AFAT: Adaptive Failure-Aware Tracker for Robust Visual Object Tracking

Tianyang Xu1 Zhen-Hua Feng1 Xiao-Jun Wu2 Josef Kittler1

tianyang.xu@surrey.ac.uk, z.feng@surrey.ac.uk,
wu_xiaojun@jiangnan.edu.cn, j.kittler@surrey.ac.uk

1 Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK
2 School of Internet of Things Engineering, Jiangnan University, Wuxi, China

Abstract. Siamese approaches have achieved promising performance in visual object tracking recently. The key to the success of Siamese trackers is to learn appearance-invariant feature embedding functions via pair-wise offline training on large-scale video datasets. However, the Siamese paradigm uses one-shot learning to model the online tracking task, which impedes online adaptation in the tracking process. Additionally, the uncertainty of an online tracking response is not measured, leading to the problem of ignoring potential failures. In this paper, we advocate online adaptation in the tracking stage. To this end, we propose a failure-aware system, realised by a Quality Prediction Network (QPN), based on convolutional and LSTM modules in the decision stage, enabling online reporting of potential tracking failures. Specifically, sequential response maps from previous successive frames as well as current frame are collected to predict the tracking confidence, realising spatio-temporal fusion in the decision level. In addition, we further provide an Adaptive Failure-Aware Tracker (AFAT) by combing the state-of-the-art Siamese trackers with our system. The experimental results obtained on standard benchmarking datasets demonstrate the effectiveness of the proposed failure-aware system and the merits of our AFAT tracker, with outstanding and balanced performance in both accuracy and speed.

Keywords: Visual Object Tracking, Convolutional Neural Network, Long Short-Term Memory, Online Adaptation

1 Introduction

Visual object tracking is a fundamental computer vision topic with the aim of automatically predicting the state of a target that is initialised with ground truth only in the first frame. With the exploration of mathematical modelling techniques, especially deep neural networks, recent advanced tracking approaches focus on employing large labelled video datasets to offline train end-to-end networks. The performance of these end-to-end deep Siamese neural networks, e.g., SINT [38], SiameseFC [3], SiamRPN [27], Siam R-CNN [40], has witnessed a continuous improvement with the help of more effective network architectures...
and more public available labelled video datasets. One typical property of the Siamese paradigm is that it considers an online tracking task as a pair-wise matching problem, where one-shot learning provides suitable mechanistic explanations. However, in the tracking stage, Siamese approaches fix the initial template model for the entire video sequences without online adaptation, leading to potential tracking shifts or failures.

In general, pair-wise offline training of Siamese networks can achieve effective feature embedding. However, it overemphasises robust appearance feature extraction for specific target categories in the training dataset [26]. Therefore, the learned model enables high tracking accuracy in the test videos where the static and dynamic appearance variations are similar to those found in the training stage. However, videos from practical surveillance and mobile devices are of high diversity with countless object categories, not to mention the corresponding variations. These videos, exhibiting unpredictable appearance content, raise challenges for one-shot learning approaches.

To mitigate the above issues, we argue that it is essential to incorporate an online alarm mechanism (failure detection) in the tracking process. We propose to integrate an online quality prediction of the tracking results at the decision level, as shown in Fig. 1. The response maps from recent successive frames are used as an input to assess the uncertainty of the tracking result of the current frame. It should be noted that the use of response maps has been widely studied in visual tracking for a variety of goals, ranging from basic target localisation to additional tracking bonus, e.g., quality assessment [41, 30, 5], re-detection condition [31, 50, 48] and update requirement [7, 4]. However, existing approaches only use hand-crafted functions or a pre-defined threshold to tap the power of response maps, neglecting their diverse variations in complex tracking situations. In contrast, we employ convolutional and Long Short-Term Memory (LSTM) modules to extract spatio-temporal fusion features from online generated response maps to perform failure flagging. Specifically, the response maps are considered as high-level representations, reflecting the probability of each position becoming the target centre. We show that by extracting implicit decision information from multi-frame response maps via convolutional and LSTM modules enables a reliable and robust prediction of potential tracking failures in a spatio-temporal perspective.

