

Received August 6, 2020, accepted August 20, 2020, date of publication August 24, 2020, date of current version September 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019206

Stably Adaptive Anti-Occlusion Siamese Region Proposal Network for Real-Time Object Tracking

FEI WU^{1,2,3}, JIANLIN ZHANG¹, AND ZHIYONG XU¹

¹Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610000, China

²School of Electrical, Electronics and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

³Key Laboratory of Optical Engineering, Chinese Academy of Sciences, Chengdu 610200, China

Corresponding author: Jianlin Zhang (jlin@ioe.ac.cn)

This work was supported by the Signal Processing Laboratory, Institute of Optics and Electronics, Chinese Academy of Sciences.


ABSTRACT Siamese region proposal network has made remarkable achievements in visual object tracking because of its balanced accuracy and speed. However, it regards tracking as a local one-shot detection task, which lose the power of updating the appearance model online thereby cannot handle the object-occlusion, fast motion and out-of-view situations. To tackle this problem, we propose a method that combines adaptive Kalman filter with Siamese region proposal network (Anti-occlusion-SiamRPN) to make full use of the object spatial-temporal information. Specifically we first extract target features through deep network and then uses adaptive Kalman filter to predict target trajectory in these difficult scenarios. Further this trajectory is used to select the candidate area of the next frame for Siamese region proposal network, which improve the searching mechanism. In this way, the introduction of adaptive Kalman filter makes the tracking process online learning which makes up for the disadvantage that Siamese region proposal network can only track offline. In addition, a hard example discrimination method (HEDM) is proposed to estimate whether the occlusion occurs and how seriously it is, which also improve Kalman filtering mechanism to make it update adaptively. Our method being evaluated with the speed of 80 FPS on five widely-applied challenging benchmarks including OTB2013, OTB2015, OTB50, VOT2016 and VOT2018. The extensive experimental results demonstrate our method achieves state-of-the-art effects and great improvement in comparison to other trackers.

INDEX TERMS Visual tracking, Siamese region proposal network, Kalman filter, occlusion judgment, real-time tracker.

I. INTRODUCTION

Visual object tracking is a fundamental problem in various tasks of computer vision, such as video surveillance, automatic driving and human-computer interactions. It aims to estimate the position of a specified object in a changing video, where the object is only identified with a rectangle in the starting frame. Although more and more valuable tracking methods have been proposed, it is still difficult to build a state-of-the-art tracker especially in challenging scenarios with occlusions, rotations, illumination variations and other conditions.

Convolutional Neural Networks (CNNs) have been widely used in various computer vision tasks including image

The associate editor coordinating the review of this manuscript and approving it for publication was Jun Li .

classification, object detection, semantic segmentation and so on. Unlike handcrafted features, CNNs exhaustively extract both the shallow appearance information and deep semantic information of the object to obtain powerful feature representation ability, which is beneficial to identify object categories. However, it is a more difficult challenge to apply it into visual object tracking task because the target objects may be arbitrary and vary dramatically in a video sequence. The methods based on correlation filtering and similarity comparison have been proposed in recent years, which have made great progress in this field.

Correlation filter based tracking algorithms use kernel trick and correlation to model the tracking process and identify the object with template cropping from current image, which have an obvious drawback that they produce strong peaks both for the object and the similar objects (distractors) due to

their limited modeling ability. They are not particularly robust to different object appearance variations and often fail on challenging tracking problems such as out-of-view, occlusion and fast motion.

Siamese networks formulate tracking process as a similarity comparison task through the learned feature cross-correlation operation between the object template and its candidates from the search region. They have obtained favorable performance in many challenging benchmarks. Siamese region proposal network (SiamRPN++ [1]) consists of a template branch and a detection branch which are offline-trained with large-scale datasets, owing to the working principle of region proposal network, it can predict the object location more precisely compared with conventional Siamese networks [9] [10] [12]. However, there are still two problems in SiamRPN++: firstly, the tracking process of SiamRPN++ is completely offline without any online learning, so it cannot update the object appearance model online which is essential to account for drastic appearance changes in tracking scenarios. Secondly, the local search strategy is employed to track object in the sequences, which will lead to the tracking failure, especially in the case of full occlusion, fast motion and out-of-view conditions.

