Supplementary Material for
STMTrack: Template-free Visual Tracking with Space-time Memory Networks

Zhihong Fu, Qingjie Liu*, Zehua Fu, Yunhong Wang
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
Hangzhou Innovation Institute, Beihang University
{fuzhihong, qingjie.liu, yhwang}@buaa.edu.cn, zehua_fu@163.com

1. Further Analyses

Here, we present additional experiments to demonstrate the effectiveness and superiority of the proposed method. First, The superiority of the pixel-level similarity computation is validated by comparing it with the feature-map-level cross correlation in Sec. 1.1. Then, comparisons with twelve competitive methods on multiple attribute subsets of LaSOT are given in Sec. 1.2. Finally, we show that our tracker requires fewer training samples than the siamese methods while achieves better performance in Sec. 1.3.

1.1. Pixel-level Similarity Computation *vs*. Feature-map-level Cross Correlation

Here, we compare the pixel-level similarity computation that is used in our proposed space-time memory network with the feature-map-level cross correlation that widely used in many siamese trackers. As shown in Fig. 1, Fig. 1(a) is the architecture of our proposed framework, and Fig. 1(b) is a typical siamese tracking framework that takes the initial frame of the tracking video as a fixed template to match the most similar region in the search frame by the depth-wise cross correlation.

To make fair comparisons, we use one memory frame in the training phase and put the initial frame of the tracking video into the memory during inference for our proposed framework. All frames are resized to be the same (*i.e.* 289×289) for both trackers, and a "Precise RoI Pooling [9]" module is applied to fix the spatial size of the template feature map f^t in the second tracker. Moreover, to make sure that the head networks of the two trackers have the same number of parameters, we increase the feature dimensionality of the cross correlation response maps from 512 to 1024 through a 1×1 convolutional layer. All hyper-parameters of the two trackers are the same as those used in the experiments of the text. Tab. 1 shows that the tracker deploying pixel-level similarity computation outperforms the one using feature-map-level cross correlation by 4% and 2.4% on OTB-2015 [17] and UAV123 [12] in terms of success (AUC) metric, respectively.

*Corresponding author.

Table 1: Performance comparisons of the tracker deploying the pixel-level similarity computation (denoted as $\mathcal{T}^P$) with the tracker deploying feature-map-level cross correlation (denoted as $\mathcal{T}^F$). Trackers are evaluated on OTB-2015 [17] and UAV123 [12] in terms of success (AUC) metric.

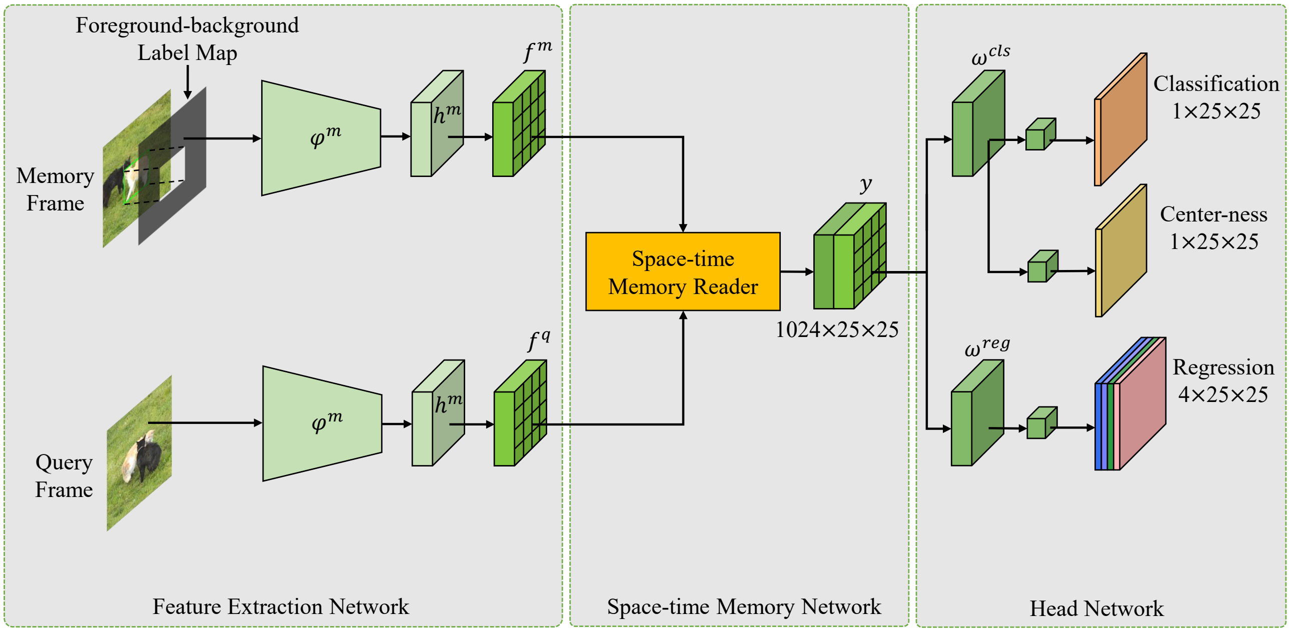
Tracker	OTB-2015	UAV123
$\mathcal{T}^P$	**0.711**	**0.632**
$\mathcal{T}^F$	0.671	0.608

