# Dynamic Attention Guided Multi-Trajectory Analysis for Single Object Tracking

Xiao Wang, Zhe Chen, *Member, IEEE*, Jin Tang, Bin Luo, *Senior Member, IEEE*, Yaowei Wang, *Member, IEEE*, Yonghong Tian, *Senior Member, IEEE*, and Feng Wu, *Fellow, IEEE*

***Abstract*— Most of the existing single object trackers track the target in a unitary local search window, making them particularly vulnerable to challenging factors such as heavy occlusions and out-of-view movements. Despite the attempts to further incorporate global search, prevailing mechanisms that cooperate local and global search are relatively static, thus are still sub-optimal for improving tracking performance. By further studying the local and global search results, we raise a question: can we allow more dynamics for cooperating both results? In this paper, we propose to introduce more dynamics by devising a dynamic attention-guided multi-trajectory tracking strategy. In particular, we construct dynamic appearance model that contains multiple target templates, each of which provides its own attention for locating the target in the new frame. Guided by different attention, we maintain diversified tracking results for the target to build multi-trajectory tracking history, allowing more candidates to represent the true target trajectory. After spanning the whole sequence, we introduce a multi-trajectory selection network to find the best trajectory that deliver improved tracking performance. Extensive experimental results show that our proposed tracking strategy achieves compelling performance on various large-scale tracking benchmarks. The project page of this paper can be found at https://sites.google.com/view/mt-track/.**

Manuscript received December 12, 2020; accepted January 19, 2021. Date of publication February 3, 2021; date of current version December 6, 2021. This work was supported in part by Key-Area Research and Development Program of Guangdong Province under Grant 2019B010155002, in part by the Post-Doctoral Innovative Talent Support Program under Grant BX20200174, in part by the China Post-Doctoral Science Foundation Funded Project under Grant 2020M682828, in part by the Australian Research Council Projects under Grant FL-170100117, and in part by the National Nature Science Foundation of China under Grant 61860206004, Grant 61825101, and Grant 62076003. The main part of this work was done when Xiao Wang have a visit at The University of Sydney in 2019. He is now a Post-Doctoral Research with the Peng Cheng Laboratory, Shenzhen, China. This article was recommended by Associate Editor Q. Huang. *(Corresponding authors: Jin Tang; Zhe Chen.)*

Xiao Wang is with the Cognitive Computing Research Center, Anhui University, Hefei 230601, China, also with the School of Computer Science and Technology, Anhui University, Hefei 230601, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: wangx03@pcl.ac.cn).

Zhe Chen is with the Faculty of Engineering, School of Computer Science, The University of Sydney, Sydney, NSW 2008, Australia (e-mail: zhe.chen1@sydney.edu.au).

Jin Tang and Bin Luo are with the Cognitive Computing Research Center, Anhui University, Hefei 230601, China, and also with the School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: tangjin@ahu.edu.cn; luobin@ahu.edu.cn).

Yaowei Wang and Yonghong Tian are with the Peng Cheng Laboratory, Shenzhen 518000, China, and also with the National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: wangyw@pcl.ac.cn; tianyh@pcl.ac.cn).

Feng Wu is with the University of Science and Technology of China, Hefei 230052, China (e-mail: fengwu@ustc.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3056684.

Digital Object Identifier 10.1109/TCSVT.2021.3056684

***Index Terms*— Visual object tracking, beam search, dynamic target-aware attention, trajectory selection network.**

## I. Introduction

With the powerful representation ability of deep neural networks, existing single object visual trackers promisingly promote state-of-the-art tracking performance one after another under tracking-by-detection framework, benefiting many other practical applications, including UAV, robotics, video surveillance, and so on. However, even using cutting-edge deep neural networks [1], [2], current tracking algorithms are still poor under some challenging scenarios, such as heavy occlusions, out-of-view movements, fast motions, scale variations. This suggests that only improving deep visual features may reach a bottleneck for boosting tracking accuracy, which drives us to devise a novel and more advanced tracking strategy to tackle these challenging factors.

According to our observation, many of the existing single object trackers that follow the *tracking-by-detection framework* generally attempt to track the target object, which is initialized in the first frame, based on a unitary local search window. However, without modeling global information appropriately, such a local search strategy can be sensitive to the aforementioned challenging factors, thus visual trackers can be prone to drifting away. To handle these issues, some researchers propose to incorporate global search strategy into visual tracking [3]–[7]. Despite that current global search approaches can handle the issues of the local search tracking strategy to some extents, they still suffer from the dilemma of selection between local and global search for each frame. Moreover, it is also difficult for current approaches that introduce global information to accurately locate the potential regions that contain the target, especially when multiple similar objects appeared in the image simultaneously.

With deeper investigation, we find that current approaches, even those that cooperate local and global search results, only model target appearance and estimate target status in a relatively static way, i.e., the tracking paradigms for each frame is fixed and static to a single bounding box. In other words, current visual trackers can be easily distracted by similar objects or noisy backgrounds when challenging factors like occlusions and out-of-view movements occur in the sequence if the target appearance modeled by the tracker is corrupted in the estimated bounding box. Furthermore, in existing trackers, the noise of an unreliable tracking result at a frame could also be continuously accumulated in later frames when the estimated target bounding box at each frame drift away from true position. This easily leads to more severe drifting. Even

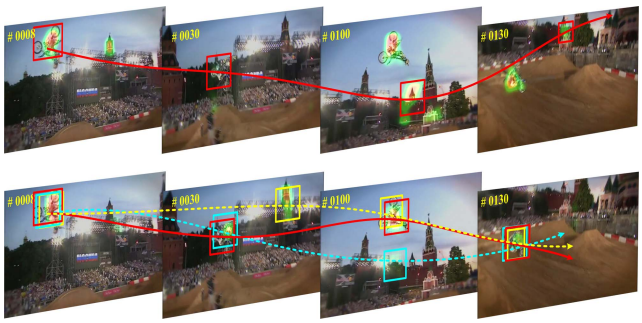



Fig. 1. **The comparison between single trajectory tracking (the first row) and our dynamic target-aware attention guided multi-trajectory tracking framework (the bottom row).** Existing trackers usually estimate one location only for each frame, which may prone to model drift in challenging scenarios even the global search scheme adopted. Our proposed multi-trajectory tracking framework maintain multiple locations guided by dynamic target-aware attention for each frame and significantly increase the dynamics of baseline trackers.

though some studies [8]–[10] introduce re-detect mechanism to correct the inaccurate tracking results once the target object is found to be lost, the already failed tracking results before re-detection will not be corrected anyway, and these erroneous tracking results still degrade the overall tracking performance. To tackle these problems, we instead seek to explore a novel search strategy that can allow more dynamics for single object tracking.

We are inspired by the recent progress of image captioning [11], [12] task, in which given images are generally embedded into deep representations, and a sequential model is applied to deliver corresponding language descriptions based on the embedded representations. In particular, by maintaining multiple words at each time step, the decoding stage provides more adequate language descriptions by applying beam search algorithm to allow more dynamics when analyzing the maintained words. We believe that the visual trackers can also take advantage of such a strategy that searches better results through different candidates, if multiple different tracking states and trajectories are maintained during tracking.

In this study, we propose a novel dynamic attention-guided multi-trajectory tracking framework to introduce more dynamics for tackling various challenging tracking issues. As shown in Fig. 2, our tracker consists of three main modules, i.e., a novel dynamic target-aware attention network to guide the global search for the target, a baseline tracker for local search about the target, and a multi-trajectory selection network. Different from existing works [4]–[6], our dynamic target-aware attention network maintains multiple target templates and fully utilizes hierarchical semantic feature representations obtained from different convolutional layers to provide more robust appearance model for the target and more accurate candidate regions when performing global search. Meanwhile, in each frame, the baseline tracker is employed to conduct local search for the target. By performing a joint local and global search for visual tracking in a parallel manner for each frame, we obtain and maintain multiple candidate states for the target. After spanning the whole video sequence, we employ a multi-trajectory selection network to find the best tracking trajectory based on different maintained tracking results for the sequence.

To sum up, the contributions of this paper can be summarized in the following three aspects:

– We propose a novel dynamic attention-guided multi-trajectory based tracking framework for single object tracking. In particular, based on multiple tracking trajectories maintained for the target, our proposed approach introduces more dynamics for cooperating local and global tracking results to help tackle the challenging tracking issues effectively.

– We propose a dynamic target-aware attention network to build a more robust and dynamic appearance model for global search based tracking. We also devise a novel multi-trajectory selection network to estimate the best tracking results according to maintained multi-trajectory information.

– Our proposed tracking strategy achieves compelling performance on several large-scale tracking benchmarks, validating the effectiveness of our proposed method.

## II. Related Work

### A. Siamese Network Based Trackers

Many trackers are developed based on Siamese network due to its high accuracy and efficiency. Tao *et al.* [13] propose to use the Siamese network to learn a fixed matching function and tracking the target object without any update. Wang *et al.* [14] further developed this framework by introducing adversarial samples for robust visual tracking. Bertinetto *et al.* [15] use a fully convolutional Siamese network for tracking by measuring the region-wise feature similarity between the target object and candidates (named SiamFC). GOTURN is proposed by Held *et al.* [16] which uses a motion prediction model developed based on Siamese network for high-speed tracking. Recently, many researchers focus on designing a more powerful network for Siamese based tracker, such as SiamDW [1], SiamRPN++ [2]. Some works improved the Siamese network based tracker by cascaded region proposal networks [17], relation reasoning [18], dynamic target template update [19], [20], re-detection [21], meta-learning [22], [23], and others [24]. Although these works all improved the Siamese network based tracker from different perspectives, however, they still adopt the greedy search based strategy for visual tracking. Therefore, these trackers may be sensitive to the challenging factors we mentioned above. In this paper, we propose a novel multi-trajectory tracking framework for visual tracking which can address issues caused by greedy search to some extent.