Considering above problems, we propose stably adaptive anti-occlusion Siamese region proposal network, where the past trajectory of the object is utilized, which can represent the object motion information effectively, to as measurement in Kalman filter for updating tracker model especially in these difficult scenarios. Firstly, we exploit the feature-representing ability of Siamese network to obtain the object feature. Then, adaptive Kalman filter is used to update the center location of search area relying on the motion trend of nonoccluded trajectory from Siamese network, which aims to ensure that the object will still be in the search area even if distractors emerge after Siamese region proposal network. Therefore, this method can enhance the reliability of the object tracking and suppress the interference of occlusion and distractor simultaneously, which helps to track the object accurately especially in the out-of-view, occlusion and fast motion situations.

Moreover, Siamese networks cannot be conscious of object-occlusion situation due to its tracking mechanisms. To address this problem, the hard example discrimination method (HEDM) is proposed to determine the current state of object. When the object is judged to be occluded according to HEDM, the tracker model will update, especially the adaptive Kalman filter based on the HEDM parameter is used to update the object search area to tackle the object-occlusion, rather than the background area generated by Siamese region proposal network individually.

To summarize, the main contributions of this work are listed below in threefold:

- A stably adaptive anti-occlusion Siamese region proposal network is proposed for challenging object tracking tasks, which adopts the adaptive Kalman filter into the Siamese region proposal network to update model by obtaining the

trend of object motion information like object trajectory to achieve rigorous real-time tracking.

- A hard example discrimination method (HEDM) is presented to efficiently judge the object state of whether being occluded, fast moving and out-of-view, and improve Kalman filtering mechanism to make it update adaptively, so that the method can adaptively decide model updating.

- Our proposed approach can achieve state-of-the-art performance with the speed of 80 FPS (frames per second) on five benchmarks, including OTB2015 [3], OTB2013 [2], OTB50, VOT2016 [42] and VOT2018 [47].

The rest of this paper is organized as follows. Section II briefly outlines some related works in visual object tracking; In Section III, our new object tracking method is described in detail; In Section IV, the new method is tested and evaluated on five widely-applied challenging benchmarks, and we illustrates several ablative studies about the presented components; Section V concludes the paper.

II. RELATED WORK

A. CORRELATION FILTER BASED TRACKING

Visual object tracking has received increasing attention over the last decades and has remained being a very active research direction. In the past few years, the Correlation Filter based tracker is particularly fast and effective because it can discriminate an arbitrary object and its 2D translations from the background, which was first proposed by Bolme *et al.* [4]. Henriques *et al.* [5] proposed KCF using circulant matrices with the Discrete Fourier Transform and efficiently incorporated multi-channel features in a Fourier domain. Several improved algorithms based on KCF have also achieved good tracking performance. For example, SAMF [6] can simultaneously detect the change of object center and scale by taking the position where the maximum response is. DSST [7] uses separate filters for translation and scaling, and SRDCF [8] reduces the boundary effect by expanding the search area and constraining the effective scope of the filter template. Although these approaches have achieved good results in some specific constrained environments, they all use hand-crafted features that are vulnerable in dynamic situations including illumination changes, occlusion, deformation, etc. Moreover, due to their limited modeling ability, they are not sensitive to distractors which have similar appearance with the object especially in the out-of-view, occlusion and fast motion situations, thus limiting their performance.

B. SIAMESE NETWORKS BASED TRACKING

Siamese trackers apply similarity comparison strategy for tracking, which was first proposed by Ran *et al.* [9], it employed a learned priori deep Siamese similarity function to search for candidates most similar to the exemplar given in the starting frame. Bertinetto *et al.* [10] proposed a fully convolutional Siamese network (SiamFC) to compare feature similarity between the exemplar given in the starting frame and the cropped current frame. RASNet [11]

proposed a Residual Attentional Network to strengthen similarity metric by learning the attention mechanism. GOTURN [12] used a deep regression network to learn a generic relationship between an object's appearance and its motion. CFNet [13] combined the correlation filters and the Siamese tracking network to achieve an end-to-end representation learning. FlowTrack [14] exploited flow information and spatial-temporal attention mechanism to improve the tracking accuracy. SiamRPN [15] introduced a region proposal network after a Siamese network, which regards the tracking process as a one-shot local detection task. C-RPN [49] consists of a sequence of RPNs cascaded from the high-level to the low-level layers in the Siamese network. SiamMask [37] involved a unified framework for visual target tracking and video object segmentation. Recently, SiamRPN++ [1] was proposed to enhance SiamRPN through depth-wise correlation and layer-wise feature aggregation, which was trained with four large datasets on AlexNet [16] and ResNet-50 [17] respectively. Although Siamese networks based trackers obtained favorable achievements because of its balanced accuracy and speed, it is worth noting that they only enhance the feature-representing ability and have not improved the search mechanism yet, which may result in detriment in accuracy for scenes of out-of-view, occlusion and fast motion.