In the data collection stage, we perform online tracking on random video segments from video datasets (TrackingNet [34] in our experiment). The response maps and corresponding quality states are collected and stored as training samples for our Quality Prediction Network (QPN). We assign the quality state for each frame based on the Intersection Over Union (IOU) measure between the predicted bounding box and the corresponding ground truth. In the offline training phase, the labelled response map sequences are fed into QPN with a cross entropy loss. As temporal order and 2D response scores are jointly considered in network training, adaptive prediction of the tracking quality can be achieved by our QPN. A forward pass of QPN takes only 1.2~1.5 ms, which can be easily integrated with existing methods. Further more, we combine QPN with the state-
Fig. 1. Illustration of the proposed adaptive failure-aware tracker. AFAT achieves online quality prediction based on spatio-temporal modelling in the decision level, enabling failure alarm and further correction. AFAT detects 5 failures in the above sequence (Bird2), i.e., #11, #48, #63, #66, #69, corresponding to severe appearance changes with lower IOU between the tracking result and ground truth. #11 and #48 are true positive alarm signals with the overlap falling rapidly. #63, #66 and #69 are false positive alarm signals as they are close to the previous valley point #61.

of-the-art Siamese trackers for robust online visual object tracking, providing an effective failure detection and further performance boosting.

The main contributions of the proposed AFAT method include:

– A new quality prediction network (QPN) for generating online tracking alarm signal, and predicting potential failures in the tracking process, achieving spatio-temporal responses fusion at the decision level. The temporal order and spatial appearance of multi-frame response maps are jointly considered to predict the reliability of the tracking results, providing additional functionality to existing one-shot learning algorithms.

– A new adaptive failure-aware tracker equipped with advanced Siamese trackers and QPN, enabling synchronous adaptive model selection in online object tracking, with a perceptual model configuration and tracking correction.

– A comprehensive evaluation of our AFAT method on a number of well-known benchmarking datasets, including OTB2015 [44], UAV123 [32], LaSOT [12], VOT2016 [21], VOT2018 [22] and VOT2019 [24]. The results demonstrate the superior performance of the proposed AFAT method over the state-of-the-art trackers.
The rest of this paper is organised as follows: In Section 2, we provide the motivation for introducing a failure-aware mechanism into tracking and discuss the related studies in recent development. A detailed construction and training scheme of the proposed quality prediction network are presented in Section 3 and furnished with an efficient adaptive failure-aware tracker in Section 4. The experimental results and the corresponding analysis are reported in Section 5. Last, the conclusions are drawn in Section 6.

2 Background

2.1 Advanced Tracking models

Recent studies in visual object tracking focus on modelling the task as online discriminative learning or offline pair-wise matching network, both achieving continuously developed performance in standard benchmarks [44, 29, 32, 12, 34] and competitions [22, 24].

Online discriminative learning was designed to distinguish the target region from its surroundings via online training and updating a classifier (or regressor). To alleviate the computational burden for online learning, Discriminative Correlation Filter (DCF) based approaches have extensively studied in recent years, with advanced efficiency and accuracy contributed by the circulant structure [17, 18, 45]. Meanwhile, DCF-based methods address the sample sufficiency issue in online discriminative learning by augmented circulant samples. To optimise the estimate filters with more intuitive requirements and desired constraints, further attempts have been performed in the DCF paradigm, e.g., spatial regularisation for boundary effect [9, 30, 20, 49], feature selection for input redundancy [17, 46], temporal sampling space for filter degeneration [10, 11, 8], and multi-response fusion for extended representation [24].

Offline pair-wise matching network benefits from the design of the Siamese network architecture. Specifically, a Siamese network maps the image pairs, i.e., template and instance, into a latent feature space, where the target regions are of salient similarity against practical appearance variations. After feature embedding, SINT [38] compares the scores of candidates for final decision making, and GOTURN [16] performs bounding box localisation using FC layers. To avoid multi-candidate calculation and FC layer learning, cross correlation is directly used for response calculation in SiameseFC [3], realising a fully convolutional structure. Besides the effective capability in centre localisation, the accuracy of the final bounding box is also essential for advanced tracking methods. To this end, recent efforts have been made on fusing bounding box refinement techniques in the offline training stage. SiamRPN utilises the Region Proposal Network (RPN) for joint high-quality foreground-background classification and bounding box regression learning [27]. While ATOM performs IOU maximisation for fine-grained overlap precision [17]. In addition, to improve the exploitation from diverse multi-layer features and Siamese layers, SiamRPN++ [26] and SiamDW [52] are proposed, which demonstrate promising performance for visual object tracking.
Despite the success in designing powerful tracking models, the significance of online tracking quality prediction and the corresponding countermeasures are underestimated by the community. In other words, the pattern of the tracking calculation process is predefined by a specified mathematical model, which always assumes specific distributions of the target appearance or motion trajectory. We argue the necessity of online predicting the tracking quality and performing effective intervention and correction.

2.2 Online tracking adaptation

Though continuous improvements have been achieved in developing diverse mathematical formulations in visual tracking, a practical tracking task can not be exactly modelled by existing methodologies accurately. To mitigate this issue, online controlling strategies are necessary to balance the characteristics of the theories reflecting simplified assumptions about the universe of target tracking and their validity in challenging video sequences with unpredictable appearance variations.