### B. Long-Term Tracking

Existing long-term tracking algorithms usually introduce *re-detection* mechanism in the tracking procedure. Kalal *et al.* [25] propose a tracking-learning-detection framework for long-term tracking. They utilize optical-flow based matcher for local search and also adopt the ensemble of weak classifiers for re-detection. Ma *et al.* [8] propose a long-term correlation filter which use KCF for local tracking and external random ferns classifier for long-term detector. Valmadre *et al.* [26] develop a long-term tracker named SiamFC+R, which also integrates a simple re-detection scheme with SiamFC and conduct re-detection when the

SiamFC's response is lower than a pre-defined threshold. Some works attempt to address the tracking task with *key point matching* [10], [27] or *global proposal mechanism* [4]–[7]. However, as noted in [3], the keypoint extractors and descriptors are not reliable in complex environments, which may limit their overall performance. Zhu *et al.* [7] propose the EBT tracking based on EdgeBox [28] and improve the baseline method significantly. Yan *et al.* [3] propose a 'Skimming-Perusal' tracking framework for real-time and long-term tracking. They use a switching mechanism to decide local or global proposals should be used in each frame and achieve better results on OxUvA dataset. Fan and H. Ling [29] propose a parallel tracking and verifying framework (PTAV) which can achieve better results on UAV20L dataset. Wang *et al.* [4]–[6] develop the target-aware attention mechanism for global proposal generation and integrate with MDNet for robust tracking. GlobalTrack is proposed in [30] which only employ global search on whole video frames also achieves good performance on long-term datasets. Dai *et al.* [31] propose offline-trained meta-updater to address the update issue in long-term tracking.

### C. Trajectory-Based Tracking

Although very few, there are still some related works that exploit trajectory information to achieve tracking. For example, MTA [32] is proposed to conduct tracking by trajectory selection using tracking results obtained from STRUCK [33]. In practice, the MTA is subject to poor hand-crafted features and empirical multi-trajectory analysis that can not be optimized. In addition to MTA, researchers also develop MHT [34] to conduct multi-object tracking based on different trajectories. To track multiple objects, the MHT designs a track tree construction and updating scheme to allow more dynamics. However, there are many differences between our proposed method and MHT. First, MHT relies on hand-crafted features and requires detected bounding boxes, while our framework is built upon more powerful deep convolutional features and does not require the detection model to provide bounding box results. Second, MHT uses tree structure to derive trajectories, while our method maintain independent tracking results and derive trajectories based on beam search strategy which can allow more trajectories to describe a single target. Also, MHT relies on detection models to encode target appearances, while our method can dynamically encode appearances. Lastly, rather than MHT that performs empirical tree construction algorithm to deliver trajectory results, our proposed multi-trajectory selection network can be optimized to select the best trajectory result.

## III. Our Proposed Approach

In this section, we will first give an introduction to the baseline tracker THOR used in this work. Then, we will give an overview of our proposed modules. After that, we will dive into the details on dynamic attention model, trajectory selection network. We mainly focus on the motivation, detailed network architecture, and advantages of our proposed modules for object tracking.

### A. Preliminary: THOR

THOR is a Siamese network based visual tracker proposed by Axel Sauer *et al.* [20]. They design a simple but effective target template update mechanism for tracking procedure. Specifically, they utilize two modules, i.e., the short-term module (STM) and long-term module (LTM) to store the tracked results. The authors define a novel diversity measure in the space of Siamese features to select the most diverse templates.

For the long-term module, they try to maximize the volume $\Gamma(f_1, \ldots, f_n)$ of the parallelotope formed by the feature vectors $f_i$ of the template $T_i$. They use convolutional operation to compute the similarity between different templates in memory and obtain a Gram matrix:

$$G(f_1, \ldots, f_n) = \begin{bmatrix} f_1 \star f_1 & f_1 \star f_2 & \ldots & f_1 \star f_n \\ \vdots & \vdots & \ddots & \vdots \\ f_n \star f_1 & f_n \star f_2 & \ldots & f_n \star f_n \end{bmatrix} \quad (1)$$

where $G$ is a square $n \times n$ matrix. Therefore, the Gram determinant (i.e., the determinant of $G$) can be written as:

$$\max_{f_1,\ldots,f_n} \Gamma(f_1, \ldots, f_n) \propto \max_{f_1,\ldots,f_n} |G(f_1, \ldots, f_n)| \quad (2)$$

The template can be incorporated into the memory, if the Gram determinant can be increased when replacing one of the allocated templates.

The short-term module is introduced to handle abrupt movements and partial occlusion. They update the memory slots of STM in a first-in, first-out manner. Different from LTM, they compute the diversity measure $\gamma$ of STM as follows:

$$\gamma = 1 - \frac{2}{N(N+1)G_{st,max}} \sum_{i<j}^{N} G_{st,ij} \quad (3)$$

The authors integrate the LTM and STM module with three Siamese network based trackers (SiamFC [15], SiamRPN [35] and SiamMask [36]), and achieve better tracking performance than the baseline methods. More details of THOR can be found in [20].

THOR works well on short-term and small datasets, the performance on large-scale and long-term tracking datasets are still unknown. This tracker also adopts a greedy search in a local search window which makes their performance unsatisfied on challenging benchmarks. In this paper, we introduce a novel dynamic target-aware attention mechanism and integrate with SiamRPN based THOR for robust tracking in the multi-trajectory manner.

### B. Overview

In general, our proposed method consists of three key components, including *the dynamic target-aware attention network*, *a local search-based baseline tracker*, and *the multi-trajectory selection network*. Given a testing video sequence, we first perform local tracking based on a Siamese tracker (e.g. THOR [20]) to search the target according to the response map of a local region. In the meantime, we perform global search to locate the target within the whole frame. To tackle the challenging factors, such as heavy occlusions, fast motions,

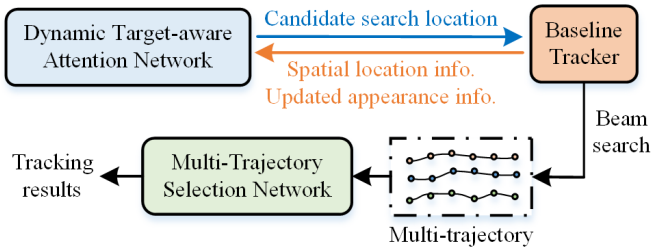



Fig. 2. The pipeline of our proposed dynamic target-aware attention guided multi-trajectory tracking.

and out-of-view movements, we introduce a novel dynamic target-aware attention network to help construct a more robust and dynamic appearance model, facilitating the global tracking to better locate the target. Moreover, rather than cooperating local and global search results in a relatively static way, we propose to maintain multiple different tracking results obtained from local and global search in a parallel manner to achieve multi-trajectory tracking with more dynamics. At the end of testing video, with the help of a carefully designed trajectory selection network, we select the trajectory with the highest confidence score to deliver a further refined final tracking results.

### C. Dynamic Target-Aware Attention Network

*1) Our Approach:* We proposed the dynamic target-aware attention network to construct dynamic appearance model by maintaining multiple target templates. In particular, we first extract feature maps from three different convolutional modules to deliver hierarchical and more powerful representations. Then, we treat target features as convolutional filters. We employ these filters to perform convolution on the feature maps of the whole image. The output features will be later fed into a gate layer which is used to control the information flow. Afterwards, the gated features will be gradually fed into a decoder network and output corresponding attention map. Inspired by [37], we incorporate the *point proposals* to make the attention map focus more on the target regions encoded by the coordinates of the bounding box corners to handle the issues caused by similar target objects. In our proposed method, we encode point proposals into horizontal and vertical feature maps with relative CoordConv [37] and regard it as prior information to estimate the current attention map. Also, we apply an RoI pooling layer to extract the point feature and feed it into an adaptive instance layer [38] to achieve instance-level attention estimation. Moreover, we employ the local search-based tracker to help collect and maintain a set of short-term and long-term target templates (i.e., the dynamic template pool in Fig. 4) to provide dynamic target feature representations during the tracking procedure. In this way, we can utilize the updated target templates to conduct dynamic target-aware attention estimation. In general, the proposed dynamic target-aware attention network can provide more accurate global search regions than the static counterpart methods [4]–[6]. An overview of our attention network is given in Fig. 3.

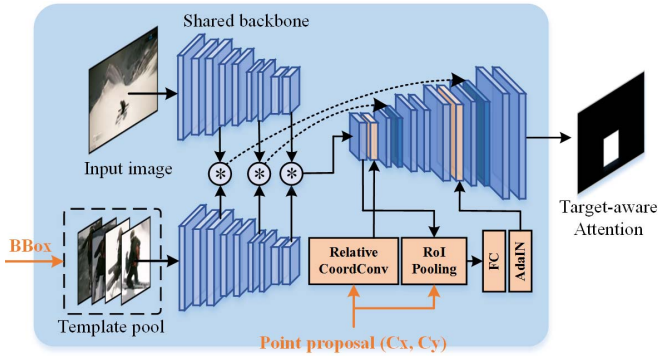



Fig. 3. Illustration of our dynamic target-aware attention generation module.