C. KALMAN FILTER IN TRACKING

In 1960, Kalman [18] described a recursive solution to the discrete-data linear filtering problem, which was used to estimate the state of a linear system whose states were assumed to be Gaussian distribution. Kalman filter keeps tracking the object by constantly updating the object's state, which is of great significance for online tracking. Reference [19] developed the tracking algorithm with Kalman filter, [20] introduced a Kalman filter to estimate the object state for further tracking the desired dynamic object and filtering the noise, [21] used an adaptive Kalman filter to construct the motion model in the tracking process. Dynamical setting Kalman filter [22] was proposed to dynamically set the optimal process error covariance matrix for a constant velocity model Kalman filter, which can track a real erratic object. In addition, adaptive Kalman filter was combined with mean shift [23] or Camshift [24] to improve the robustness. However, all these improved tracking methods tend to failure when the object moves drastically fast owing to the fact that these methods greatly depended on the prediction of Kalman filter without reliable external supervision.

In summary, correlation filter based trackers and Siamese network based trackers are basically relied on the feature representation ability of the tracking system and they are liable to failure especially in some complex situations with object appearance incompleteness or uncertainty. While the Kalman filter based trackers are tend to lose the fast moving object owing to its limited prediction based on linear system model. In order to overcome these problems, we propose a stably adaptive anti-occlusion Siamese region proposal network (Anti-occlusion-SiamRPN), which combines the deep

network with the adaptive Kalman filter for robust object tracking. It is described in more detail in the following section.

III. TRACKING BASED ON ANTI-OCCLUSION SIAMESE REGION PROPOSAL NETWORK

To tracking the occluded object and the fast moving object, a new framework is presented, named as stably adaptive anti-occlusion Siamese region proposal network (Anti-occlusion-SiamRPN), which exploits the feature-representing ability of Siamese network for identifying drastically moving object to further establishes the object motion trajectory through adaptive Kalman filter. As a result, when the object is in the challenging scenarios like occlusion, fast motion and out-of-view, the proposed method can accurately predict the object location while the Siamese network based trackers or correlation filter based trackers cannot identify the object correctly. Further, the hard example discrimination method (HEDM) is proposed to fully monitor the tracking process when the object encounter these difficult conditions. With the help of HEDM, our method achieve promising tracking results owing to the combination of the deep network and the prediction of adaptive Kalman filter.

In this section, we introduce the proposed Anti-occlusion-SiamRPN tracker. The overall framework is overviewed in Section 1). Section 2) introduces the algorithm principle of Siamese region proposal network. How to use the adaptive Kalman filter to deal with the object-occlusion, fast motion and out-of view condition is introduced in Section 3). Last in Section 4), the working principle of HEDM is demonstrated.

1) THE OVERVIEW FRAMEWORK OF ANTI-OCCLUSION-SIAMRPN

The tracking framework of Anti-occlusion-SiamRPN is shown in Fig.1. It consists of two branches, one is called normally inference branch which applies the Siamese region proposal network to predict the object position and to crop the search area for the next frame, and this object trajectory is used as measurement in Kalman filter simultaneously. The other is called abnormally inference branch which uses adaptive Kalman filter to predict the object position and the center position of the next frame's search area instead of the background area generated by the network when the object is judged occluded or moved fast by HEDM.

The proposed architecture consists of a Siamese tracking network, an HEDM module and an adaptive Kalman filter motion trajectory estimation module. The input of the tracking network is a pair which consists of a tracking template cropped from the starting frame (denoted as z) and a search area cropped from the current T frame image (denoted as x), whose sizes are 127×127 and 287×287 respectively. In the Siamese tracking network, the template and the current frame's search area are respectively fed to CNN to obtain response map and object trajectory. The HEDM module will evaluate the state confidence of the being tracked object, which can effectively determine whether the target is in the

