✓	**65.8**	**69.2**	67.3	**67.9**	**45.5**
Centerness	✓	62.7	66.3	67.4	64.4	41.3
Classification · Centerness	✓	63.7	67.8	**68.7**	65.8	45.3

Table 3. Impact of centerness scores on training and inference.

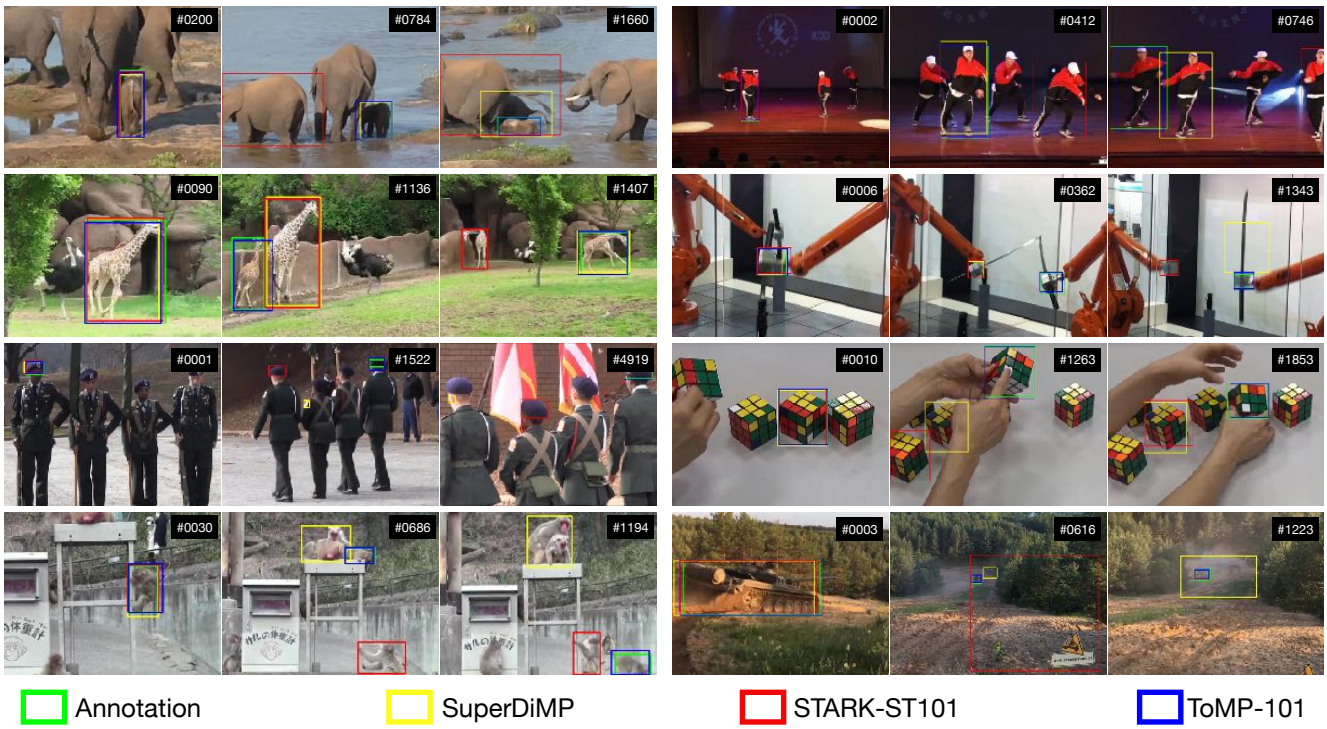


Figure 2. Visual comparison of different trackers (ToMP-101, SuperDiMP [11] and STARK-ST101 [48]) on different LaSOT [19] sequences.

We use eight heads and a hidden dimension of 2048 for the feed-forward network. Furthermore, we use Dropout with probability 0.1 and layer normalization. The Transformer settings are adopted from DETR [5]. The predicted target model weights for classification and bounding box regression consist of a single 1×1 filter with 256 channels. The bounding box regression CNN consists of four convolution-instance-normalization-ReLU layers and a final convolution layer, followed by an exponential activation. The MLP for target extent encoding ϕ consists of three layers ($4 \rightarrow 64 \rightarrow 256 \rightarrow 256$) where each layer consists of a linear projection, batch normalization and ReLU activation except the last that only consist of a linear projection. The region-encoding tokens e_{fg} and e_{test} are 256 dimensional learnable embeddings.