To achieve the above goal, many tracking approaches propose to analyse the reliability of the current result for potential update and re-detection. To alleviate tracking shift caused by radical update, ATOM [7] and Dimp [4] adjust the learning rate based on the response score. TLD [19] utilised an independent classifier to re-detect the target with a distance threshold-based failure detector. Similar threshold-based quality measurement is widely used in recent online discriminative learning and Siamese trackers, e.g., LCT [31], SPLT [50] and Siamese-LT [26], performing re-detection when the response score is under a threshold. Besides, quality prediction also achieved wide attention for online adaptation in visual tracking. LMCF [41] utilised the response map for multi-modal target detection to reduce the impact from similar appearance. Then, a quality function is designed for high-confidence update, suppressing the contaminated samples. Other attempts in quality prediction focus on adaptive multi-response fusion. CSR-DCF [30] assigns more reliability to the single modal response channels and UPDT [5] simultaneously constrains the fusion response to be narrow and uni-modal distributed.

However, the above processing techniques only use a simple threshold or hand-crafted function to measure the quality of the current tracking result, which are unable to satisfy the practical tracking challenges. Therefore, we propose to employ neural network to predict the tracking quality. To better extract the potential information from response maps, we construct a spatio-temporal quality prediction network (QPN) with convolutional and LSTM modules. Different from existing measurement, we collect the response maps from successive multiple frames to enhance the sequential order. By training from the generated labelled data, QPN enables adaptive classification between successful and failed result in the decision level.
2.3 Meta-learning for tracking

From the perspective of learning framework, Siamese trackers belong to one-shot learning category, using meta-learning formulation to learn the best embedding network that highlights the target region while suppresses the background. In this paradigm, Learnet [1] is the seminal work using meta-learning to generate parameters for visual tracking. Besides predicting the tracking parameters, Meta-tracker [35] proposed to jointly learn the learning rate. While UpdateNet [51] proposed to replace the handcrafted update function with a method which learns to update. In addition, MLT [6] utilised a meta-learner network to provide the Siamese network with new appearance of the target. To explore the potential power of meta-learning, we propose to train a quality prediction network for a given tracker.

3 Quality prediction network

In this section, we present the proposed Quality Prediction Network (QPN). As shown in Fig. 2, the proposed QPN consists of convolutional and Long Short-Term Memory (LSTM) modules. Sequential response maps and the corresponding quality labels are used to train QPN in an end-to-end fashion. QPN utilises the current and previous temporal information to provide quality feedback, i.e.,
whether the current tracking result is successful or not. The forward pass of QPN can be formulated as:

\[ q_t = QPN(f_{t-K+1}, f_{t-K+2}, \ldots, f_t), \]

where \( f_i \in \mathbb{R}^{C \times N \times N} \) is the response maps of the \( i \)-th frame. We select SiameseRPN++ [26] (with mobilev2 backbone) as our base tracker, which can produce the response map of \( 5 \times 25 \times 25 \) size in each frame, where 5 and \( 25 \times 25 \) are corresponding to the anchor number and spatial resolution, respectively. \( K \) is the sequential length for our spatio-temporal inference. We set \( K = 20 \) in our experiment. We describe the construction of the QPN with more details as follows.

### 3.1 Spatial feature extraction

Different from raw image patches, we use response maps as the input data for our QPN. Response maps are output by an online tracker, reflecting the decision-making ability of the model in the current tracking situation. Similar to an image, considering the spatial distribution, we propose to use CNN for basic spatial feature extraction from response maps. Based on existing studies and early analysis, we have the following observations: (1) an ideal response map should be uni-modal with narrow peak; and (2) a tracking failure always occurs with clutters in the response map. To make the convolutional network suitable for our response input, we circularly shift the response map and scatter the maximal peak into the four corners, enhancing the context information. Then three convolutional layers and one FC layer are employed to extract the representations \( \phi(f_i) \) for each frame \( i \) in the response map sequence, where \( \phi \) denotes the forward pass of the CNN for spatial feature extraction.

### 3.2 Temporal feature fusion

Existing quality measurements in advanced tracking approaches only consider the response maps obtained by the current frame, neglecting the previous temporal information. In general, a basic assumption in visual tracking is that the target always smoothly changes in appearance and position between adjacent frames. Therefore, we argue the significance of exploiting temporal ordered data for tracking quality prediction. To this aim, we propose to fuse the sequential response maps from recent successive frames for quality prediction using two LSTM networks [13]. The extracted spatial features \( \Phi = \{\phi(f_{t-K+1}), \phi(f_{t-K+2}), \ldots, \phi(f_t)\} \) are fed into the two LSTM modules to predict the quality of the tracking result in frame \( t \), \( q_t = g(\Phi) \).