*2) Network Architecture:* Given a video frame $I$ and a target template $T$, we first resize them into $300 \times 300$ and $100 \times 100$, respectively. Then, we put them into an encoder network with two branches. Each branch is a residual network (ResNet-18 used in this paper) which is pre-trained on ImageNet classification task. To obtain a more effective feature representation, we extract the hierarchical semantics from feature maps of both inputs from three convolutional blocks whose dimension are $[128 \times 38 \times 38;\ 256 \times 19 \times 19;\ 512 \times 10 \times 10]$ and $[128 \times 13 \times 13;\ 256 \times 7 \times 7;\ 512 \times 4 \times 4]$, respectively. We use $F_i^T$ and $F_i^I$ to denote the $i^{th}, i \in \{1, 2, 3\}$ feature map of target object and global image. Different from previous works which directly concatenate the two feature maps, we conduct convolutional operation with target feature $F_i^T$ on global image features $F_i^I$ to boost their interaction:

$$F_i^{TI} = F_i^T \star F_i^I \tag{4}$$

where $\star$ denotes the convolutional operator. Then, $F_i^{TI}$ is fed into a gate layer to achieve controllable information flow which is widely used in recurrent neural network (such as forget gate, input gate in LSTM [39]). This process can be written as:

$$\hat{F_i^{TI}} = F_i^{TI} \odot \sigma(F_i^{TI}) \tag{5}$$

where $\sigma$ is a sigmoid gate layer and the $\odot$ denote dot product between two matrix. These features will be fed into a decoder network gradually using skip connections to predict corresponding target-aware attention map.

Global search-based tracking that only utilizes static target templates could suffer from the issue of the existence of multiple similar target objects in the image. Since the global search module will provide multiple search regions for different similar target objects, it can be very difficult for the tracker to recognize the true target. In this paper, we propose to utilize *point proposal* $(x, y)$ to help target-aware attention estimation for a more robust global search. The center location of the previous tracking result is considered as the point proposal. Specifically, the point proposal is used in two ways: the *location prior* of target object and *"characteristic" embedding*. For the *location prior*, existing works [37], [40], [41] all shown that the pixel coordinates are important to disambiguate different object instances. Liu *et al.* [41] proposed

a simple CoordConv layer to encode the pixel coordinates into feature maps by creating a tensor of the same size as input that contains pixel coordinates normalized to $[-1, 1]$. The authors of [37] further proposed a *Relative CoordConv* block which can utilize standard backbone networks for their task, such as pre-trained ResNet. Following [37], we also construct two coordinate feature maps, one for $x$-coordinates and one for $y$-coordinates. The values of constructed $x$-map vary from $-1$ to $+1$, where $-1$ and $+1$ corresponding to $x - R$ and $x + R$, respectively. The $y$-map has similar attributes and the $R$ is a hyper-parameter used to denote the radius of Relative CoordConv. The constructed prior map is concatenated with feature maps produced by the regular convolutional network. Besides, we also attain the "*characteristic*" embedding based on the point proposal. More specifically, given the point proposal $(x, y)$, we extract a $1 \times 1 \times 512$ feature $Q(x, y)$ with RoI pooling on the feature maps generated from the third convolutional layer of decoder network. This feature vector is then fed into two fully connected layers and the AdaIN layer [38] to achieve instance selection. The output feature will be integrated into the decoder network.

Using the template representation collected and maintained by the local search-based baseline tracker to achieve Siamese tracking, our attention estimation can be implemented in a *dynamic* manner, while previous works are all *static* target-aware attention. In other word, we can borrow the updated target representation in the tracking procedure for more accurate attention estimation. Adaptively switching between local and global search is one intuitive approach for robust tracking as many previous trackers do [3]–[6]. To maximize the benefits of dynamic appearance model for the whole video sequence, our tracker runs in a batch manner.

*3) Advantages of Our Approach:* Compared with original target-aware attention based trackers, the highlights of our proposed dynamic attention model can be listed as follows: **Firstly**, the original method exploits the features of the target object and global image based on simple concatenation operation without considering the relationships and interactions between the two types of features. This may limit the representational power of the model. **Secondly**, they only utilize the feature map from the last convolutional layer of their encoder, which can not fully utilize the hierarchical semantic information for better attention prediction. **Thirdly**, they only utilize the feature map of the target object initialized in the first frame. However, since the appearance of the target is generally continuously changing in a tracking sequence, such a fixed appearance model is almost infeasible to track the target that undergoes significant appearance variations in the corresponding video sequence. **Last but not the least**, their attention model is also too primitive to handle multiple similar target objects. In practice, there would be multiple high response regions which may easily distract the tracker on the similar but non-target objects.

As a result, the aforementioned issues inspire us to design an advanced new attention scheme for tracking in challenging scenarios, especially for long-term object tracking. In particular, we propose the dynamic target-aware attention mechanism, as shown in Fig. 3, to introduce more dynamics in the global search-based tracking procedure. By further addressing the issues of the original target-aware attention methods [4], [5] [6] in this study, our method constructs a more robust and more dynamic target-aware attention model for global search-based tracking.

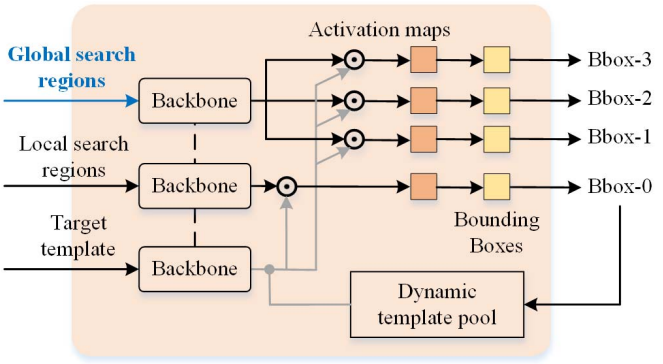



Fig. 4. Illustration of our attention guided multi-trajectory tracking module.

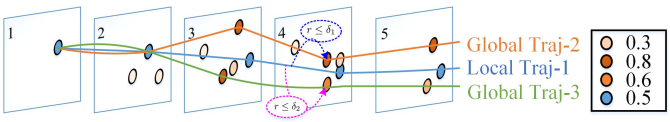



Fig. 5. Illustration of tracking by multi-trajectory selection. Best viewed by zooming in.

### *D. Multi-Trajectory Selection Network*

Based on aforementioned dynamic attention model, we explore the beam search strategy to replace greedy search policy for visual tracking. Specifically, we keep multiple different tracking results during tracking and find the best trajectory to approach robust tracking. In particular, in each frame, instead of simply selecting the location/proposal with the highest confidence score as the final tracking result, we keep record of multiple candidate tracking results in a *beam search* manner: first, we obtain the candidate regions from the global search module, i.e., the dynamic target-aware attention network. Then, we locate the target from these regions with the local search-based tracker and select the most reliable top-$k$ results as the current search results, obtaining $k$ potential trajectories, as shown in Fig. 4. Similar operations are executed for the subsequent video frames until the end of the video. After this procedure is completed, we measure the quality of tracking results in each frame with the trajectory selection network. A new trajectory with a maximum selection score will be chosen as the final tracking result of the current test video.

For example, suppose that a video has T frames and each frame has 3 candidates, we always maintain 3 trajectories for tracking, since introducing more candidates will cost excessively large computational time. As illustrated in Fig. 5, two threshold parameters $\delta_1$ and $\delta_2$ are used to measure the quality of current tracking results for 3 trajectories. We denote Traj-1 as the local search result. We also denote $r$ as the confidence score of the local search based tracker. For each frame, if $r \leq \delta_1$, the search region will be switched into global

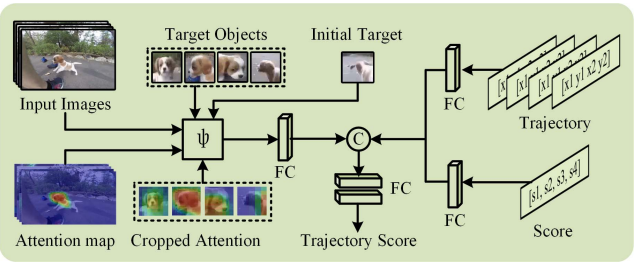



Fig. 6. Our proposed trajectory selection module.

attention region with best similarity (i.e., the dark orange points in Fig. 5, different colors means various confidence), therefore we have a new trajectory Traj-2. Meanwhile, if $r \leq \delta_2$, we search the target object from another global attention region if existed, therefore, we can attain the Traj-3. The Traj-1, 2, 3 then form the 3-trajectory search. This procedure shares a similar idea with beam search which is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set and widely used in image caption [42]. Our proposed multi-trajectory selection policy and beam search all attempt to maintain multiple candidates for final selection. It is easy to find that our tracker can conduct joint local and global search in an adaptive manner for robust tracking and our model can still work well on long-term videos.

After performing multi-trajectory based tracking through the whole video, as shown in Fig. 6, we take all the available information as the input of the multi-trajectory selection network to accurately predict the quality of each trajectory. The encoded information includes input video frames, initial target template, predicted attention map, the cropped tracking result and corresponding target-aware attention patch, the specific value of predicted bounding box, and its similarity score predicted by the Siamese tracker. For all the input patches, we resize them into $300 \times 300$ and then feed them into a residual network. By aggregating these features via concatenation and reshaping, we send the aggregated features into a fully connected layer for dimension reduction. For bounding boxes and scores, we utilize two fully connected layers to encode them into corresponding feature vectors, respectively. Then, these feature vectors will be concatenated and fed into two fully connected layers for regression. The ground truth for this regression task is the IoU (Intersection over Union) between the predicted trajectory and ground truth annotations. In our implementation, we adopt GIoU [43] which is an improved evaluation metric between two bounding boxes.