A.2. Inference Details

In order to decide whether a previous tracking result should be used for training of not we use the maximal value of the target score map produced by the target model. In particular, we select the sample if its confidence value is above a certain threshold η. Tab. 1 shows that the chosen threshold of 0.9 leads to high performance on LaSOT [19], NFS [22] and OTB-100 [46]. Furthermore, we follow SuperDiMP [11] and enter in the *target not found* state if the maximal value of the target score map is bellow 0.25. Moreover, we use the same spatial resolution of the target scores of 18×18 and the same search area scale factor of 5.0 during inference and training.

Furthermore, we study the effect of using more than two training frames stored in the sample memory. Instead of using only one initial and one recent training frame to predict the network weights we test the impact of increasing the number of recent training frames and of using multiple initial training frames. We increase the number of initial training frames with ground truth bounding box annotations using an augmentation (vertical flipping and random translation). Tab. 2 shows the results for different combinations of multiple initial and recent training frames. Note, that we use the same network weights for all experiments trained with one initial and one recent recent frame in all cases. We observer that using more training frames can improve the tracking performance but decreases the run-time. Furthermore, we observe that the tracker greatly benefits from including at least one recent frame for training.

A.3. Centerness

Our proposed bounding box regression component is inspired by FCOS [41] but in contrast to FCOS we omit an auxiliary centerness branch. The classification head of FCOS is trained to predict a high score for almost every region inside the bounding box. The centerness branch is

therefore needed to identify the center location of the object, used to select the bounding box offsets. In contrast, our classification branch is directly trained to accurately locate the object's center. The additional centerness branch is therefore redundant. Nonetheless, we train our best model with a centerness head and $L_{\text{centerness}}$ and report the results in Tab. 3 (2nd-4th rows). The 1st row shows the performance when omitting centerness for training. We achieve comparable results when using the model trained with centerness but applying only the classification scores to localize the target (2nd row). Using only the centerness scores decreases the performance (3rd row) because centerness often fails to identify the target among distractors (see Fig. 3). Finally,

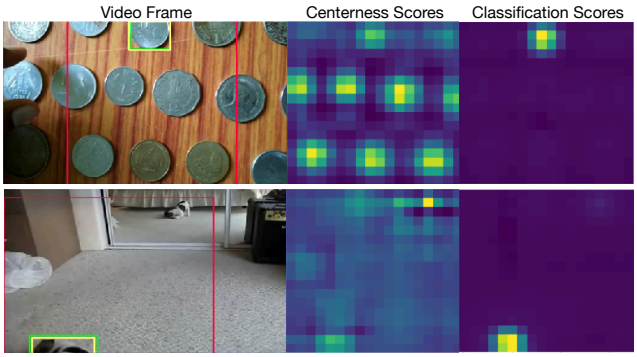


Figure 3. Visual Comparison between centerness and classification scores.

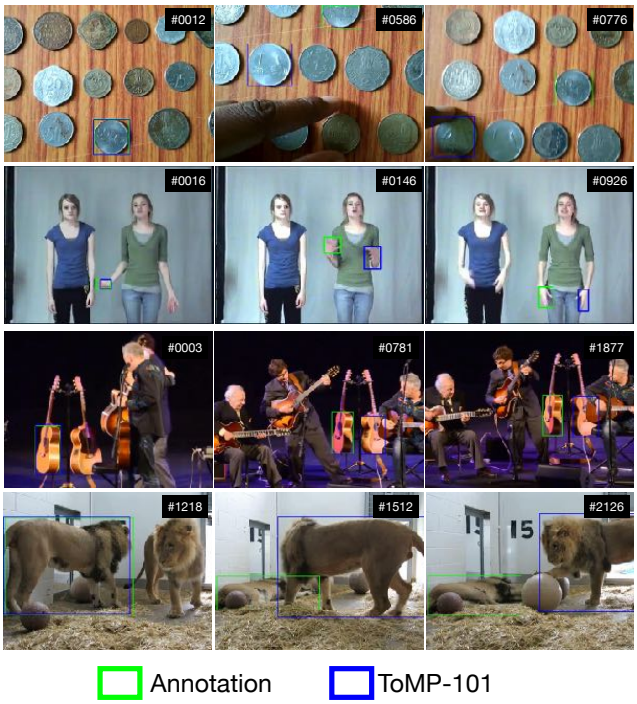


Figure 4. Visualization of failure cases of our tracker.