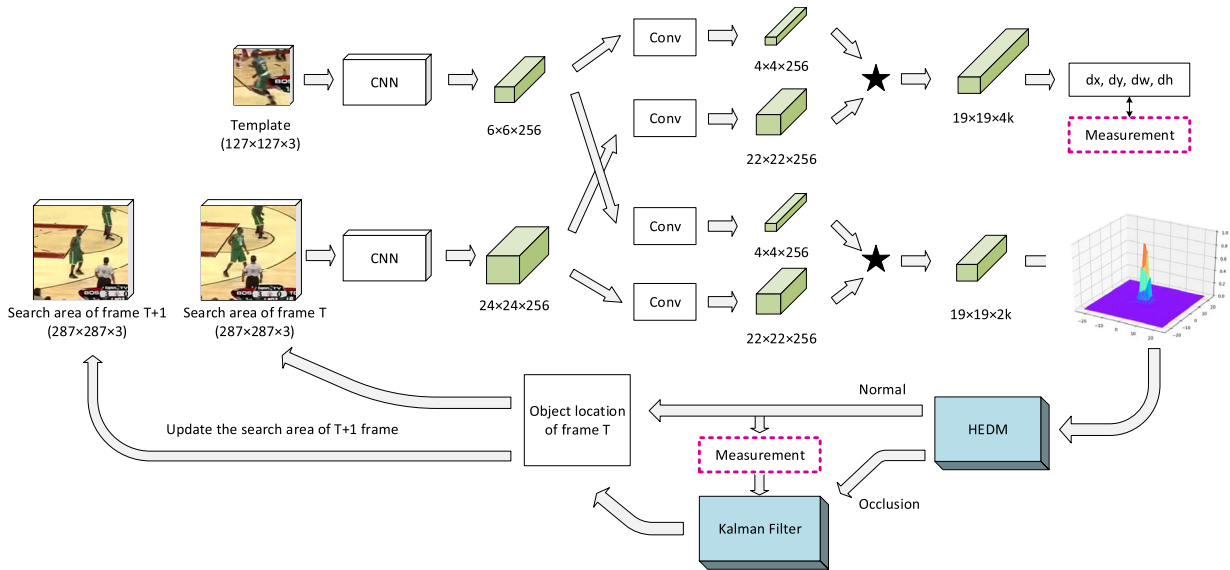


FIGURE 1. The overview of our proposed anti-occlusion-SiamRPN framework. In the figure, k denotes the anchor number, \star denotes depthwise correlation operator.

difficult scenes such as object-occlusion, fast motion, and out of view.

With the low confidence, the proposed method get into normally inference branch to track the object, and further collect the object’s trajectory by the adaptive Kalman filter module, so as to accumulate the long-term trajectory for more accurately positioning. Meanwhile, with the high confidence, which indicates the object may be in occlusion, fast motion, and out of view condition, the proposed method will turn to abnormally inference branch to accurately track the object in these challenging situations. As a result, owing to the complementary mechanism of our framework, through establishing the reliable object trajectory estimation and accurately positioning the search area, the proposed method can smoothly address the challenging issues in visual tracking.

Supposing the object of the T frame was in occlusion or fast motion condition, when we take the cropped search area of the T frame as input into Siamese region proposal network, the HEDM module will declare the T frame is a hard example, then the tracking output will not be generated by the deep network’s background trajectory, on the contrary, the proposed method will utilize the adaptive Kalman filter to correctly predict the object trajectory of the T frame according to the accumulated measurements of the deep network’s well-tracking outputs from 1 to $T-1$ frames, at the same time, this corrected object trajectory will also be used to replace the center position of the search area in the $T+1$ frame to change the wrong selection way under challenging scenes, which improve the searching mechanism. That is, our method aims to learn the motion trend of object trajectory and accurately predict the object motion information in the case of object-occlusion, fast motion and out-of-view.

Fig.2 demonstrates the schematic flowchart of our method, the HEDM module is employed to determine the object state, which consists of the occlusion index (TOI) and the max

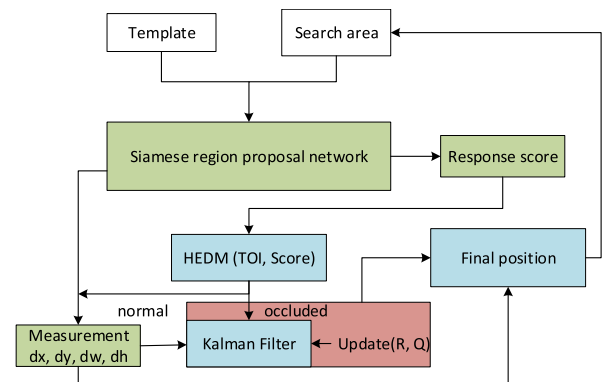


FIGURE 2. The schematic flowchart of the proposed approach.

score (Score) in the response score map. Details of the above are given in Sections 3)-4).