## IV. Experiments

### A. Dataset and Evaluation Metric

In this paper, we train our dynamic target-aware attention network on two tracking datasets: TLP [44] and DTB [45], which totally contain 120 video sequences. The trajectory selection network is trained on the training subset of GOT-10k dataset [46]. We test our model on several popular tracking benchmarks, including OTB-2015 [47], GOT-10k [46], OxUvA [26], LaSOT [48], VOT2018-LT [49], UAV123 [50] and UAV20L [50]. A brief introduction to these benchmark datasets are given below.

**OTB-2015 [47]** contains 100 video sequences and also defines 9 attributes such as *Illumination Variation, Scale Variation, Occlusion, Deformation*. It is one of the most widely used benchmark datasets for visual tracking since its release in 2015.

**GOT-10k [46]** is constructed based on the backbone of WordNet structure [51]. It populates the majority of over 560 classes of moving objects and 87 motion patterns. It contains 10,000 videos totally, with more than 1.5 million manually labeled bounding boxes. The authors select 280 videos as the test subset and the rest of videos are used for training.

**OxUvA [26]** is developed for the training and evaluation of long-term trackers. It comprises 366 sequences spanning 14 hours of video which can be categorized into 22 classes. This dataset is divided into train and testing subset, which contains 200 and 166 videos respectively.

**LaSOT [48]** is the currently the largest long-term tracking dataset which contains 1400 video sequences with more than 3.5M frames in total. The average video length is more than 2,500 frames and each video contains challenging factors deriving from the wild, e.g., out-of-view, scale variation. It provides both natural language and bounding box annotations which can be used for the explorations of integrating visual and natural language features for robust tracking. For the evaluation of LaSOT dataset, *Protocol I* employs all 1400 sequences for evaluation and *Protocol II* uses the testing subset of 280 videos.

**VOT2018-LT [49]** is a long-term dataset which contains 35 videos with a total length of 146817 frames. It is calculated that the target object will disappear for 12 times and each lasting on average 40 frames for each video.

**UAV123 [50]** and **UAV20L [50]** is an aerial video dataset which designed for low altitude UAV target tracking. It is consisted of 123 videos comprising more than 110K frames. They also provide a high-fidelity real-time visual tracking simulator for evaluation. The authors also merge these subsequences and pick the 20 longest sequences for long-term evaluation, also termed UAV20L.

For OTB-2015 [47], GOT-10k [46], LaSOT [48], UAV123 [50] and UAV20L [50], **Precision Plots** and **Success Plots** are adopted for the evaluation (also termed **PR** and **SR**). The first evaluation metric illustrates the percentage of frames where the center location error between the object location and ground truth is smaller than a predefined threshold (20-pixel threshold are usually adopted). The second one demonstrates the percentage of frames the Intersection over Union (IoU) of the predicted and the ground truth bounding boxes is higher than a given ratio. It is worthy to note that the **AO** is also adopted for the evaluation of GOT-10k [46] dataset. The AO denotes the average of overlaps between all ground truth and estimated bounding boxes. The VOT2018-LT [49] dataset adopts **Precision, Recall** and **F1-score** for the evaluation. Specifically, the definition

TABLE I
TRACKING RESULTS ON GOT-10K BENCHMARK

| **Tracker** | KCF [53] | SRDCF [54] | DAT [55] | MDNet [56] | ECO [57] | GOTURN [16] | SiamFC [15] |
|---|---|---|---|---|---|---|---|
| **AO** | 0.203 | 0.236 | 0.251 | 0.299 | 0.316 | 0.347 | 0.348 |
| **SR**$_{0.50}$ | 0.177 | 0.227 | 0.242 | 0.303 | 0.309 | 0.375 | 0.353 |
| **Tracker** | ATOM [58] | RT-MDNet [59] | THOR [20] | THOR+GS | THOR+BS-2T | THOR+BS-3T | THOR+BS-3T-TSN |
| **AO** | 0.547 | 0.342 | 0.447 | 0.453 | 0.458 | 0.461 | 0.462 |
| **SR**$_{0.50}$ | 0.628 | 0.356 | 0.538 | 0.545 | 0.550 | 0.554 | 0.556 |

of these metrics are:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

where TP, FP and FN are used to denote the True Positive, False Positive and False Negative, respectively. For the OxUvA [26] long-term tracking benchmark, the **TPR, TNR** and **MaxGM** are adopted for the evaluation. Specifically, the TPR gives the fraction of present objects that are reported *present* and correctly located, while TNR gives the fraction of absent objects that are reported *absent*. The MaxGM is a single measure of tracking performance which can be formulated as:

$$MaxGM = \max_{0 \le p \le 1} \sqrt{((1-p) * TPR)((1-p) * TNR + p)}. \tag{8}$$

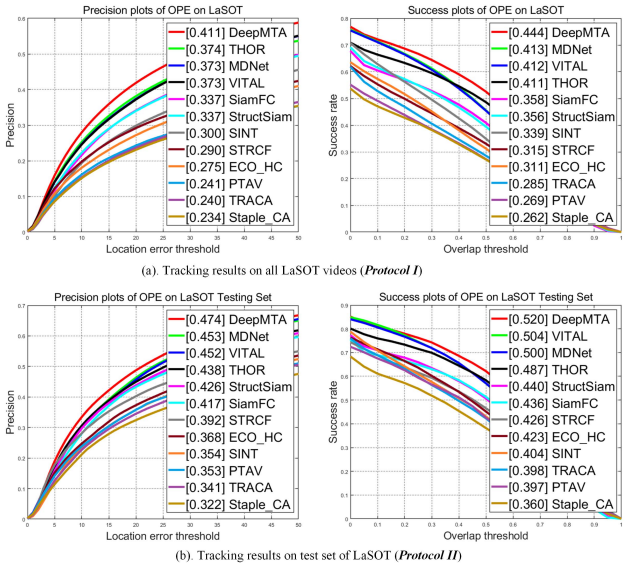
### B. Implementation Details

In this paper, we utilize binary cross-entropy loss to train the dynamic target-aware attention network. The ground truth mask is obtained from existing tracking datasets by whiting the target object and black the background regions as previous works do [4]–[6]. The batch size is 20, learning rate is $1e-4$. For the training of the trajectory selection network, we first run our target-aware attention guided THOR to collect the predicted trajectories and attention maps on selected $3k$ video sequences from the training subset of the GOT-10k dataset. After that, we train this network for 50 epochs on the collected dataset. The initial learning rate is 0.001, the batch size is 10 and the Adagrad [52] optimizer is selected to optimize these two networks. For each time step, we input 12 frames for the TSN network for the trajectory evaluation due to the limitation of our GPU. $\delta_1$ and $\delta_2$ are experimentally set as 0.5 and 0.6 in all the experiments. Our code is developed based on PyTorch and the experiments are conducted on a server with Ubuntu 16.04.3 LTS, Intel(R) Xeon(R) CPU E5-2620 v4@2.10GHz, GeForce RTX 2080. The source code is available at https://github.com/wangxiao5791509/DeepMTA_PyTorch.

### C. Comparison on Public Benchmarks

In this section, we will report the tracking results of our method and other trackers on LaSOT, GOT-10k, UAV20L, OTB-2015, UAV123 and OxUvA datasets, respectively. It is worthy to note that we utilize our **THOR+BS-3T** version (termed DeepMTA) to compare with other trackers, although higher performance can be obtained with more trajectories.

**LaSOT [48]:** As shown in Fig. 7 (a), our tracker achieve 0.411 and 0.444 on precision plots and success plots based on




(a). Tracking results on all LaSOT videos (***Protocol I***)

(b). Tracking results on test set of LaSOT (***Protocol II***)

Fig. 7. **Tracking results on LaSOT benchmarks.** DeepMTA is our proposed tracker in this paper.

the $ProtocolI$, which are significantly better than the baseline THOR (0.374/0.411) and other trackers. For the $ProtocolII$, our results are also better than these trackers as illustrated in Fig. 7 (b). This fully demonstrates the effectiveness of our proposed tracker.

**GOT-10k [46]:** As we can see from Table I, our tracker achieves better results than the baseline method THOR and also most of the other recent and popular trackers, such as MDNet and ECO.[1] Specifically, the baseline tracker THOR achieve 0.447, 0.538 on the $AO$ and $SR_{0.50}$, respectively, while our tracker which utilize the dynamic target-aware attention guided multi-trajectory tracking obtained better results (0.462 and 0.556 on the evaluation metric).

**OTB-2015 [47]:** As shown in Table II, THOR achieves 0.787/0.641 on the OTB-2015 dataset which are already better than most of the compared trackers, such as CFNet, SiamFC and Staple. Thanks to the dynamic target-aware attention, which can provide us global search regions for more robust tracking, we achieve 0.799/0.650 on this dataset. Although our overall performance is not better than some trackers (ECO: 0.910/0.691) on this dataset, however, our results are significantly better than these trackers on large-scale tracking benchmarks, such as LaSOT. We think this may be caused by overfitting of their tracker on this small and short-term dataset.

**UAV123 [50]:** As shown in Table II, the baseline tracker THOR achieves 0.758/0.697 on PR and SR, which is

[1]The related tracking results on GOT-10k are adopted from the leader board from http://got-10k.aitestunion.com/leaderboard.