	ToMP 101+AR	**ToMP 50 +AR**	RPT [28, 37]	STARK ST50+AR [48]	STARK ST101+AR [48]	Ocean Plus [7, 28]	Alpha Refine [28, 49]	AFOD [28]	LWTL [4, 28]	Fast Ocean [28]
EAO	0.497	0.496	**0.530**	0.505	0.497	0.491	0.482	0.472	0.463	0.461
Accuracy	0.750	0.754	0.700	**0.759**	**0.763**	0.685	0.754	0.713	0.719	0.693
Robustness	0.798	0.793	**0.869**	0.817	0.789	**0.842**	0.777	0.795	0.798	0.803

Table 4. Comparison to the state of the art of segmentation only methods on VOT2020ST [28] in terms of EAO score.

we follow FCOS and multiply the classification and centerness scores point-wise to retrieve the target object (4th row). We conclude that omitting the centerness branch for training and during inference to localize the target achieves the best tracking performance.

B. Visual Results

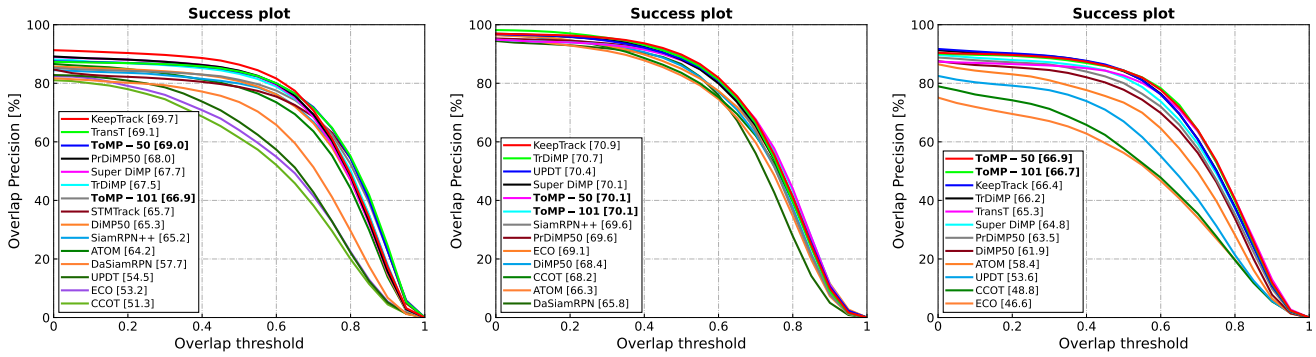
In this part we provide visual results of our tracker. First, we show three frames of different sequences where our tracker outperforms the state of the art. Secondly, we compare the produced target score map of our tracker with score maps obtained by optimization based model prediction. Finally, we show some failure cases of our tracker.

B.1. Visual Comparison to the State of the Art

Fig. 2 shows three frames of eight different LaSOT [19] sequences where each frame contains the ground truth annotation of the target object and the predictions of three different trackers: SuperDiMP [11], STARK-ST101 [48] and ToMP-101. We observe that our tracker produces in most sequences more robust and in some more accurate bounding box predictions than the related methods. In particular it achieves solid robustness for scenarios where distractors are present but the target object is at least partially visible and not undergoing a full occlusion.

B.2. Target Model Prediction

Fig. 1 shows the target score maps produced by the target model when using two different model predictors for three different sequences. In detail we compare the target score map produced by SuperDiMP [11] that adopts the DiMP [1] model predictor with optimized settings. In particular it uses a slightly smaller search area factor of 6 instead of 5 and a target score resolution of 22 instead of 18. Note, that our tracker uses 5 and 18 similar to DiMP [1] as stated Sec. A.2. We observe that our model predictor leads to much cleaner and unambiguous target localization than DiMP. While the former often produces multiple local maxima for distractors, our methods is able to almost fully suppress these. An important design choice that enables this is the transductive model weight and test feature prediction produced by our Transformer based model predictor. However, the cleaner score maps come with the risk, that once the target is lost and a distractor is tracked instead recovering is less likely since our tracker effectively suppresses



(a) UAV123 [39] (b) OTB-100 [46] (c) NFS [22]

Figure 5. Success plots on the UAV123 [39], OTB-100 [46] and NFS [22] datasets in terms of overall AUC score, reported in the legend.