2) ADAPTIVE REPRESENTATION WITH SIAMESE REGION PROPOSAL NETWORK

Inspired by the recently proposed SiamRPN++ [1], we firstly pre-offline trained a Siamese region proposal network to adaptively represent the object feature by general similarity learning. The Siamese region proposal network consists of a feature extraction subnetwork and a region proposal subnetwork. There are two branches in the feature extraction subnetwork. One is called the template branch which cropped from the starting frame as input (denoted as z). The other is called the detection branch which cropped from the current frame as input (denoted as x). These two branch are used to share parameters in the modified AlexNet [16] where the groups from conv2 and conv4 are removed [10], so that both the information from the template branch and the detection branch can be implicitly encoded through the same transformation. After that, the region proposal subnetwork

is used for proposal extraction by the correlation operation between response map of these two branches, it have k anchors, and produces $2k$ channels for classification and $4k$ channels for regression. For convenience, we denote $\varphi(z)$ and $\varphi(x)$ as the output feature maps of the feature extraction subnetwork, so the output of the Siamese region proposal network can be defined as follows:

$$A_{w \times h \times 2k}^{cls} = \text{Corr}([\varphi(x)]_{cls}, [\varphi(z)]_{cls}) \quad (1)$$

$$A_{w \times h \times 4k}^{reg} = \text{Corr}([\varphi(x)]_{reg}, [\varphi(z)]_{reg}) \quad (2)$$

where $\varphi(*)$ is the CNN representation, $\text{Corr}(*)$ is the depth-wise correlation operation that use the template feature maps $[\varphi(z)]_{cls}$ and $[\varphi(z)]_{reg}$ as kernel, it is further improved with kernel function. Then, Siamese region proposal network is optimized by minimizing the average loss function as follows:

$$\text{loss} = L_{cls} + \lambda L_{reg} \quad (3)$$

where we use the Cross-entropy loss as classification loss (L_{cls}) and we employ the smooth_{L1} loss for bounding box regression (L_{reg}), λ is hyper-parameter to balance these two parts. For bounding box regression, the center point, the height and the width of the anchor boxes are denoted as x, y, w, h while those of the ground truth boxes are denoted as x_g, y_g, w_g, h_g . The normalized distance can be expressed as:

$$\delta[0] = (x_g - x) / w_g \quad (4)$$

$$\delta[1] = (y_g - y) / h_g \quad (5)$$

$$\delta[2] = \ln(w_g / w) \quad (6)$$

$$\delta[3] = \ln(h_g / h) \quad (7)$$

The mentioned regression loss can be formulated as:

$$L_{reg} = \sum_{i=0}^3 \text{smooth}_{L1}(\delta[i], \sigma) \quad (8)$$

where the formula of smooth_{L1} loss is:

$$\text{smooth}_{L1}(\omega, \sigma) = \begin{cases} 0.5\omega^2\sigma^2, & |\omega| < \frac{1}{\sigma^2} \\ |\omega| - \frac{1}{2\sigma^2}, & |\omega| \geq \frac{1}{\sigma^2} \end{cases} \quad (9)$$

We use this Siamese region proposal network to generate response map and object trajectory, which as key information in the HDEM and the adaptive Kalman filter respectively. When the object is judged to be occluded according to the HDEM, our presented method which combined the deep network with the adaptive Kalman filter is used for accurately positioning object and selecting the search area in the next frame simultaneously, replacing the background area generated by the deep network individually, and making the tracking process online learning to overcome the problem of not being able to stably track fast moving and occluded targets in Siamese region proposal network.

In order to directly demonstrate the progressiveness of our method, the failure screenshot generated by SiamRPN++ on

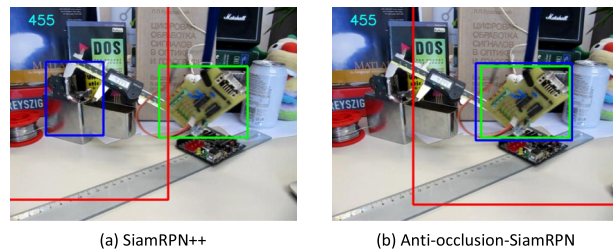


FIGURE 3. The tracking screenshot of SiamRPN++ and anti-occlusion-SiamRPN on board challenging sequence.

Board [3] challenging sequence is shown in Fig.3a, where the blue and red bounding box denote the object and the searching area respectively, and the green bounding box denotes its ground truth. This example indicates once the object's search area selected incorrectly, it is impossible to identify the object. On the contrary, our method, as shown in Fig3b, which accurately positioning the object due to the corrected searching.