TABLE II
TRACKING RESULTS ON OTB-2015, UAV123 AND UAV20L DATASET

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **OTB-2015** | CFNet | SiamFC | Staple | DSST | PTAV | ECO | CREST | THOR | Ours |
| **PR/SR** | 0.748/0.568 | 0.771/0.582 | 0.784/0.581 | 0.680/0.513 | 0.848/0.634 | 0.910/0.691 | 0.838/0.623 | 0.787/0.641 | 0.799/0.650 |
| **UAV123** | DSST | SAMF | SRDCF | ECO | SiamRPN | DaSiamRPN | RT-MDNet | THOR | Ours |
| **PR/SR** | 0.586/0.356 | 0.592/0.396 | 0.676/0.464 | 0.741/0.525 | 0.748/0.527 | 0.796/0.586 | 0.772/0.528 | 0.758/0.697 | 0.814/0.746 |
| **UAV20L** | PTAV | SiamFC | MUSTer | SRDCF | MEEM | fDSST | RT-MDNet | THOR | Ours |
| **PR/SR** | 0.624/0.423 | 0.626/0.403 | 0.514/0.329 | 0.507/0.343 | 0.482/0.295 | 0.422/0.300 | 0.583/0.461 | 0.533/0.436 | 0.715/0.570 |

TABLE III
TRACKING RESULTS ON OXUVA LONG-TERM BENCHMARK

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Tracker** | SPLT [3] | MBMD [9] | SiamFC+R [26] | TLD [25] | DaSiam LT [60] | LCT [8] | LTSINT [13] | MDNet [56] |
| **MaxGM** | 0.622 | 0.544 | 0.454 | 0.431 | 0.415 | 0.396 | 0.363 | 0.343 |
| **TPR** | 0.498 | 0.609 | 0.427 | 0.208 | 0.689 | 0.292 | 0.526 | 0.472 |
| **Tracker** | SINT [13] | ECO-HC [57] | SiamFC [15] | EBT [7] | BACF [61] | Staple [62] | THOR [20] | Ours |
| **MaxGM** | 0.326 | 0.314 | 0.313 | 0.283 | 0.281 | 0.261 | 0.320 | 0.340 |
| **TPR** | 0.426 | 0.395 | 0.391 | 0.321 | 0.316 | 0.273 | 0.410 | 0.463 |

comparable with existing trackers like RT-MDNet, SiamRPN and DaSiamRPN. Our tracker attains the best performance on this benchmark compared with these trackers, which are 0.814/0.746 on PR/SR, respectively. These experiments also fully demonstrate the effectiveness of our proposed dynamic target-aware attention guided multi-trajectory tracking framework.

**UAV20L [50]:** As shown in Table II, our tracker achieves 0.715/0.570 on the PR and SR, respectively, which are significantly better than the baseline tracker THOR and also some recent strong trackers like PTAV, RT-MDNet and SRDCF. This experiment fully validated the effectiveness of our proposed multi-trajectory tracking framework.

**OxUvA [26]:** As shown in Table III, we report the results on the test dataset of OxUvA which contains 166 long-term videos. The baseline method THOR achieves 0.320, 0.410 on MaxGM and TPR, while our tracker obtains 0.340 and 0.463.

### *D. Ablation Study*

In this section, the following notations are needed to be watchful to better understand our model. Specifically, the MDC is short for the modules of convolutional operators between the feature maps of target template and global images. PP denotes the point proposal module, GS is short for global search mechanism, and BS means the beam search scheme for robust tracking, i.e., the multi-trajectory analysis module.

*1) Analysis on Dynamic Target-Aware Attention:* To check the effectiveness of each component of our target-aware attention model, we implement the following component analysis:

1). ResNet+Concat: naive version for target-aware attention estimation. We directly concatenate the feature map of the target object and global image as previous work does [4].

2). ResNet+MDC: we conduct convolutional operations on multiple hierarchical feature maps, to check the effectiveness of interactions between target template and global images.

3). ResNet+MDC+PP: we integrate the point proposal to check the influence of spatial coordinates.

In this section, we utilize MAE (Mean Absolute Error), which is a widely used evaluation metric in the salient object detection community [63], [64], to measure the quality of predicted attention maps of each model. As shown in Table IV, the MAE of ResNet+Concat is 4.27, while the ResNet+MDC and ResNet+MDC+PP are 3.99 and 2.23, respectively. It is easy to find that the introduced convolutional operation and spatial coordinates are all contributed to the dynamic target-aware attention estimation. Some attention maps predicted with these models can be found in Fig. 8, and more results are provided in Fig. 12.

TABLE IV
MAE OF ATTENTION PREDICTION ON OTB-2015 DATASET

| Model | ResNet+Concat | ResNet+MDC | ResNet+MDC+PP |
|---|---|---|---|
| MAE | 4.27 | 3.99 | 2.23 |

*2) Analysis on Multi-Trajectory Tracking:* To check the effectiveness of our proposed multi-trajectory inference strategy, we conduct the following component analysis:

i). THOR: the baseline method used in this paper;

ii). THOR+GS: we introduce target-aware attention for global search and integrate with THOR to check the effectiveness of target-aware attention;

iii). THOR+BS-2T/3T: we utilize *double/triplet-trajectory* search strategy for component-ii to check the influence of multi-trajectory tracking;

iv). THOR+BS-3T-TSN: we use the trajectory selection network (TSN) for final trajectory selection to check the effectiveness of this module.

As shown in Table I, the baseline tracker THOR achieves 0.447/0.538 on AR and SR, respectively. When integrating our dynamic target-aware attention module into THOR, the tracking performance improved to 0.453/0.545, which fully demonstrates the effectiveness of dynamic target-aware attention for global search. We also conduct tracking based on our proposed multi-trajectory tracking framework and further improve the tracking results. For example, the THOR+BS-2T/3T improves the tracking results from 0.453/0.545 to 0.458/0.550 and 0.461/0.554, respectively. These experiments fully validated the effectiveness of our multi-trajectory tracking framework.


(a). Results with different version of our attention model

(b). More visualization of our final attention model (i.e. ResNet + MDC + PP)

Fig. 8. Attention maps generated with different attention models.

TABLE V
GENERALIZATION ANALYSIS ON THE GOT-10K AND VOT2018-LT DATASET

| **GOT-10k** | **SiamFC++** [65] | **Ours** | **MDNet** [56] | **Ours** | **SiamMask** [36] | **Ours** | **DiMP** [66] | **Ours** | **ATOM** [58] | **Ours** | **SiamRPN++** [2] | **Ours** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AO** | 0.604 | 0.609 | 0.299 | 0.392 | 0.451 | 0.461 | 0.673 | 0.674 | 0.547 | 0.557 | 0.453 | 0.473 |
| **SR** | 0.734 | 0.738 | 0.303 | 0.433 | 0.541 | 0.555 | 0.785 | 0.788 | 0.628 | 0.645 | 0.537 | 0.561 |
| **VOT2018LT** | **SiamFC++** [65] | **Ours** | **MDNet** [56] | **Ours** | **SiamMask** [36] | **Ours** | **DiMP** [66] | **Ours** | **ATOM** [58] | **Ours** | **SiamRPN++** [2] | **Ours** |
| **Precision** | 0.689 | 0.684 | 0.479 | 0.562 | 0.610 | 0.615 | 0.660 | 0.670 | 0.619 | 0.625 | 0.646 | 0.659 |
| **Recall** | 0.457 | 0.544 | 0.324 | 0.355 | 0.407 | 0.498 | 0.583 | 0.585 | 0.485 | 0.507 | 0.419 | 0.442 |
| **F1-score** | 0.549 | 0.606 | 0.387 | 0.435 | 0.488 | 0.550 | 0.619 | 0.625 | 0.544 | 0.560 | 0.508 | 0.529 |

It is also worthy to note that the aforementioned trackers are all select the final trajectory based on their response score only. However, the response score can not reflect the true results in some complex scenarios. The performance can be further improved by using our proposed trajectory selection network, which not only considers the response score from tracker but also the target-aware attention and consistency of the tracked target object. The results of THOR+BS-3T-TSN fully validated the effectiveness of this module.

From a methodological point of view, more trajectories mean more dynamics we can capture with our model. As shown in Table I, we can find that the tracking results can be further improved by introducing more trajectories. Specifically, the tracking results of 1, 2 and 3 trajectories on the GOT-10k dataset are 0.453/0.545, 0.458/0.550 and 0.461/0.554 which are consistent with our views. In addition, we also evaluate different trajectories (3, 4, 5 trajectories are tested) based on OTB-2015 dataset and we get 0.799/0.650, 0.799/0.650, 0.799/0.651, respectively. These experiments all demonstrate that beam search based multi-trajectory tracking achieves better results than regular greedy search strategy.

*3) Analysis on the Generalization:* In our experiments, we select the THOR as our baseline tracker for tracking. It is worthy to note that our proposed algorithm can also be integrated with other tracking algorithms due to it is a generic module. In this section, we test our module by integrating them with SiamFC++ [65], SiamMask [36], MDNet [56], SiamRPN++ [2], DiMP [66], and ATOM [58] on the GOT-10k and VOT2018LT dataset to check the generalization.

As shown in Table V, we can find that our module can improve all the baseline approaches on both datasets. Specifically, the SiamFC++ achieves 0.604/0.734 on the AO/SR respectively, while our method attain 0.609/0.738 on the GOT-10k dataset. We also improve the SiamMask from 0.451/0.541 to 0.461/0.555, the MDNet from 0.299/0.303 to 0.392/0.433. The improvement of SiamFC++ and SiamMask on the GOT-10k dataset is relatively little, due to the videos are short, therefore, there is little room for our module to improve final results. For the long-term benchmark VOT2018LT which contains 35 videos, we can find that the improvements are significant. More detail, the SiamFC++ achieves 0.689, 0.457, 0.549 on precision/recall/F1-score respectively, meanwhile, we improve these metrics to 0.684, 0.544, 0.606. For the minor decrease of SiamFC++ on the Precision, we think this maybe caused by the tradeoff between local search and global search mechanism. The experimental results based on MDNet, SiamMask, DiMP, ATOM, and SiamRPN++ trackers also demonstrate that our proposed modules are effective for tracking task, especially on the long-term video sequences.