	ToMP 101	**ToMP 50**	Keep Track [38]	STARK ST101 [48]	Tr DiMP [44]	TransT [6]	SAOT [58]	STARK ST50 [48]	Super DiMP [11]	Pr DiMP [15]	Siam R-CNN [42]	STM Track [21]	DiMP [1]	KYS [2]	Siam RPN++ [32]	ATOM [12]	UPDT [3]	Retina MAML [43]	FCOS MAML [43]
UAV123	66.9	69.0	**69.7**	68.2	67.5	**69.1**	**69.1**	–	67.7	68.0	64.9	64.7	65.3	–	61.3	64.2	54.5	–	–
OTB-100	70.1	70.1	70.9	68.1	71.1	69.4	68.5	71.4	70.1	69.6	70.1	**71.9**	68.4	69.5	69.6	66.9	70.2	71.2	70.4
NFS	**66.7**	**66.9**	66.4	66.2	66.2	65.7	65.2	65.6	64.8	63.5	63.9	–	62.0	63.5	–	58.4	53.7	–	–

	Ocean [55]	STN [36]	Auto Match [54]	Auto Track [34]	Siam BAN [8]	Siam CAR [25]	ECO [13]	DCFST [57]	PG-NET [35]	CRACT [20]	GCT [23]	Siam GAT [24]	CLNet [17]	TLPG [33]	Siam AttN [53]	Siam FC++ [47]	MDNet [40]	CCOT [14]	DaSiam RPN [59]
UAV123	–	64.9	–	67.1	63.1	61.4	53.2	–	–	66.4	50.8	64.6	63.3	–	65.0	–	–	51.3	57.7
OTB-100	68.4	69.3	71.4	–	69.6	–	69.1	70.9	69.1	**72.6**	64.8	71.0	–	69.8	71.2	68.3	67.8	68.2	65.8
NFS	–	–	–	–	59.4	–	46.6	64.1	–	62.5	–	–	54.3	–	–	–	41.9	48.8	–

Table 5. Comparison with state-of-the-art on the OTB-100 [46], NFS [22] and UAV123 [39] datasets in terms of overall AUC score.

	ToMP 101	**ToMP 50**	STARK ST101 [48]	Keep Track [38]	STARK ST50 [48]	Alpha Refine [49]	TransT [6]	Siam R-CNN [42]	Tr DiMP [44]	Super Dimp [11]	SAOT [58]	STM Track [21]	DTT [52]	Pr DiMP [15]	DM Track [56]	Auto Match [54]	TLPG [33]	TACT [9]	LTMU [10]
LaSOT	**68.5**	**67.6**	67.1	67.1	66.4	65.3	64.9	64.8	63.9	63.1	61.6	60.6	60.1	59.8	58.4	58.3	58.1	57.5	57.2

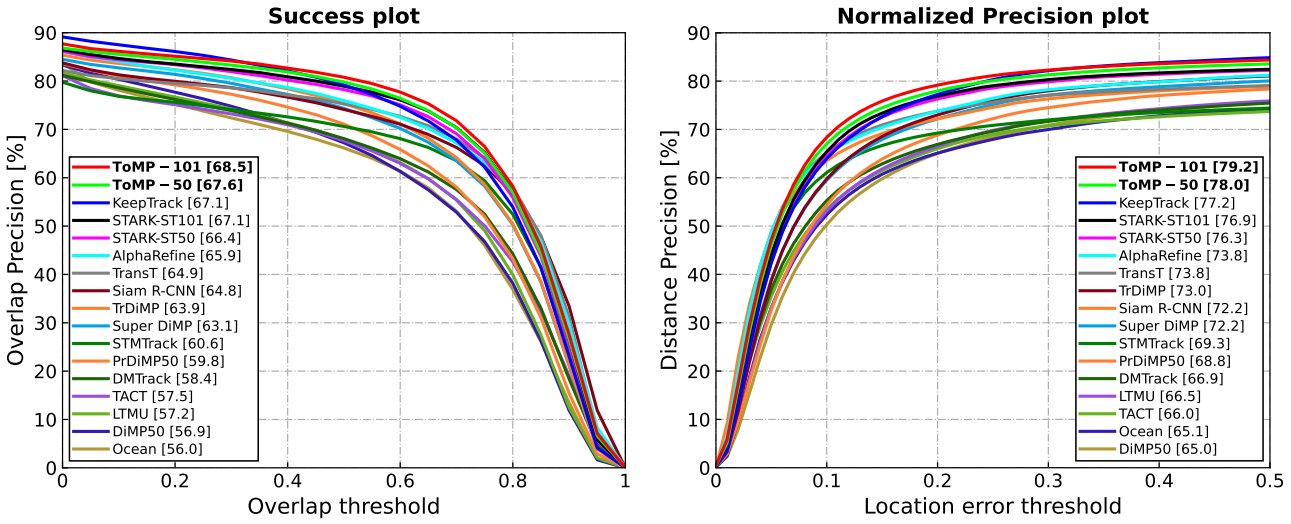
	DiMP [1]	Ocean [55]	Siam AttN [53]	CRACT [20]	Siam FC++ [47]	Siam GAT [24]	PG NET [35]	FCOS MAML [43]	Global Track [27]	ATOM [12]	DaSiam RPN [59]†	Siam BAN [8]	Siam CAR [25]	CLNet [17]	Siam RPN++ [32]†	Retina MAML [43]	Siam Mask [45]†	ROAM++ [51]	SPLT [50]
LaSOT	56.9	56.0	56.0	54.9	54.4	53.9	53.1	52.3	52.1	51.5	51.5	51.4	50.7	49.9	49.6	48.0	46.7	44.7	42.6