1.2. Per-attribute Results on LaSOT

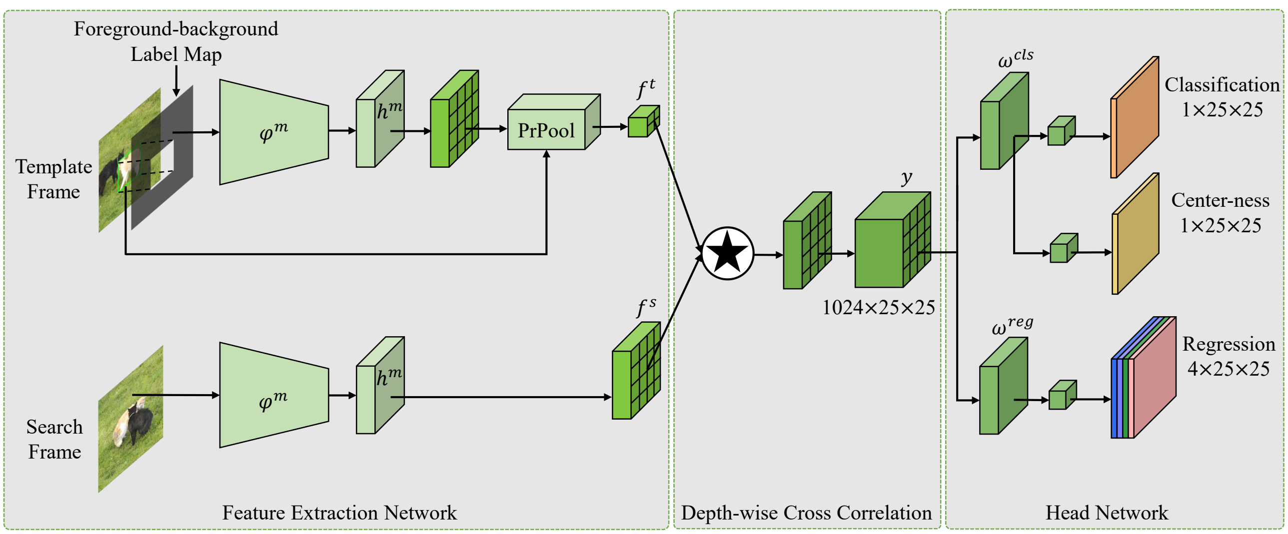
We test our tracker on the *testing* set of LaSOT [5], and compare it with twelve competitive methods: LTMU [3], DiMP-50 [1], Ocean [21], SiamFC++ [18], GlobalTrack [8], SiamCAR [6], ATOM [4], SiamBAN [2], SiamRPN++ [10], UpdateNet [20], ROAM++ [19], and VITAL [16]. Fig. 2 shows results on different attribute videos of the LaSOT *testing* set. It can be observed that our tracker has significant advantages when targets suffer from deformations (DEF), rotations (ROT), scale variations (SV), partial occlusions (POC), and illumination variations (IV). Specifically, it surpasses the second place methods by 4.0%, 5.2%, 4.1%, 3.4%, and 6.1% in scenarios of DEF, ROT, SV, POC, and IV, respectively. These advantages can be mainly attributed to the pixel-level similarity computation used in our proposed space-time memory network.