TABLE VI
EFFICIENCY ANALYSIS OF OUR MODEL

| **Tracker** | **THOR-I** | **THOR-II** | **Ours-2** | **Ours-3** |
|---|---|---|---|---|
| **FPS** | 112 | 79.73 | 12.10 | 12.07 |

*4) Efficiency Analysis:* The baseline tracker THOR running at 112 FPS reported in their original paper (named as THOR-I in Table VI); while it running at 79.73 FPS on a Laptop with CPU Intel I7 and GPU NVIDIA RTX 2070 (i.e., the THOR-II in Table VI). This speed is tested on the whole OTB-2015 dataset. Our tracker can run at 12.10 and 12.07 FPS respectively when 2 and 3 trajectories are adopted. It is also worthy to note that this running time includes both the *multi-trajectory tracking* and *multi-trajectory selection*. We believe that our tracker can obtain better running efficiency, if better GPU is utilized, such as NVIDIA RTX 2080.

*5) Influence of Threshold Parameters:* In this work, two parameters $\delta_1$ and $\delta_2$ are very important for final tracking. We report the tracking results on OTB-2015 dataset with different settings in Table VII. Specifically speaking, we first

TABLE VII
RESULTS WITH DIFFERENT THRESHOLD PARAMETERS $\delta_1$ AND $\delta_2$ ON OTB-2015 DATASET

| $\delta_1$ | Baseline | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|
| PR | 0.787 | 0.799 | 0.799 | 0.796 | 0.795 | 0.795 | 0.786 | 0.769 |
| SR | 0.641 | 0.650 | 0.650 | 0.649 | 0.648 | 0.647 | 0.642 | 0.629 |
| $\delta_2$ | Baseline | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| PR | 0.787 | 0.791 | 0.794 | 0.796 | 0.799 | 0.798 | 0.795 | 0.788 |
| SR | 0.641 | 0.645 | 0.646 | 0.649 | 0.650 | 0.649 | 0.648 | 0.643 |

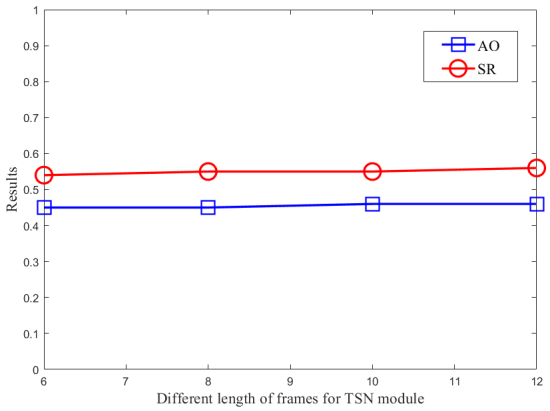



Fig. 9. Results of different number of frames for TSN module on GOT-10k dataset.

fix $\delta_2$ as 0.6, and test the performance with various values of $\delta_1$, i.e., from 0.3 to 0.9. We can find that the results will be better when the $\delta_1$ is 0.4. Then, we fix the $\delta_1$ as 0.4, and check the results with various values of $\delta_2$. Finally, it is easy to find that we can attain the best performance when $\delta_1$ and $\delta_2$ are set as 0.4 and 0.6. Meanwhile, we can also find that our tracker is not so sensitive to these two parameters. Because the results are relatively stable when the value changing from 0.3 to 0.7.

*6) Analysis on Different Number of Frames for TSN Module:* In our experiments, the TSN network is used for trajectory evaluation by processing the whole trajectory in a batch manner, i.e., we set the batch size as 12, due to the limited memory of our GPU. Much larger frame number can also be used, if the GPU meets the demand. To analyse the influence of this parameter, we also tested other values on GOT-10k dataset. From Fig. 9, we find that the overall results are relatively stable when different values are used, i.e., 6, 8, 10, 12.

*7) Attribute Analysis:* In this section, we report the results of our tracker and some state-of-the-art trackers on each attribute as shown in Fig. 10 (Success plots) and Fig. 11 (Precision plots). It is easy to find that our tracker achieves the best performance on most of attributes, including out-of-view, low resolution, aspect ratio change. These experiments fully validated the effectiveness and robustness of our tracker when facing challenging factors.

*8) Visualization:* In this section, we give some visualization of our attention maps and tracking results respectively.

*a) Attention maps:* As we can see from Fig. 12, our dynamic target-aware attention network can locate the attention regions which most related to the target object we want to track for each video. Our model show good robustness to *clutter background, heavy occlusion, motion blur* and *view rotation*, *etc*. Specifically speaking, our attention model can still locate the target object accurately in the clutter background, such as the *shark* and *person* in the first and second row, respectively. This fully validated the effectiveness of our convolutional operation based target-aware attention prediction. It is also worthy to note that our attention can reflect the occluded target object to some extent, such as the *shark* in the 0026 and 1054 frame at the first row, the *dog* in the 1030 at the fourth row. These amazing attention results are brought by our dynamic point proposal, which can provide spatial coordinate information for a more accurate target object location. Besides, the *fox* and *car* in the third and fifth row demonstrate that our dynamic attention model is robust to the view variation of the target object.

According to the aforementioned analysis, we can conclude that our dynamic attention model shows good robustness to challenging factors, such as motion blur, heavy occlusion, out-of-view, clutter background. The tracking results on each attribute on the LaSOT dataset also proved the robustness of our model, as shown in Section IV-D.7.

*b) Tracking results:* As shown in Fig. 13, we give some visualizations of our tracker and other state-of-the-art trackers on the LaSOT dataset. It is intuitive to find that our tracker is robust to challenging factors such as *out-of-view, clutter background, scale variation*. For example, the *flying kite* in the first row will become out of the view, when it occurred back, our tracker can still capture its location due to the utilize of joint local and global search scheme. However, many other trackers failed to locate the target object due to only the local search mechanism used in their procedure. This also demonstrates the importance of accurate prediction of dynamic target-aware attention maps.

For the second and third row, we can find that our tracker (red bounding box) can locate the target object more accurately than the baseline tracker THOR (green bounding box). This fully validates the effectiveness of our tracking algorithm. For the fourth row, we can find that our tracker can still work well in challenging scenarios, while many other trackers (including the baseline tracker) are easily influenced by the clutter background.

### E. Discussion

The baseline tracker THOR only use local search under tracking-by-detection framework; the target-aware attention model provides the global attention map which can be used for global search for baseline tracker. On the other hand, the baseline tracker can provides tracking results to dynamically modeling the target object in the target-aware attention module. Therefore, these two modules are complementary to each other. Tracking by switching between local and global search is an intuitive way for robust object tracking, however, this may still confuse trackers when challenging factors occurred. Therefore, the multi-trajectory analysis module is introduced

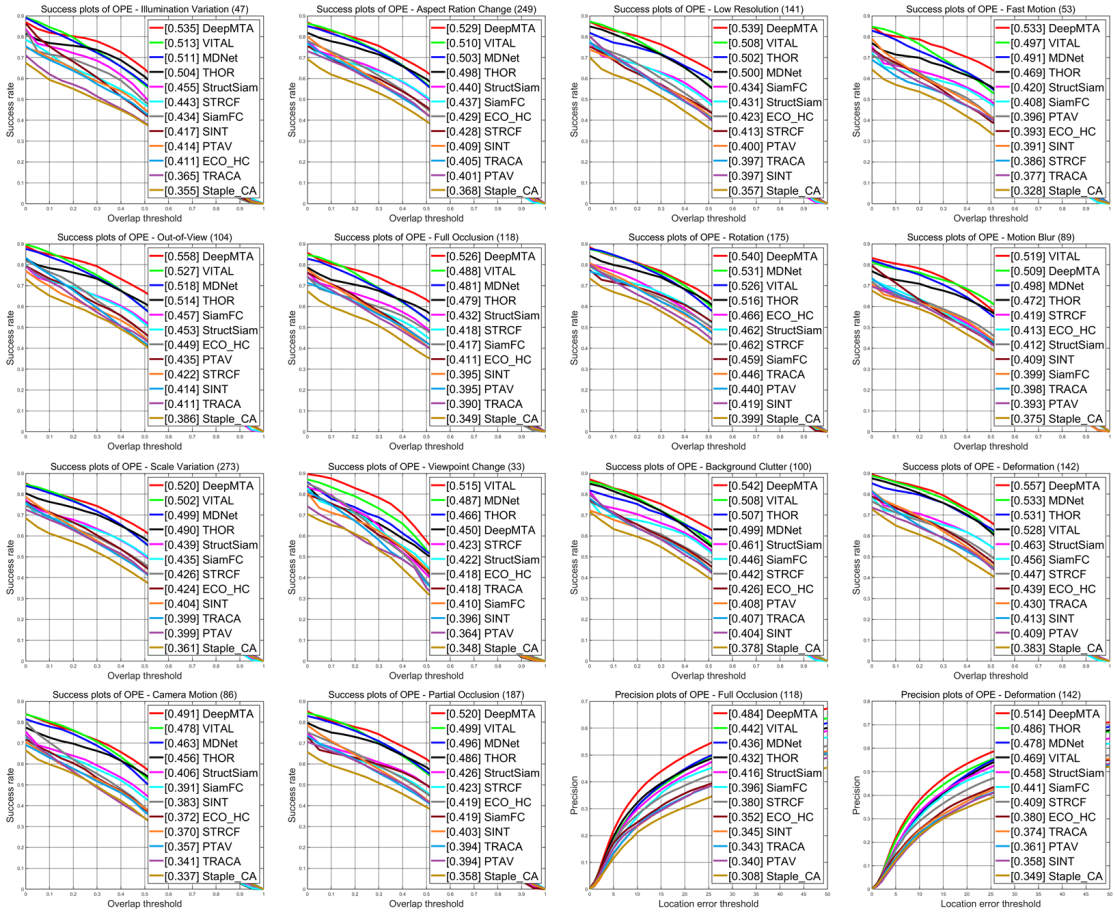



Fig. 10. **Success plots of attribute analysis on LaSOT benchmarks.** Two sub-figures from Precision plots are moved in this figure for aesthetics.