Table 6. Comparison with state-of-the-art on the LaSOT [19] test set in terms of overall AUC score. The symbol † marks results that were produced by Fan *et al*. [19] otherwise they are obtained directly from the official paper.

distractors. Similarly, our method learns to produce a score map containing a Gaussian such that overall the maximum score values are higher than by SuperDiMP. Thus, we chose a relatively high threshold to decide whether to use a previous prediction as training sample or not.

B.3. Failure Cases

Fig. 4 shows failure cases of our tracker. In particular, it shows three frames of four different LaSOT [19] sequences containing the ground truth annotations and the predicted bounding boxes of our tracker using a ResNet-101 [26] as backbone. To summarize, our tracker typically fails if object similar to the targets so called distractors are present. While the sole presence of distractors typically does not lead to tracking failure, our tracker shows difficulties in sequences where the target is occluded and distractors are present (1st and 3rd row). Instead of detecting that the target is occluded the tracker starts to track a distractor instead. Another challenging scenario are sequences where the target and a distractor approach each other (2nd row in Fig. 4) or one occludes the other (4th row in Fig. 4). The model then detects only a single object instead of two in both scenarios. Once they diverge again and the tracker detects two objects it typically fails to reliably differentiate between the target and the distractor.



(a) Success

(b) Normalized Precision

Figure 6. Success and normalized precision plots on LaSOT [19]. Our approach outperforms all other methods by a large margin in AUC, reported in the legend.

C. Experiments

We provide more detailed experiments to complement the comparison to the state of-the art performed in the main paper. And provide results for the VOT2020ST [28] challenge when using AlphaRefine [49] on top of our method in order to compare with methods that produce a segmentation mask as output.