1.3. Amount of Training Data

We list the amount of training data used by our tracker and some top-performance siamese methods [2, 6, 18, 10] in Tab. 2, where YT-BB [14], TrackingNet [13], GOT-10k [7], ILSVRC VID [15], and LaSOT [5] are video datasets, and ILSVRC DET [15] and COCO [11] are image datasets. It

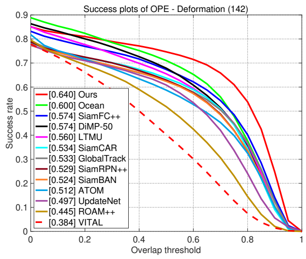


(a) A special case of our proposed framework, in which the number of memory frames is set to 1 during training and inference. For a fair comparison, the memory branch and the query branch share the same backbone φ^m and the same non-linear convolutional layer h^m.

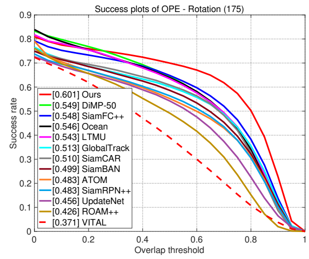


(b) A typical siamese tracking framework that uses a fixed template to match the most similar region in the search frame by the depth-wise cross correlation. For a fair comparison, we set the input size of the template frame to be the same as the input size of the search frame, and we also use a foreground-background label map in the template branch. We then utilize the precise RoI pooling [9] (denoted as PrPool in this figure) to fix the spatial size of the template feature map. The feature dimensionality of the cross correlation response map is increased from 512 to 1024 by a 1×1 convolutional layer to ensure that the head network has the same number of parameters as the one in Fig. 1(a). Here f^t and f^s are the feature maps of the template frame and the search frame, respectively. "$\star$" denotes the depth-wise cross correlation, and y is the cross correlation response map whose feature dimensionality is increased to 1024.

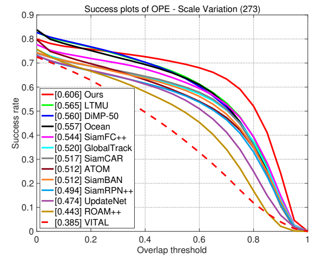
Figure 1: Two visual tracking frameworks. Fig. 1(a) is a special case of our proposed tracking framework that deploys the pixel-level similarity computation (a key operation in our proposed space-time memory network), and Fig. 1(b) is a conventional siamese tracking framework that deploys the feature-map-level cross correlation. In Fig. 1(a) and Fig. 1(b), φ^m is a backbone for the feature extraction, h^m is a non-linear convolutional layer for the feature dimensionality reduction, and ω^{cls} and ω^{reg} are two lightweight convolutional networks for the foreground-background classification and the target bounding box regression, respectively.