3) CANDIDATE AREA PREDICTING WITH ADAPTIVE KALMAN FILTER

Object tracking is a complex task which typically involves the temporal and spatial information. SiamRPN++ [1] can get accurate bounding boxes by applying box refinement procedure, but it only takes advantage of spatial features and cannot deal with fast motion or occlusion. Therefore, we propose Anti-occlusion-SiamRPN to integrate the adaptive Kalman filter and Siamese region proposal network to achieve robust visual tracking. It combined the motion trend learned from temporal information by adaptive Kalman filter and reliable spatial features generated from the deep network, and this adaptive Kalman filter is constructed with HEDM, which adaptively changes the reliability of the measurement values for the Kalman filter. Furthermore, our presented method has two strengths. The first is that Kalman filter is more efficient than the other complicated online learning approach, especially for real-time tracking. The second is that the strong feature-representing ability of deep network can help to discriminate the object from the background effectively, which increases the reliability of predictions.

In the tracking process, the Kalman filter is used to capture the object motion trajectory which aims to make full use of the temporal information of the object to successfully cope with occlusions. In general, the Kalman filter can approximately estimate the object state even its observed measurements with some uncertainties over time, it is a recursive process that consists of two stages: prediction and update. And the Kalman filter can be divided into two models: the process model and the measurement model.

$$x(k) = F_k x(k-1) + B_k u(k-1) + \omega_{k-1} \quad (10)$$

$$z(k) = Hx(k) + v(k) \quad (11)$$

where $x(k)$ is the process state at time k , $z(k)$ is the measurement vector, F_k is the state transition matrix applied to the

previous state vector $x(k-1)$ which is the optimal result of the previous state, H is the measurement matrix, B_k is the control-input matrix applied to the control vector $u(k-1)$ (if there is no control, it can be zero), ω_{k-1} is the process noise vector, $v(k)$ is the measurement noise vector.

The prediction stage including state prediction and covariance prediction.

$$\dot{x}(k) = F_k \hat{x}(k-1) \quad (12)$$

$$\dot{p}(k) = F_k \hat{p}(k-1) F_k^T + Q_{k-1} \quad (13)$$

where $\dot{x}(k)$ is the prior state estimate at time k , $\hat{x}(k-1)$ is the optimal state at time $k-1$, $\hat{p}(k-1)$ is a posteriori estimate error covariance, $\dot{p}(k)$ is a priori estimate error covariance, Q_{k-1} is the covariance matrix of system process noise.

The update stage including calculating the Kalman gain, status update and covariance update, also known as the correction phase.

$$g_k = \dot{p}(k) H^T [H \dot{p}(k) H^T + R_k]^{-1} \quad (14)$$

$$\hat{x}(k) = \dot{x}(k) + g_k [z(k) - H \dot{x}(k)] \quad (15)$$

$$\hat{p}(k) = (I - g_k H) \dot{p}(k) \quad (16)$$

where g_k represents the Kalman gain, R_k is the covariance matrix of measurement noise, I is a unit matrix.

We introduce the hard example discrimination method (HEDM) to fully supervise the credibility of the Kalman filter, specifically, the value of the Q_{k-1} and R_k can be described by HEDM, where the Q_{k-1} and R_k denotes the reliability of the predicted value and the measurement value in the Kalman filter. When the HEDM value is less than the threshold what we set, the Q_{k-1} and R_k are performed as:

$$Q_{k-1} = \text{HEDM} \quad (17)$$

$$R_k = (1 - 1/\text{HEDM})^2 \quad (18)$$

Otherwise, the Q_{k-1} and R_k are performed as:

$$Q_{k-1} = 0 \quad (19)$$

$$R_k = \infty \quad (20)$$

The introduction of HEDM can not only judge the occlusion situation, but also improve Kalman filtering mechanism to make it update adaptively. Being different from Siamese network based trackers, instead of using the center of the object position in the previous frame, our method uses the object position predicted by the adaptive Kalman filter for cropping, which changes the search mechanism.

The long-term corrected trajectory is adopted as the tracked location, rather than the output from the deep network in that a tracker without an online learning process cannot handle occlusion problems. The working principle of the hard example discrimination method (HEDM) is demonstrated in the next section.