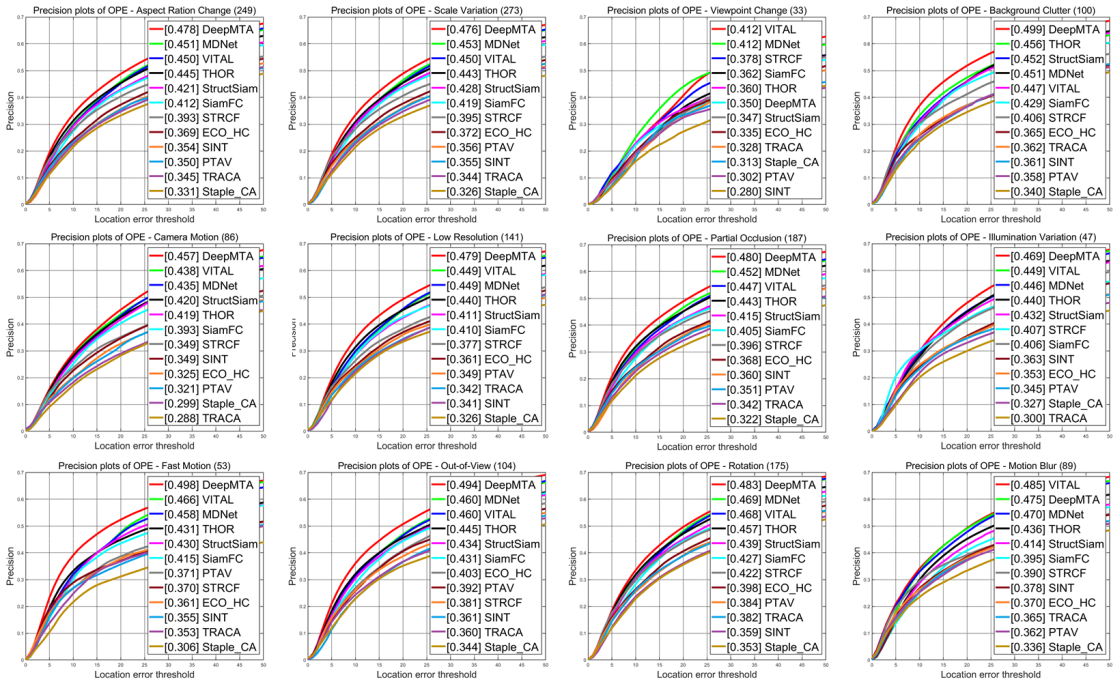



Fig. 11. Precision plots of attribute analysis on LaSOT benchmarks.

to collect multi-trajectories and conduct tracking by multi-trajectory selection. This procedure is mainly implemented by the trajectory selection network. It is also worthy to note that our algorithm tracking the target object in an *batch* manner which will be beneficial for specific applications, such as video analysis in sport and surveillance.

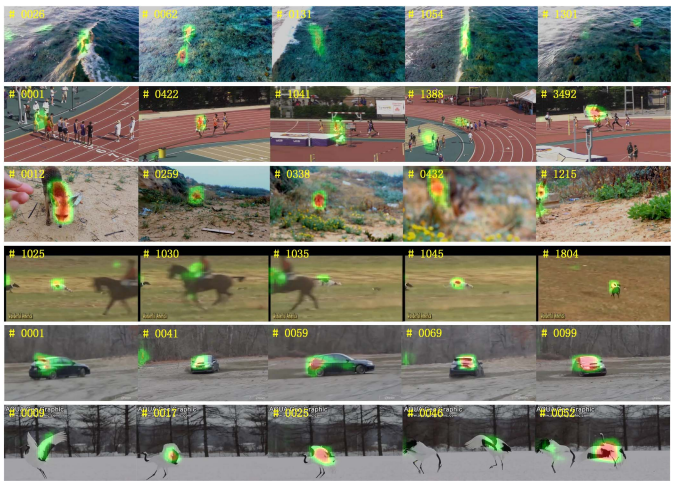


Fig. 12. Attention predicted by our dynamic target-aware attention network.

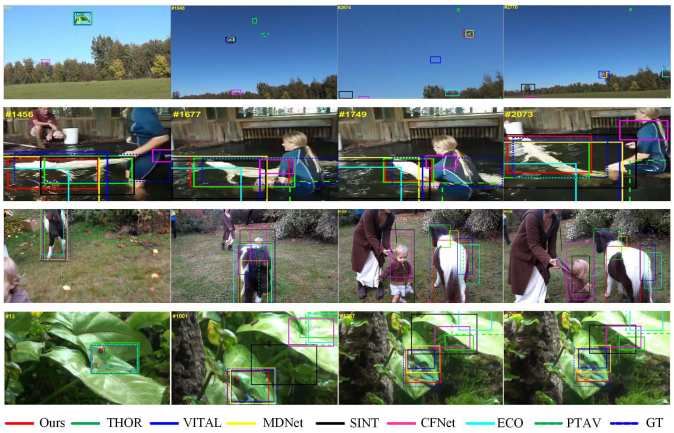



Fig. 13. Visualization of tracking results on videos from LaSOT benchmarks.

## V. Conclusion and Future Work

In this paper, we propose a novel multi-trajectory tracking framework, which significantly increased the dynamics of visual tracking. Specifically, we maintain multiple tracking results for each frame based on joint local and global search. To conduct a more accurate global search, we design a novel dynamic target-aware attention module which receives dynamic target templates and coordinates as condition and estimate target location from global views. After all the video frames are processed, we select the best trajectory with our proposed trajectory selection network, which considers multiple information, such as attention maps, response scores, and coordinates of BBox. Extensive experiments are conducted on multiple tracking dataset, including short-term and long-term tracking datasets.

In our implementation, we simply select the best-scored trajectory as our tracking result, but different trajectories may have their own good tracking result clips. How to design an efficient and effective trajectory fusion scheme to achieve better tracking performance is a worthy study problem. The efficiency of our tracker can also be improved by adaptively choosing the number of trajectories. In other words, limited trajectories are needed for simple videos and more trajectories can be employed for challenging videos. We will focus on these two issues in our future works.

## References

[1] Z. Zhang and H. Peng, "Deeper and wider siamese networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4591–4600.

[2] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4282–4291.

[3] B. Yan, H. Zhao, D. Wang, H. Lu, and X. Yang, "'Skimming-perusal' tracking: A framework for real-time and robust long-term tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2385–2393.

[4] X. Wang, C. Li, R. Yang, T. Zhang, J. Tang, and B. Luo, "Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking," 2018, *arXiv:1811.10014*. [Online]. Available: http://arxiv.org/abs/1811.10014

[5] X. Wang, T. Sun, R. Yang, and B. Luo, "Learning target-aware attention for robust tracking with conditional adversarial network," in *Proc. 30th Brit. Mach. Vis. Conf.*, 2019, p. 131.

[6] R. Yang, Y. Zhu, X. Wang, C. Li, and J. Tang, "Learning target-oriented dual attention for robust RGB-T tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3975–3979.

[7] G. Zhu, F. Porikli, and H. Li, "Beyond local search: Tracking objects everywhere with instance-specific proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 943–951.

[8] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5388–5396.

[9] Y. Zhang, D. Wang, L. Wang, J. Qi, and H. Lu, "Learning regression and verification networks for long-term visual tracking," 2018, *arXiv:1809.04320*. [Online]. Available: http://arxiv.org/abs/1809.04320

[10] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, "MUlti-store tracker (MUSTer): A cognitive psychology inspired approach to object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 749–758.

[11] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 118:1–118:36, Feb. 2019.

[12] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, Oct. 2018.

[13] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1420–1429.

[14] X. Wang, C. Li, B. Luo, and J. Tang, "SINT++: Robust visual tracking via adversarial positive instance generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4864–4873.

[15] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2016, pp. 850–865.

[16] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 fps with deep regression networks," in *Proc. 14th Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 749–765.

[17] H. Fan and H. Ling, "Siamese cascaded region proposal networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7952–7961.

[18] J. Gao, T. Zhang, and C. Xu, "Graph convolutional tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4649–4659.

[19] T. Yang and A. B. Chan, "Learning dynamic memory networks for object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 152–167.

[20] A. Sauer, E. Aljalbout, and S. Haddadin, "Tracking holistic object representations," in *Proc. BMVC*, 2019, pp. 1–12.

[21] P. Voigtlaender, J. Luiten, P. H. S. Torr, and B. Leibe, "Siam R-CNN: Visual tracking by re-detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6578–6588.