### 3.3 End-to-end training of QPN

To generate sequential response maps for training our QPN, we run visual tracking using our base tracker (SiameseRPN++, with mobilev2 backbone) on TrackingNet [34]. There are 30132 labelled video sequences in TrackingNet, and we
select the first 20000 videos to generate the training data, and select the rest 10132 videos for validation. Specifically, for each piece of data, we collected the response maps via running the base tracker on successive 20 frames from a randomly selected video. In addition, we label the response maps based on the IOU between the predicted bounding box and the corresponding ground truth for each frame,

\[
y_i = \begin{cases} 
  \text{success}, & \text{if } \text{IOU} > 0.5, \\
  \text{lost}, & \text{if } \text{IOU} < 0.1, \\
  \text{unassigned}, & \text{others.}
\end{cases}
\] (2)

The detailed number of our generated training samples is listed in Table 1.

During the training stage of QPN, sequential response maps \((f_i, y_i)\) are fed into the network, where \(f_i \in \mathbb{R}^{20 \times 5 \times 25 \times 25}\) and \(y_i \in \mathbb{R}\). We train QPN from scratch using the Stochastic Gradient Descent (SGD) optimiser. For network optimisation, the Cross-entropy loss function is used in our design. As shown in Table 1, the training data is imbalanced between the success and lost categories. To balance the volume between these two types of training data, we set the loss weight as 0.002 and 1 for success and lost, respectively. The hidden state and cell state in the two LSTM networks are initialised with zeros for each forward pass. We train the network for 25 epochs in total. The learning rate is decreased from 0.01 to 0.001 since the 16-th epoch. We report the detailed loss and accuracy in each epoch in Fig. 3. During the training stage, QPN converges with the accuracy of 0.917 and 0.904 on the validation set for success and lost, respectively. This demonstrates the effectiveness of the proposed QPN method in failure detection for an online object tracker. In addition, a forward pass of QPN takes only 1.2~1.5 ms, which can be easily integrated with existing methods.
Algorithm 1 AFAT tracking algorithm.

**Initialisation:** Initialise SiameseRPN++$_{mobilev2}$ and SiameseRPN++$_{resnet50}$ with the ground truth in the initial frame.

**Input:** Image frame $I_t$, model SiameseRPN++$_{mobilev2}$ and SiameseRPN++$_{resnet50}$, response list $I_t$, target centre coordinate $p_{t-1}$ and scale size $w \times h$ from frame $t-1$.

**Tracking:**
1. Extract search windows from $I_t$ at $p_{t-1}$;
2. Calculate response maps $f$ using SiameseRPN++$_{mobilev2}$;
3. Add $f$ into $I_t$ if the length of $I_t$ exceeds 20, remove the early observations;
4. Record current tracking result from model SiameseRPN++$_{mobilev2}$;
5. Forward pass QPN and obtain the quality prediction $q_t$;
6. If $q_t = \text{lost}$, change current tracking result using model SiameseRPN++$_{resnet50}$ and empty $I_t$.

**Output:** Target bounding box (centre coordinate $p_t$ and current scale size $w \times h$).

4 Adaptive failure-aware tracker

We propose the Adaptive Failure-Aware Tracker (AFAT) based on the SiameseRPN++ tracker [26] and the proposed QPN in Section 3. The tracking framework is summarised in Algorithm 1. We directly use two backbones, i.e., mobilev2 and resnet50, in our AFAT. We employ SiameseRPN++$_{mobilev2}$ as the base tracker and SiameseRPN++$_{resnet50}$ as the correction tracker. Without bells and whistles, the default parameters from the original SiameseRPN++ are used in the proposed AFAT method.

5 Evaluation

5.1 Implementation details and experimental settings

We implement our AFAT using PyTorch 1.0.1 on a platform with an Intel(R) Xeon(R) Gold 6134 CPU and three NVIDIA GeForce RTX 2080Ti GPU cards. For speed test, we only use a single GPU card. The code will be made publicly available. We set $K = 20$ for temporal sequential data collection. We use the same parameters as SiameseRPN++ [26] for the backbone networks, mobilev2 and resnet50, released by the original authors for all the experiments (no additional optimised parameters are used). We compare the proposed AFAT on several benchmarks, including LaSOT [12], UAV123 [32], OTB2015 [44], VOT2016 [21], VOT2018 [22] and VOT2019 [24], with a number of state-of-the-art trackers, i.e., ATOM [7], Meta-Tracker [35], SiamRPN++ [26], DaSiam [53], SiamRPN [27], LADCF [47], CREST [36], BACF [20], CFNet [39], CACF [33], CSR-DCF [30], C-COT [11], Staple [2], SiameseFC [3], SRDCF [9] and other top-ranking trackers in VOT challenges.