C.1. VOT2020 with AlphaRefine

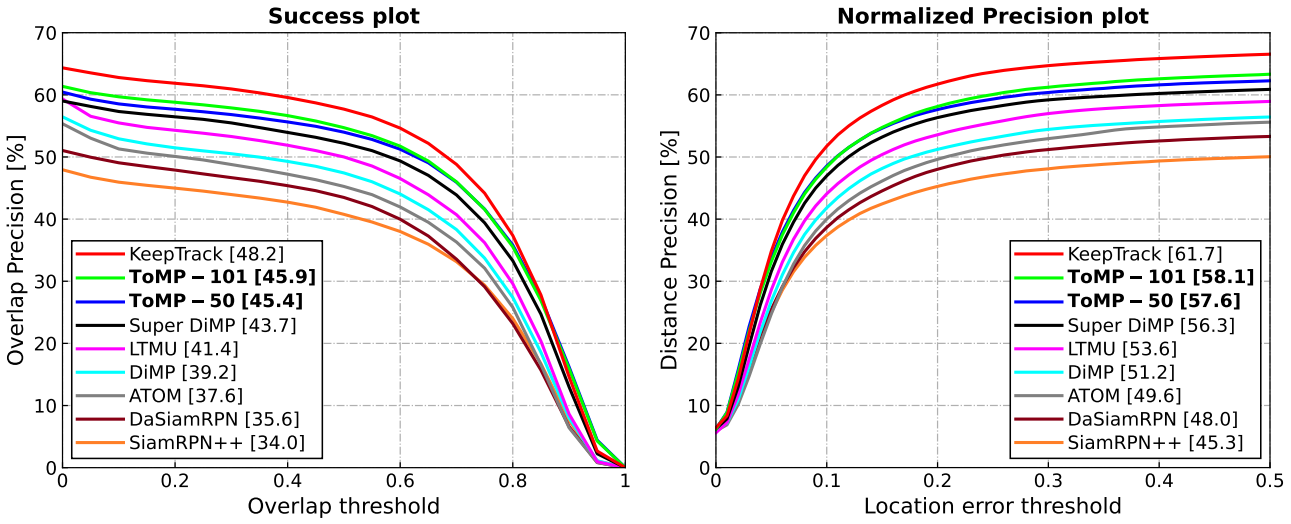
In contrast to previous years where the sequences in the VOT short-term challenge were annotated with bounding boxes [29, 30] the sequences of the more recent challenges contain segmentation mask annotations [28, 31] of the target in each frame. In the main paper we compare our method with methods that produce bounding boxes. Thus, in addition, we compare our method on the VOT2020 short-term challenge to methods that produce a segmentation mask in each frame. Since our method produces only a bounding box, we use AlphaRefine [49] that is able to produce a segmentation mask give the bounding box. Tab. 4 shows that our method achieves competitive results. In particular ToMP-101 achieves the same EAO (for more details on EAO we refer the reader to [28]) as STARK-ST101+AR [48] that employs AlphaRefine too. Nonetheless, RPT [37] achieves higher EAO than our tracker. In particular it scores a higher robustness but a lower accuracy than our trackers.

C.2. UAV123, OTB-100 and NFS

To complement the results detailed in the paper, we provide the success plots for the UAV123 [39] dataset in Fig. 5a, the OTB-100 [46] dataset in Fig. 5b and the NFS [22] dataset in Fig. 5c. Fig. 5a shows that KeepTrack [38] and PrDiMP50 [15] achieve higher robustness than our tracker ($T < 0.6$) but that our trackers together with TransT [6] reaches the highest accuracy among all trackers ($T > 0.7$) compensating for the lower robustness. Fig. 5b reveals similar conclusions on OTB-100. For NFS Fig. 5c shows that our tracker is almost as robust as KeepTrack [38] but achieves superior accuracy leading to a new state of the art. While we reported only the methods with the highest performances on these datasets in the main paper, we compare our method in Tab. 5 with additional related methods.

C.3. LaSOT and LaSOTExtSub

In addition to the success plots, we provide the normalized precision plots on the LaSOT [19] test set in Fig. 6 the LaSOTExtSub [18] test set in Fig. 7. The normalized precision score NPr_D measures the percentage of frames where the normalized distance (relative to the target size) between the predicted and ground-truth target center location is less than a threshold $D \in [0, 0.5]$. The ranking is determined by computing the AUC of each tracker. The AUC is reported in the legend of Figs. 6b and 7b. We compare our tracker on LaSOT with the state of the art in Tab. 6 and show their performance if available in Fig. 6. In Fig. 7 we show results of methods produced by Fan *et al*. [18] except KeepTrack [38] and SuperDiMP [11] that we obtained from Mayer *et al*. [38].



(a) Success

(b) Normalized Precision

Figure 7. Success and normalized precision plots on LaSOTExtSub [18]. Our approach outperforms all other methods by a large margin in AUC, reported in the legend.