[22] G. Wang, C. Luo, X. Sun, Z. Xiong, and W. Zeng, "Tracking by instance detection: A meta-learning approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6288–6297.

[23] J. Choi, J. Kwon, and K. M. Lee, "Deep meta learning for real-time target-aware visual tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 911–920.

[24] Z. Chen, J. Li, Z. Chen, and X. You, "Generic pixel level object tracker using bi-channel fully convolutional network," in *Proc. 24th Int. Conf. Neural Inf. Process.* Guangzhou, China: Springer, Nov. 2017, pp. 666–676.

[25] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[26] J. Valmadre *et al.*, "Long-term tracking in the wild: A benchmark," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 670–685.

[27] G. Nebehay and R. Pflugfelder, "Clustering of static-adaptive correspondences for deformable object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2784–2791.

[28] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. 13th Eur. Conf. Comput. Vis.* Zürich, Switzerland: Springer, Sep. 2014, pp. 391–405.

[29] H. Fan and H. Ling, "Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5486–5494.

[30] L. Huang, X. Zhao, and K. Huang, "Globaltrack: A simple and strong baseline for long-term tracking," in *Proc. AAAI*, 2019, pp. 11037–11044.

[31] K. Dai, Y. Zhang, D. Wang, J. Li, H. Lu, and X. Yang, "High-performance long-term tracking with meta-updater," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6298–6307.

[32] D.-Y. Lee, J.-Y. Sim, and C.-S. Kim, "Multihypothesis trajectory analysis for robust visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5088–5096.

[33] S. Hare *et al.*, "Struck: Structured output tracking with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2096–2109, Oct. 2016.

[34] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4696–4704.

[35] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[36] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1328–1338.

[37] K. Sofiiuk, K. Sofiyuk, O. Barinova, A. Konushin, and O. Barinova, "AdaptIS: Adaptive instance selection network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7355–7363.

[38] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] D. Novotny, S. Albanie, D. Larlus, and A. Vedaldi, "Semi-convolutional operators for instance segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 86–102.

[41] R. Liu *et al.*, "An intriguing failing of convolutional neural networks and the CoordConv solution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9605–9616.

[42] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5561–5570.

[43] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[44] A. Moudgil and V. Gandhi, "Long-term visual object tracking benchmark," in *Proc. 14th Asian Conf. Comput. Vis.* Perth, WA, Australia: Springer, Dec. 2018, pp. 629–645.

[45] S. Li and D.-Y. Yeung, "Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models," in *Proc. 21st AAAI Conf. Artif. Intell.*, Feb. 2017, pp. 1–7.

[46] L. Huang, X. Zhao, and K. Huang, "GOT-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 4, 2019, doi: 10.1109/TPAMI.2019.2957464.

[47] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[48] H. Fan *et al.*, "LaSOT: A high-quality benchmark for large-scale single object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5374–5383.

[49] M. Kristan *et al.*, "The sixth visual object tracking VOT2018 challenge results," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–53.

[50] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. 14th Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 445–461.

[51] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[52] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.

[53] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[54] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1430–1438.

[55] S. Pu, Y. Song, C. Ma, H. Zhang, and M.-H. Yang, "Deep attentive tracking via reciprocative learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1931–1941.

[56] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.

[57] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6638–6646.

[58] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4660–4669.

[59] I. Jung, J. Son, M. Baek, and B. Han, "Real-time MDNet," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 83–98.

[60] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 101–117.

[61] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1135–1143.

[62] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1401–1409.

[63] Z. Chen, J. Zhang, and D. Tao, "Recursive context routing for object detection," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 142–160, 2020.

[64] Z. Chen, W. Ouyang, T. Liu, and D. Tao, "A shape transformation-based dataset augmentation framework for pedestrian detection," *Int. J. Comput. Vis.*, pp. 1–18, Jan. 2021, doi: 10.1007/s11263-020-01412-0.

[65] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines," in *Proc. AAAI*, 2020, pp. 12549–12556.

[66] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6182–6191.

**Xiao Wang** received the B.S. degree from West Anhui University, Luan, China, in 2013, and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 2019. He is currently a Post-Doctoral Researcher with the Peng Cheng Laboratory, Shenzhen, China. From 2015 to 2016, he was a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He also have a visiting at the UBTECH Sydney Artificial Intelligence Centre, the Faculty of Engineering, the University of Sydney, in 2019. His current research interests mainly about computer vision, machine learning, pattern recognition, and deep learning. He serves as a Reviewer for a number of journals and conferences, such as IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IJCV, PR, CVPR, ICCV, ECCV, and AAAI.

**Zhe Chen** (Member, IEEE) received the B.S. degree in computer science from the University of Science and Technology of China, Hefei, China, in 2014, and the Ph.D. degree with the UBTECH Sydney Artificial Intelligence Centre, Faculty of Engineering, The University of Sydney, in 2019. His research interests include object detection, computer vision, and deep learning. His studies were published in IEEE CVPR, ICONIP, ECCV, and JAS. He also serves as a Reviewer for a number of journals and conferences, such as IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), and IEEE TRANSACTIONS ON CYBERNETICS (T-CYB).

**Jin Tang** received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor with the School of Computer Science and Technology, Anhui University. His current research interests include computer vision, pattern recognition, and machine learning.

**Bin Luo** (Senior Member, IEEE) received the B.Eng. degree in electronics and the M.Eng. degree in computer science from Anhui University, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, U.K., in 2002. He is currently a Professor with Anhui University. He also chairs the IEEE Hefei Subsection. He has published more than 200 papers in journals and refereed conferences. He served as a Peer Reviewer for international academic journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PAMI), *Pattern Recognition*, and *Pattern Recognition Letters*. His current research interests include random graph-based pattern recognition, image and graph matching, and spectral analysis.

**Yaowei Wang** (Member, IEEE) received the Ph.D. degree in computer science from the Graduate University of Chinese Academy of Sciences, in 2005.

He was a Professor with the National Engineering Laboratory for Video Technology Shenzhen (NELVT), Peking University Shenzhen Graduate School, in 2019. From 2014 to 2015, he worked as an academic Visitor with the Vision Laboratory, Queen Mary University of London. He worked with the Department of Electronics Engineering, Beijing Institute of Technology, from 2005 to 2019. He is currently an Associate Professor with the Peng Cheng Laboratory, Shenzhen, China. He is the author or coauthor of more than 70 refereed journals and conference papers. His research interests include machine learning, multimedia content analysis, and understanding. He was a recipient of the Second Prize of the National Technology Invention in 2017 and the First Prize of the CIE Technology Invention in 2015. His team was ranked as one of the best performers in the TRECVID CCD/SED tasks from 2009 to 2012 and in PETS 2012. He is a member of CIE, CCF, and CSIG.

**Yonghong Tian** (Senior Member, IEEE) is currently a Boya Distinguished Professor with the Department of Computer Science and Technology, Peking University, China, and also the Deputy Director of the Artificial Intelligence Research Center, PengCheng Laboratory, Shenzhen, China. He is the author or coauthor of more than 200 technical articles in refereed journals, such as IEEE TPAMI/TNNLS/TIP/TMM/TCSVT/TKDE/TPDS, ACM CSUR/TOIS/TOMM, and conferences, such as NeurIPS/CVPR/ICCV/AAAI/ACMMM/WWW. His research interests include neuromorphic vision, brain-inspired computation, and multimedia big data. He is a Senior Member of CIE and CCF, a member of ACM. He was/is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) since January 2018, IEEE TRANSACTIONS ON MULTIMEDIA (TMM) from August 2014 to August 2018, *IEEE Multimedia Magazine* since January 2018, and IEEE ACCESS since January 2017. He co-initiated IEEE International Conference on Multimedia Big Data (BigMM) and served as the TPC Co-Chair for BigMM 2015, and also served as the Technical Program Co-Chair for IEEE ICME 2015, IEEE ISM 2015, and IEEE MIPR 2018/2019, and the General Co-Chair for IEEE MIPR 2020 and ICME2021. He has been a Steering Member of IEEE ICME since 2018 and IEEE BigMM since 2015, and is a TPC Member of more than ten conferences, such as CVPR, ICCV, ACM KDD, AAAI, ACM MM, and ECCV. He was a recipient of the Chinese National Science Foundation for Distinguished Young Scholars in 2018, two National Science and Technology awards and three ministerial-level awards in China, and obtained the 2015 EURASIP Best Paper Award for Journal on Image and Video Processing, and the best paper award of IEEE BigMM 2018.

**Feng Wu** (Fellow, IEEE) received the B.S. degree in electrical engineering from Xidian University, in 1992, and the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, in 1996 and 1999, respectively. He is currently a Professor with the University of Science and Technology of China. Before that, he was a Principle Researcher and the Research Manager with Microsoft Research Asia. He has authored or coauthored more than 200 high-quality papers (including several dozens of IEEE TRANSACTION papers) and top conference papers on MOBICOM, SIGIR, CVPR, and ACM MM. He has 77 granted U.S. patents. His 15 techniques have been adopted into international video coding standards. His research interests include image and video compression, media communication, and media analysis and synthesis. As a coauthor, he got the best paper award in IEEE T-CSVT 2009, PCM 2008, and SPIE VCIP 2007. He got the IEEE Circuits and Systems Society 2012 Best Associate Editor Award. He also serves as the TPC Chair for MMSP 2011, VCIP 2010, and PCM 2009, and the Special sessions Chair for ICME 2010 and ISCAS 2013. He serves as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEM FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, and several other International journals.