To measure the tracking performance, we follow the corresponding protocols [44, 23, 25]. We use precision plot and success plot [43] for OTB2015, UAV123 and LaSOT. Specifically, the precision plot measures the percentage of frames
Table 2. Ablation study of the proposed AFAT on VOT2016, VOT2018, VOT2019, OTB2015, UAV123 and LaSOT.

|             | EAO  | A   | FPS | EAO  | A   | FPS | EAO  | A   | FPS |
|-------------|------|-----|-----|------|-----|-----|------|-----|-----|
| **SiamRPN++** | 0.455 | 0.624 | 96.4 | 0.410 | 0.586 | 95.8 | 0.291 | 0.579 | 93.0 |
| **mobilev2** | 0.464 | 0.642 | 38.7 | 0.414 | 0.600 | 37.9 | 0.287 | 0.594 | 38.2 |
| **AFAT**     | 0.482 | 0.642 | 74.0 | 0.419 | 0.605 | 70.5 | 0.295 | 0.599 | 70.9 |

|             | Failure detections | Total frames | Failure detection rate |
|-------------|--------------------|--------------|-----------------------|
| **SiamRPN++** | 6632               | 21455        | 30.9%                 |
| **mobilev2**  | 7007               | 21356        | 32.8%                 |
| **AFAT**      | 6814               | 19935        | 34.2%                 |

|             | EAO  | A   | FPS | EAO  | A   | FPS | EAO  | A   | FPS |
|-------------|------|-----|-----|------|-----|-----|------|-----|-----|
| **SiamRPN++** | 0.658 | 0.864 | 93.2 | 0.602 | 0.802 | 95.1 | 0.450 | 0.525 | 95.2 |
| **resnet50**  | 0.663 | 0.875 | 38.5 | 0.611 | 0.804 | 37.9 | 0.496 | 0.571 | 38.9 |
| **AFAT**      | 0.663 | 0.874 | 86.8 | 0.612 | 0.811 | 80.4 | 0.492 | 0.574 | 77.7 |

|             | Failure detections | Total frames | Failure detection rate |
|-------------|--------------------|--------------|-----------------------|
| **SiamRPN++** | 2881               | 59035        | 4.88%                 |
| **resnet50**  | 4513               | 112578       | 4.01%                 |
| **AFAT**      | 57192              | 685360       | 8.34%                 |

with the distance between the tracking results and ground truth less than a certain number of pixels. The success plot measures the proportion of successful frames with the threshold ranging from 0 to 1 (a result is considered successful if the overlap of the predicted and ground-truth bounding boxes exceeds a pre-defined threshold). Three numerical values, i.e. distance precision (DP), normalised precision (NP) and area under curve (AUC), are further employed to measure the performance. DP is the corresponding precision plot value (illustrated in the legend of precision plot) when the threshold set to 20 pixels. AUC is the expected success rate (illustrated in the legend of success plot) in terms of overlap evaluation. For all the VOT (2016, 2018 and 2019) datasets, we employ expected average overlap (EAO), accuracy (A) value and robustness (R) to evaluate the performance [23].

5.2 Ablation study

To validate the effectiveness of the proposed AFAT, we perform ablation studies on recent public datasets including VOT2016, VOT2018, VOT2019, OTB2015, UAV123 and LaSOT. As reported in Table 2, the proposed AFAT achieves improved performance compared to the base tracker, SiamRPN++mobilev2, with light sacrifice on speed. Specifically, though SiamRPN++resnet50 is employed as a correction tracker in our design, AFAT outperforms SiamRPN++resnet50 on EAO and A on all the three VOT datasets. Due to the failure-aware system, some challenging frames are recognised by our QPN with follow-up corrections, alleviating the possibility of being judged as failures under VOT su-
Fig. 4. Illustration of the detected failure alarms during online tracking by our AFAT in sequences DragonBoy, Couple, Bolt2, Human9, Gym, Jogging2, MotorRolling, and Man [44]. Red ovals denote the detected failure alarms by the proposed QPN.

supervised evaluation. The overall performance of SiamRPN++resnet50 is ahead of SiamRPN++mobilev2 on VOT2016 and VOT2018, while shows the opposite on VOT2019. However, the detailed accuracy of these two trackers is diverse for each individual sequence, supporting the conclusion that the correction tracker is not necessarily stronger than the base tracker. In addition, the AUC values on OTB2015 and UAV123 of AFAT are also better than SiamRPN++resnet50, demonstrating the advantage of the proposed quality prediction network under unsupervised evaluation. But AFAT cannot outperform SiamRPN++resnet50 on LaSOT with AUC metric, as LaSOT dataset shares a longer average sequence length than others, with more demands on long-term robustness.
Table 3. Tracking results on VOT2018. (The best three results are highlighted by red, blue and brown.)