	Illumination Variation	Partial Occlusion	Deformation	Motion Blur	Camera Motion	Rotation	Background Clutter	Viewpoint Change	Scale Variation	Full Occlusion	Fast Motion	Out-of-View	Low Resolution	Aspect Ration Change	Total
LTMU	56.5	54.0	57.2	55.8	61.6	55.1	49.9	56.7	57.1	49.9	44.0	52.7	51.4	55.1	57.2
PrDiMP50	63.7	56.9	60.8	57.9	64.2	58.1	54.3	59.2	59.4	51.3	48.4	55.3	53.5	58.6	59.8
STMTrack	65.2	57.1	64.0	55.3	63.3	60.1	54.1	58.2	60.6	47.8	42.4	51.9	50.3	58.8	60.6
SuperDiMP	67.8	59.7	63.4	62.0	68.0	61.4	57.3	63.4	62.9	54.1	50.7	59.0	56.4	61.6	63.1
TrDiMP	67.5	61.1	64.4	62.4	68.1	62.4	58.9	62.8	63.4	56.4	53.0	60.7	58.1	62.3	63.9
Siam R-CNN	64.6	62.2	65.2	63.1	68.2	64.1	54.2	65.3	64.5	55.3	51.5	62.2	57.1	63.4	64.8
TransT	65.2	62.0	67.0	63.0	67.2	64.3	57.9	61.7	64.6	55.3	51.0	58.2	56.4	63.2	64.9
AlphaRefine	**69.4**	62.3	66.3	**65.2**	70.0	63.9	58.8	63.1	65.4	57.4	53.6	61.1	58.6	64.1	65.3
STARK-ST50	66.8	64.3	66.9	62.9	69.0	66.1	57.3	67.8	66.1	58.7	53.8	62.1	59.4	64.9	66.4
STARK-ST101	67.5	**65.1**	68.3	64.5	69.5	66.6	57.4	**68.8**	66.8	58.9	54.2	63.3	59.6	65.6	67.1
KeepTrack	**69.7**	64.1	67.0	**66.7**	**71.0**	65.3	**61.2**	66.9	66.8	**60.1**	**57.7**	**64.1**	**62.0**	65.9	67.1
ToMP-50	66.8	64.9	**68.5**	64.6	70.2	**67.3**	59.1	67.2	**67.5**	**59.3**	56.1	**63.7**	61.1	**66.5**	**67.6**
ToMP-101	69.0	**65.3**	**69.4**	**65.2**	**71.7**	**67.8**	**61.5**	**69.2**	**68.4**	59.1	**57.9**	**64.1**	**62.5**	**67.2**	**68.5**

Table 7. LaSOT [19] attribute-based analysis. Each column corresponds to the results computed on all sequences in the dataset with the corresponding attribute.

C.3.1 Attributes

To support the attribute based analysis in the main paper, where we compared the performance of our tracker with other Transformer based trackers, we provide the detailed analysis for multiple trackers and ToMP in Tab. 7. ToMP-101 achieves the best performance on all but three. It achieves the second best results for *Motion Blur* behind KeepTrack [38] and similar to AlphaRefine [49]. Further ToMP-101 achieves the third best for *Full Occlusion* behind KeepTrack [38] and ToMP-50. Similarly it scores third for *Illumination Variation* behind KeepTrack [38] and AlphaRefine [49]. We further observe, that discriminative model prediction based methods such as TrDiMP [44], SuperDiMP [11], AlphaRefine [49], KeepTrack [38] and ToMP all outperform STARK [48] on the attribute *Background Clutter* showing the advantage of using full training samples during tracking instead of cropped templates that mainly cover the centered target.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 3, 4

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 4

[3] Goutam Bhat, Joakim Johnander, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Unveiling the power of deep tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 4

[4] Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *Proceedings of the European Conference on Computer Vision ECCV*, August 2020. 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229, August 2020. 2

[6] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 4, 5

[7] Yiwei Chen, Jingtao Xu, Jiaqian Yu, Qiang Wang, Byungin Yoo, and Jae Joon Han. AFOD: Adaptive focused discriminative segmentation tracker. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, August 2020. 3

[8] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[9] Janghoon Choi, Junseok Kwon, and Kyoung Mu Lee. Visual tracking by tridentalign and context embedding. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 4

[10] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[11] Martin Danelljan and Goutam Bhat. PyTracking: Visual tracking library based on PyTorch. https://github.com/visionml/pytracking, 2019. Accessed: 1/05/2021. 2, 3, 4, 5, 6

[12] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4

[13] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: efficient convolution operators for tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2017. 4

[14] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, October 2016. 4

[15] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4, 5

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1

[17] Xingping Dong, Jianbing Shen, Ling Shao, and Fatih Porikli. Clnet: A compact latent network for fast adjusting siamese trackers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 4

[18] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Mingzhen Huang, Juehuan Liu, Yong Xu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision (IJCV)*, 129(2):439–461, 2021. 5, 6

[19] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 4, 5, 6

[20] Heng Fan and Haibin Ling. Cract: Cascaded regression-align-classification for robust visual tracking. *arXiv preprint arXiv:2011.12483*, 2020. 4

[21] Zhihong Fu, Qingjie Liu, Zehua Fu, and Yunhong Wang. Stmtrack: Template-free visual tracking with space-time memory networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 4