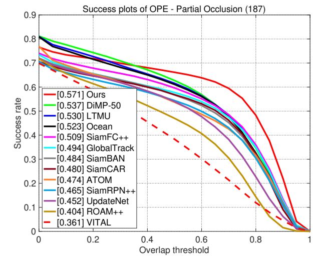
(a) Deformation



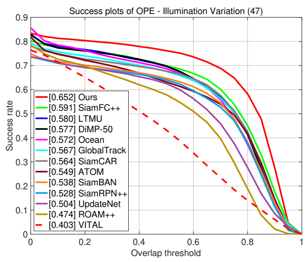
(b) Rotation



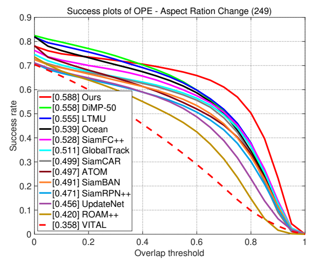
(c) Scale variation



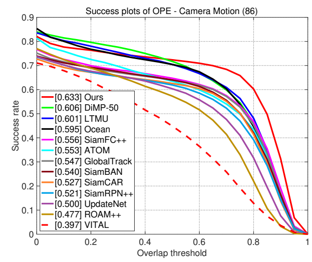
(d) Partial occlusion



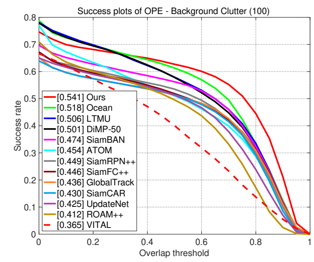
(e) Illumination variation



(f) Aspect ratio change



(g) Camera motion



(h) Background clutter

Figure 2: Performance comparisons of our proposed tracker with numerous competitive methods on several subsets with different attributes from the LaSOT *testing* set.

Table 2: A training data usage comparison of our proposed tracker with some top-performance siamese methods [2, 6, 18, 10]. YT-BB is the abbreviation for YouTube BoundingBoxes [14]. #Vids + #Imgs: number of videos plus number of additional static images.

Tracker	Videos					Additional Images		Total
	YT-BB 380k	TrackingNet 30k	GOT-10k 9k	ILSVRC VID 4k	LaSOT 1k	ILSVRC DET 457k	COCO 119k	#Vids + #Imgs
Ours		✓	✓	✓	✓	✓	✓	44k + 576k
SiamBAN	✓		✓	✓	✓	✓	✓	394k + 576k
SiamCAR	✓			✓		✓	✓	384k + 576k
SiamFC++	✓		✓	✓	✓	✓	✓	394k + 576k
SiamRPN++	✓			✓		✓	✓	384k + 576k

can be seen that, compared with these siamese methods, our tracker requires much fewer training samples yet achieves better performance.

2. Qualitative Results

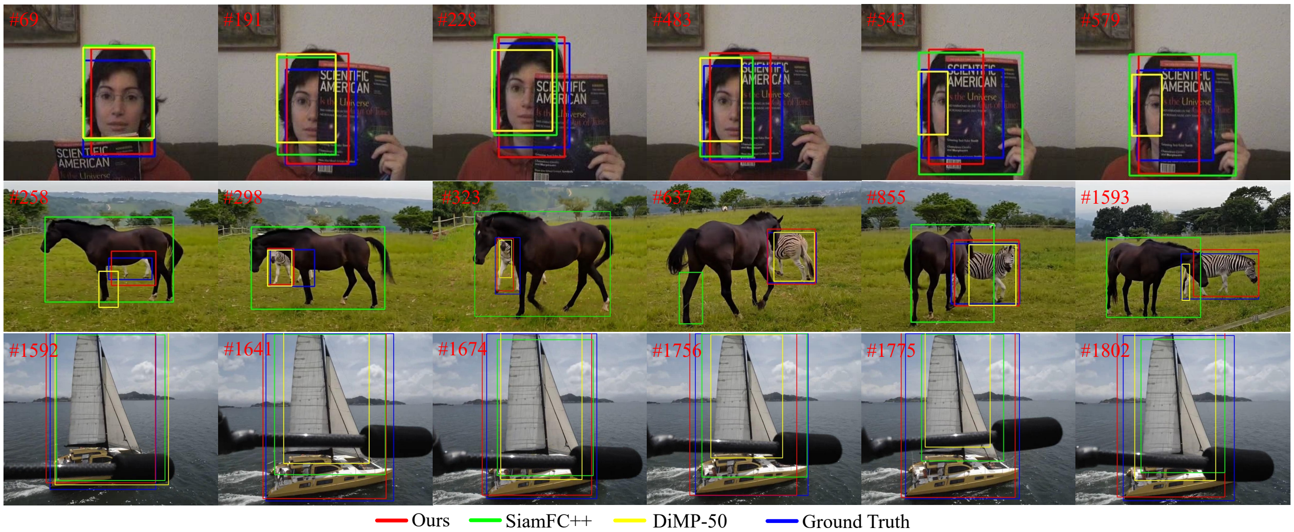
We provide additional qualitative results of our tracker (shown in red) in Fig. 3. Video sequences are collected from OTB-2015 [17] and LaSOT [5]. For intuitive comparisons, the results of two state-of-the-art trackers SiamFC++ [18] (shown in green), DiMP-50 [1] (shown in yellow), and the corresponding ground truth (shown in blue) are also visualized in each snapshot. All visualized video sequences are challenging, as described below:

- Fig. 3(a) shows the accuracies of trackers when the targets suffer from partial occlusions.
- Fig. 3(b) illustrates the semantic awareness of trackers when the targets suffer from non-rigid deformations.
- Fig. 3(c) demonstrates the discriminative ability of trackers when the targets distracted by similar objects and backgrounds are cluttered.

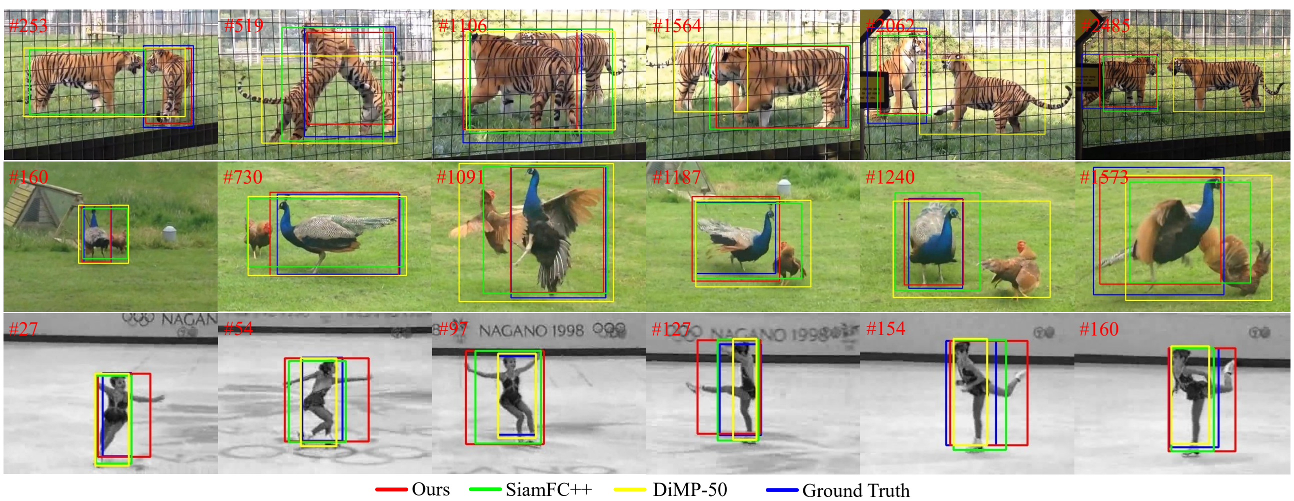
References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, pages 6182–6191, 2019. 1, 3