| Method          | EAO  | A    | R    | EAO  | A    | R    |
|-----------------|------|------|------|------|------|------|
| ECO             | 0.280| 0.483| 0.276| 0.286| 0.509| 0.281|
| CFCF            | 0.303| 0.484| 0.267| 0.323| 0.493| 0.218|
| CFWCR           | 0.323| 0.536| 0.218| 0.378| 0.586| 0.184|
| LSART           | 0.378| 0.586| 0.184| 0.383| 0.503| 0.276|
| UPDT            | 0.050| 0.600| 0.340| 0.383| 0.503| 0.276|
| SiamRPN         | 0.383| 0.600| 0.340| 0.383| 0.503| 0.276|
| LADCF           | 0.389| 0.605| 0.340| 0.389| 0.503| 0.276|

Gradnet DaSiamRPN GFS-DCF ATOM SiamRPN++ SiamRPN++ AFAT

| MobileV2        | 0.247| 0.507| 0.375| 0.326| 0.560| 0.340|
|-----------------|------|------|------|------|------|------|
| Resnet50        | 0.397| 0.511| 0.143| 0.401| 0.590| 0.204|
| EAO             | 0.401| 0.590| 0.143| 0.410| 0.600| 0.234|
| A               | 0.410| 0.600| 0.234| 0.414| 0.605| 0.239|
| R               | 0.414| 0.605| 0.239| 0.410| 0.600| 0.234|

Fig. 5. The experimental results on UAV123. The precision plots with DP reported in the figure legend (left) and the success plots with AUC reported in the figure legend (right) are presented.

The failure detection numbers are also reported for each dataset in Table 2. It is interesting that our AFAT gives more failure detection results on the VOT datasets, with the detection rate ranging from 30.9% to 34.2%. While our QPN detects more successful results on OTB2015, UAV123 and LaSOT, which is consistent with the insight that VOT datasets generates more difficulties for short-term visual tracking. In addition, we illustrate the detected failure alarms during online tracking in Fig. 4. Though there are false alarms during the tracking procedure, many positive alarms are detected by our QPN with a conservative detection strategy in the training stage. With further correction, a number of
Table 4. Tracking results on VOT2019. (The best three results are highlighted by red, blue and brown.)

|          | SiamRPN | SiamMask | GFS-DCF | ATOM  | SiamRPN++ | SiamRPN++ | AFAT |
|----------|---------|----------|---------|-------|-----------|-----------|------|
| EAO      | 0.285   | 0.287    | 0.291   | **0.292** | **0.292** | 0.287     | 0.295 |
| A        | 0.599   | 0.594    | 0.513   | **0.603** | 0.580     | 0.595     | **0.599** |
| R        | 0.482   | 0.461    | 0.453   | **0.411** | **0.446** | 0.467     | 0.450 |

Fig. 6. The experimental results on LaSOT. The normalised precision plots with NP reported in the figure legend (left) and the success plots with AUC reported in the figure legend (right) are presented.

potential failures can be corrected to a certain degree. The above results and analysis demonstrate the efficiency of AFAT for failure detection.

5.3 Comparison with the state-of-the-art

The VOT2018 Dataset: We first evaluate our AFAT on the recent VOT2018 dataset [22] and compare the tracking performance with 13 state-of-the-art trackers. The VOT2018 public dataset contains 60 challenging video sequences for single object tracking research. Following the standard protocol [25], we report the detailed results in Table 5. AFAT achieves the best EAO score, 0.419, and Accuracy, 0.605, outperforming recent advanced trackers, e.g., LADCF, ATOM and SiamRPN++. In addition, the reported speed of AFAT on VOT2018 is 70.5 FPS (Table 3), demonstrating the effectiveness and efficiency of the proposed adaptive failure-aware system.
The VOT2019 Dataset: We also evaluate our AFAT on the VOT2019 dataset and compare the tracking performance with 6 state-of-the-art trackers. The VOT2019 public dataset contains 60 challenging video sequences for single object tracking research, introducing more challenging factors than VOT2018. Following the standard protocol, we report the detailed results in Table 4. According to the table, the proposed AFAT method achieves the best EAO score, 0.295, outperforming recent advanced trackers, e.g., ATOM (0.292) and SiamRPN++ (0.292). In addition, the reported Accuracy (0.599) and Robustness (0.450) scores of AFAT are also within the top three, demonstrating the effectiveness of the proposed adaptive failure detection and correction framework.