4) HEDM FOR OCCLUSION UPDATING

The learning process of Siamese region proposal network is completely offline. The template for the similarity comparison with the search area is always the ground

truth given in the starting frame and never updates in its original framework, which brings in the drawback that once the object encounters large deformation or occlusion, the Siamese region proposal network cannot stably tracking. Through combing the Siamese region proposal network with online learning process, our method effectively compensate for these defects, the tracking failure can be avoided especially in the challenging scenes of object-occlusion, fast motion and out-of-view.

The occlusion index (TOI) utilizes the peak signal-to-noise ratio of the Siamese region proposal network's response map, which determines occlusion state through discriminating the difference between the response map of the current frame with the following equation:

$$\text{TOI} = (R_{\text{peak}} - R_{\text{low}}) / R_{\text{std}} \quad (21)$$

where R_{peak} is the maximum value in the response map, R_{low} is the minimum value in the response map and R_{std} is the standard deviation of the response map. The denominator implies the undulation of the response map, which means the stability of confidence in the current frame. The numerator implies the maximum difference of the response map, which means the dependability of these object candidates.

David3 is one of sequences in OTB2015 [3], which has some challenge scenes like occlusion, background clutters, etc. As shown in Fig.4, when the object is not occluded (Fig.4a, c), the response map shows one sharp peak, which indicates the object candidates with high confidence. By contrast, there is a shallow peak response map in Fig.4b, which means the object suffer from serious occlusion, and the TOI increases significantly due to its average response difference descending.

The max response score and the TOI value of David3 challenging sequence from 1st frame to 150th frame are shown in the third row of Fig.4. They have opposite variation trends when the object is occluded (79th frame to 84th frame), which also corresponding to the scenarios in the first row of Fig.4. In order to indicate the confidence of object-occlusion judgment based on the max response score and the TOI value above, we define the hard example discrimination method (HEDM) as follows:

$$\text{HEDM} = \theta_1 * (1/\text{score}_{\text{max}}) + \theta_2 * \text{TOI} \quad (22)$$

When the HEDM value, which consist of the max response score and the TOI with ratios θ_1 , θ_2 , is higher than the threshold what we set, the object is judged to be occluded, under these conditions, our proposed method will determine the tracking output through combining the adaptive Kalman filter with the deep network, and simultaneously use this accurate position to select the search area of the next frame, thereby avoiding the incorrect searching from the previous object-occlusion's center position. To adequately justify the effectiveness of the HEDM, Section IV illustrates several ablative studies about it.