[22] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *ICCV*, 2017. 2, 4, 5

[23] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4

[24] Dongyan Guo, Yanyan Shao, Ying Cui, Zhenhua Wang, Liyan Zhang, and Chunhua Shen. Graph attention tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 4

[25] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[27] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, February 2020. 4

[28] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, Linbo He, Yushan Zhang, Song Yan, Jinyu Yang, Gustavo Fernández, and et al. The eighth visual object tracking vot2020 challenge results. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, August 2020. 3, 5

[29] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pfugfelder, Luka Cehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, Gustavo Fernandez, and et al. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, September 2018. 5

[30] Matej Kristan, Jirí Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Luka Cehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, Abdelrahman Eldesokey, Jani Käpylä, Gustavo Fernández, and et al. The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, October 2019. 5

[31] Matej Kristan, Jiří Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Cehovin, Alan Lukežič, Ondrej Drbohlav, Jani Käpylä, Gustav Häger, Song Yan, Jinyu Yang, Zhongqun Zhang, and Gustavo Fernández. The ninth visual object tracking vot2021 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2711–2738, October 2021. 5

[32] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4

[33] Siyuan Li, Zhi Zhang, Ziyu Liu, Anna Wang, Linglong Qiu, and Feng Du. Tlpg-tracker: Joint learning of target localization and proposal generation for visual tracking. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI*, July 2020. 4

[34] Yiming Li, Changhong Fu, Fangqiang Ding, Ziyuan Huang, and Geng Lu. Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[35] Bingyan Liao, Chenye Wang, Yayun Wang, Yaonong Wang, and Jun Yin. Pg-net: Pixel to global matching network for visual tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 4

[36] Yuan Liu, Ruoteng Li, Yu Cheng, Robby T. Tan, and Xiubao Sui. Object tracking using spatio-temporal networks for future prediction location. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 4

[37] Ziang Ma, Linyuan Wang, Haitao Zhang, Wei Lu, and Jun Yin. RPT: learning point set representation for siamese visual tracking. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, August 2020. 3, 5

[38] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13444–13454, October 2021. 4, 5, 6

[39] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, October 2016. 4, 5

[40] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[41] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[42] Paul Voigtlaender, Jonathon Luiten, Philip H.S. Torr, and Bastian Leibe. Siam R-CNN: Visual tracking by redetection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[43] Guangting Wang, Chong Luo, Xiaoyan Sun, Zhiwei Xiong, and Wenjun Zeng. Tracking by instance detection: A meta-learning approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[44] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 4, 6

[45] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H.S. Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4

[46] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(9):1834–1848, 2015. 2, 4, 5

[47] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, February 2020. 4

[48] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10448–10457, October 2021. 2, 3, 4, 5, 6

[49] Bin Yan, Xinyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Alpha-refine: Boosting tracking performance by precise bounding box estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 3, 4, 5, 6

[50] Bin Yan, Haojie Zhao, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 'skimming-perusal' tracking: A framework for real-time and robust long-term tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 4

[51] Tianyu Yang, Pengfei Xu, Runbo Hu, Hua Chai, and Antoni B. Chan. Roam: Recurrently optimizing tracking model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[52] Bin Yu, Ming Tang, Linyu Zheng, Guibo Zhu, Jinqiao Wang, Hao Feng, Xuetao Feng, and Hanqing Lu. High-performance discriminative tracking with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9856–9865, October 2021. 4

[53] Yuechen Yu, Yilei Xiong, Weilin Huang, and Matthew R. Scott. Deformable siamese attention networks for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[54] Zhipeng Zhang, Yihao Liu, Xiao Wang, Bing Li, and Weiming Hu. Learn to match: Automatic matching network design for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13339–13348, October 2021. 4

[55] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 4

[56] Zikai Zhang, Bineng Zhong, Shengping Zhang, Zhenjun Tang, Xin Liu, and Zhaoxiang Zhang. Distractor-aware fast tracking via dynamic convolutions and mot philosophy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 4

[57] Linyu Zheng, Ming Tang, Yingying Chen, Jinqiao Wang, and Hanqing Lu. Learning feature embeddings for discriminant model based tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 4

[58] Zikun Zhou, Wenjie Pei, Xin Li, Hongpeng Wang, Feng Zheng, and Zhenyu He. Saliency-associated object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9866–9875, October 2021. 4

[59] Zheng Zhu, Qiang Wang, Li Bo, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 4