The UAV123 Dataset: UAV123 dataset contains 123 video sequences, recorded by a camera embedded in a UAV system. The average sequence length of UAV123 is 915 frames. We report the experimental results in Fig. 5 with precision and success plots. Our AFAT outperforms recent DCF and Siamese trackers: ECO, CSRDCF, DaSiam and SiamRPN++. Specifically, the proposed AFAT method achieves improved performance in terms of AUC (0.612), DP (0.811) and speed (80.4 FPS), against SiamRPN++ with AUC (0.611), DP (0.804) and speed (37.9 FPS).

The LaSOT Dataset: LaSOT is a recent dataset including 1400 sequences in 70 object categories, with an average sequence length of more than 2500 frames. We evaluate the tracking performance on the designed LaSOT test set that contains 280 video sequences, with 4 videos of each category. Fig. 6 reports the normalised precision plots and success plots of the proposed AFAT and recent state-of-the-art trackers, e.g., LADC, MDNet, VITAL, and SiamRPN++. AFAT achieves 0.574 in the NP metric, outperforming SiamRPN++ (0.571). But SiamRPN++ achieves better performance in terms of AUC than AFAT.

To summarise, our AFAT method achieves advanced performance, as compared with the state-of-the-art trackers. In addition, the speed of our AFAT is ranging from 70 FPS to 87 FPS, depending on the number of detected failures. It should be highlighted that a forward pass of our QPN takes only 1.2~1.5 ms, demonstrating its feasibility in real-time visual object tracking.

6 Conclusion

We proposed an effective online Quality Prediction Network (QPN) by learning sequential response maps in the decision level, enabling adaptive quality perception of an online visual object tracker. With spatial feature extraction and temporal feature fusion, the quality of the current tracking result can be trained from sequential response maps in a supervised manner, achieving online failure detection. Furthermore, we combined our QPN with the recent SiamRPN++ method to construct our AFAT tracker. The significance of realising online failure detection and correction was examined in datasets with diverse challenges. The extensive experimental results obtained on a number of benchmarking datasets demonstrate the effectiveness and robustness of the proposed method, as compared with the state-of-the-art trackers.
References