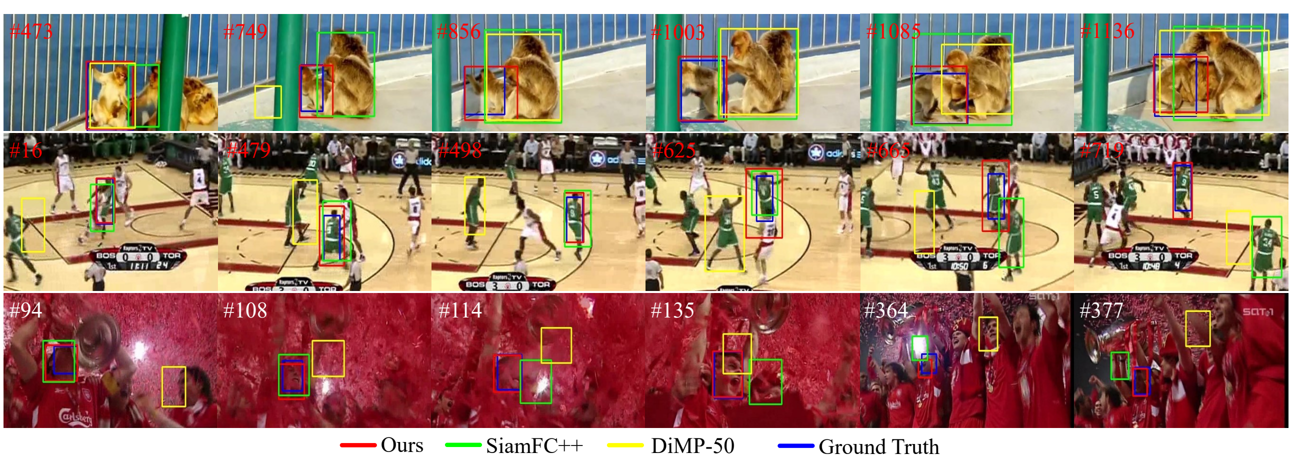
[2] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *CVPR*, pages 6668–6677, 2020. 1, 3



(a) Video sequences in which the targets suffer from partial occlusions. Here our proposed tracker shows higher accuracies.



(b) Video sequences in which the targets suffer from non-rigid deformations. Here our proposed tracker shows stronger semantic awareness.



(c) Video sequences in which the targets are distracted by similar objects and backgrounds are cluttered. Here our proposed tracker shows stronger discriminative ability.

Figure 3: Qualitative examples in three difficult challenges: partial occlusion, non-rigid deformation, and background clutter.

[3] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *CVPR*, pages 6298–6307, 2020. 1
[4] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, pages 4660–4669, 2019. 1
[5] 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 *CVPR*, pages 5374–5383, 2019. 1, 3
[6] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *CVPR*, pages 6269–6277, 2020. 1, 3
[7] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *TPAMI*, 2019. 1
[8] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. In *AAAI*, volume 34, pages 11037–11044, 2020. 1
[9] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, pages 784–799, 2018. 1, 2
[10] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, pages 4282–4291, 2019. 1, 3
[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1
[12] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *ECCV*, pages 445–461, 2016. 1
[13] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, pages 300–317, 2018. 1
[14] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, pages 5296–5305, 2017. 1, 3
[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1
[16] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson WH Lau, and Ming-Hsuan Yang. Vital: Visual tracking via adversarial learning. In *CVPR*, pages 8990–8999, 2018. 1
[17] Y. Wu, J. Lim, and M. Yang. Object tracking benchmark. *TPAMI*, 37(9):1834–1848, 2015. 1, 3
[18] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *AAAI*, pages 12549–12556, 2020. 1, 3
[19] Tianyu Yang, Pengfei Xu, Runbo Hu, Hua Chai, and Antoni B Chan. Roam: Recurrently optimizing tracking model. In *CVPR*, pages 6718–6727, 2020. 1
[20] Lichao Zhang, Abel Gonzalez-Garcia, Joost van de Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Learning the model update for siamese trackers. In *ICCV*, pages 4010–4019, 2019. 1
[21] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *ECCV*, 2020. 1