1. Bertinetto, L., Henriques, J.F., Valmadre, J., Torr, P., Vedaldi, A.: Learning feed-forward one-shot learners. In: NIPS. pp. 523–531 (2016)
2. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.: Staple: Complementary learners for real-time tracking. In: CVPR. pp. 1401–1409 (2016)
3. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: ECCV. pp. 850–865 (2016)
4. Bhat, G., Danelljan, M., Good, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: ICCV. pp. 6182–6191 (2019)
5. Bhat, G., Johndyer, J., Danelljan, M., Shahbaz Khan, F., Felsberg, M.: Unveiling the power of deep tracking. In: ECCV. pp. 483–498 (2018)
6. Choi, J., Kwon, J., Lee, K.M.: Deep meta learning for real-time target-aware visual tracking. In: ICCV. pp. 911–920 (2019)
7. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: CVPR. pp. 4660–4669 (2019)
8. Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M.: Eco: Efficient convolution operators for tracking. In: CVPR. pp. 6638–6646 (2017)
9. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: ICCV. pp. 4310–4318 (2015)
10. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. In: CVPR. pp. 1430–1438 (2016)
11. Danelljan, M., Robinson, A., Khan, F.S., Felsberg, M.: Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: ECCV. pp. 472–488 (2016)
12. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: CVPR. pp. 5374–5383 (2019)
13. Gers, F.A., Schmidhuber, J., Cummins, F.: (1999)
14. Gundogdu, E., Alatan, A.A.: Good features to correlate for visual tracking. IEEE TIP 27(5), 2526–2540 (2018)
15. He, Z., Fan, Y., Zhuang, J., Dong, Y., Bai, H.: Correlation filters with weighted convolution responses. In: ICCV. pp. 1992–2000 (2017)
16. Held, D., Thrun, S., Savarese, S.: Learning to track at 100 fps with deep regression networks. In: ECCV. pp. 749–765 (2016)
17. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. In: ECCV. pp. 702–715 (2012)
18. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE TPAMI 37(3), 583–596 (2014)
19. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE TPAMI 34(7), 1409–1422 (2011)
20. Kiani Galooahhi, H., Fagg, A., Lucey, S.: Learning background-aware correlation filters for visual tracking. In: ICCV. pp. 1135–1143 (2017)
21. Kristan, M., Leonardis, A., Matas, J., et al.: The visual object tracking vot2016 challenge results. Springer (Oct 2016), http://www.springer.com/gp/book/9783319488806
22. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pfugfelder, R., Cehovin Zajc, L., Vojir, T., Bhat, G., Lukezic, A., Eldesokey, A., et al.: The sixth visual object tracking vot2018 challenge results. In: ECCV. pp. 1–52 (2018)
23. Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernandez, G., Vojíř, T., Hager, G., Nebehay, G., Pfugfelder, R.: The visual object tracking vot2015 challenge results. In: ICCVW. pp. 1–23 (2015)
24. Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pfugfelder, R., Kamarainen, J.K., Cehovin Zajc, L., Drbohlav, O., Lukezic, A., Berg, A., et al.: The seventh visual object tracking vot2019 challenge results. In: ICCVW. pp. 1–36 (2019)
25. Kristan, M., Matas, J., Leonardis, A., et al.: A novel performance evaluation methodology for single-target trackers. IEEE TPAMI 38(11), 2137–2155 (2016)
26. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: CVPR. pp. 4282–4291 (2019)
27. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: CVPR. pp. 8971–8980 (2018)
28. Li, P., Chen, B., Ouyang, W., Wang, D., Yang, X., Lu, H.: Gradnet: Gradient-guided network for visual object tracking. In: ICCV. pp. 6162–6171 (2019)
29. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: Algorithms and benchmark. IEEE TIP 24(12), 5630–5644 (2015)
30. Lukezic, A., Vojíř, T., Cehovin Zajc, L., Matas, J., Kristan, M.: Discriminative correlation filter with channel and spatial reliability. In: CVPR. pp. 6309–6318 (2017)
31. Ma, C., Yang, X., Zhang, C., Yang, M.H.: Long-term correlation tracking. In: CVPR. pp. 5388–5396 (2015)
32. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: ECCV. pp. 445–461 (2016)
33. Mueller, M., Smith, N., Ghanem, B.: Context-aware correlation filter tracking. In: CVPR. pp. 1396–1404 (2017)
34. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: ECCV. pp. 300–317 (2018)
35. Park, E., Berg, A.C.: Meta-tracker: Fast and robust online adaptation for visual object trackers. In: ECCV. pp. 569–585 (2018)
36. Song, Y., Ma, C., Gong, L., Zhang, J., Lau, R.W., Yang, M.H.: Crest: Convolutional residual learning for visual tracking. In: ICCV. pp. 2555–2564 (2017)
37. Sun, C., Wang, D., Lu, H., Yang, M.H.: Learning spatial-aware regressions for visual tracking. In: CVPR. pp. 8962–8970 (2018)
38. Tao, R., Gavves, E., Smeulders, A.W.: Siamese instance search for tracking. In: CVPR. pp. 1420–1429 (2016)
39. Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., Torr, P.H.: End-to-end representation learning for correlation filter based tracking. In: CVPR. pp. 2805–2813 (2017)
40. Voigtlaender, P., Luiten, J., Torr, P.H., Leibe, B.: Siam r-cnn: Visual tracking by re-detection. arXiv preprint arXiv:1911.12836 (2019)
41. Wang, M., Liu, Y., Huang, Z.: Large margin object tracking with circulant feature maps. In: CVPR. pp. 4021–4029 (2017)
42. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. In: CVPR. pp. 1328–1338 (2019)
43. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: CVPR. pp. 2411–2418 (2013)
44. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE TPAMI 37(9), 1834–1848 (2015)
45. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: An accelerated correlation filter tracker. Pattern Recognition p. 107172 (2019)
46. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: Joint group feature selection and discriminative filter learning for robust visual object tracking. In: ICCV. pp. 7950–7960 (2019)
47. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. IEEE TIP 28(11), 5596–5609 (2019)
48. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: Learning low-rank and sparse discriminative correlation filters for coarse-to-fine visual object tracking. IEEE TCSVT (2019)
49. Xu, T., Wu, X.J., Kittler, J.: Non-negative subspace representation learning scheme for correlation filter based tracking. In: ICPR. pp. 1888–1893 (2018)
50. Yan, B., Zhao, H., Wang, D., Lu, H., Yang, X.: ’skimming-perusal’ tracking: A framework for real-time and robust long-term tracking. In: ICCV. pp. 2385–2393 (2019)
51. Zhang, L., Gonzalez-Garcia, A., Weijer, J.v.d., Danelljan, M., Khan, F.S.: Learning the model update for siamese trackers. In: ICCV. pp. 4010–4019 (2019)
52. Zhang, Z., Peng, H.: Deeper and wider siamese networks for real-time visual tracking. In: CVPR. pp. 4591–4600 (2019)
53. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware siamese networks for visual object tracking. In: ECCV. pp. 101–117 (2018)