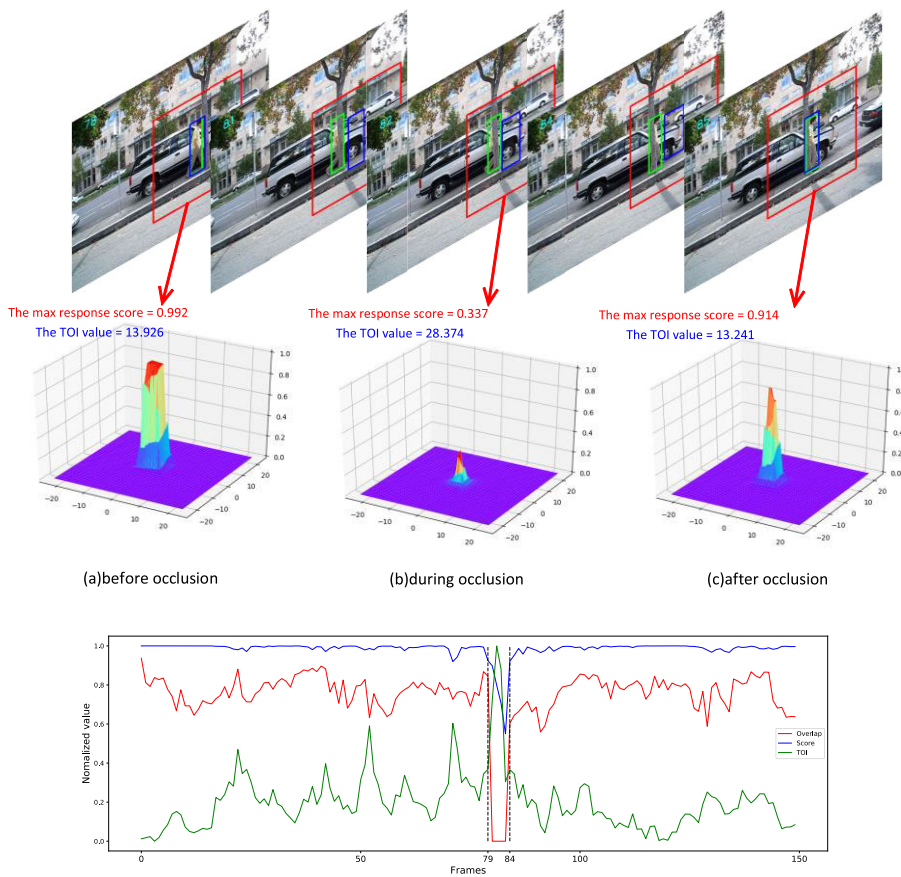


FIGURE 4. Three typical shots of David3 sequence and their response maps. The first row shows the scenarios (a) before occlusion, (b) during occlusion and (c) after occlusion, where the blue and red bounding box denote the object and the searching area of Siamese region proposal network, respectively, and the green bounding box denotes its ground truth. The second row shows the response map corresponding to the scenario above. The third row shows the scores, normalized TOI values and according overlaps from the 1st frame to the 150th frame.

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

For a fair comparison with the baseline SiamRPN++ [1] which using modified AlexNet [10] as backbone, we set the same hyper-parameters as the baseline to train our proposed method (Anti-occlusion-SiamRPN). We use color images offline training on the training sets of COCO [25], ImageNet DET [26], ImageNet VID [26], and YouTube-BoundingBoxes Dataset [27] to learn a similarity comparison standard between general objects for visual tracking. In both training and testing, the sizes of the template patches and the searching regions are set to 127 pixels and 287 pixels respectively, and the optimization method is SGD. We use a warm-up learning rate of 0.001 for the first 5 epochs to train the RPN branches. For the last 45 epochs, the whole network is end-to-end trained with learning rate exponentially decayed from 0.01 to 0.0005. Weight decay of 0.0005 and momentum of 0.9 are used. The anchor number is 5, the ratios of anchor is set to 0.33, 0.5, 1, 2, 3, respectively, the aspect ratio penalty for scale change is 0.16 and the influence of the cosine window is set to 0.4.

In the inference phase, we carried out a simple experimental test without a too-fine parameter search to select the hyper-parameters of Kalman filter. Under considering the distribution of Kalman filter parameters while kept other hyper-parameters no change, in this way we determine the optimal parameters of the threshold value and adjustment coefficients. To set the hyper-parameters for the hard example discrimination method (HEDM), under considering the distribution of HEDM value while kept other hyper-parameters on change, ratios θ_1, θ_2 are set to 6.7 and 0.9 respectively. Then these hyper-parameters were evaluated on several challenging video sequences on OTB benchmarks.

We implemented our tracker in the Pytorch 1.0.0 framework. The experimental results were obtained on a PC equipped with an Intel®Core™i5-2400 @ 3.10GHz CPU and a NVIDIA GeForce GTX 1080 GPU.

B. ABLATION ANALYSIS

To verify the effectiveness of the adaptive Kalman filter and the HEDM components in our method, we conduct the ablation study. We compare the proposed method

TABLE 1. Ablation study of the effectiveness of tracking components on OTB using the area under the curve (AUC).

Tracker	OTB2013	OTB2015	OTB50
SiamRPN++	65.0	66.6	62.1
SiamRPN++-Kalman	17.3	19.4	14.1
SiamRPN++-Kalman -Score	65.7	67.0	62.9
SiamRPN++-Kalman -PSR	66.1	67.2	63.1
Anti-occlusion-SiamRPN	67.6	67.9	64.3

(Anti-occlusion-SiamRPN) with the baseline (SiamRPN++) on OTB2013, OTB2015 and OTB50, respectively. There are four variants to the baseline, the first is to combine Kalman filter with the baseline directly, the second is to add the adaptive Kalman filter and the Peak to Sidelobe Ratio (PSR) which was first proposed in MOSSE [4] on the baseline, the third is to add the adaptive Kalman filter and the max response score on the baseline, and the fourth represents the final algorithm of our proposed method which equipped with the adaptive Kalman filter and HEDM component on the Siamese region proposal network. We set the same parameters both for the Kalman filter and the occlusion-judgment methods. We use the widely applied tracker evaluation method one pass evaluation (OPE) to test the results, and the area under curve (AUC) of these variants on the OTB benchmarks are shown in Table 1. As we can see, using Kalman filter to monitor the tracking process directly is impractical and it is necessary to add the occlusion-judgment method on it. The PSR and the max response score can reflect the object-occlusion state partly. By comparison, only our proposed method achieve greatly better performance than the baseline by 2.6%, 1.3% and 2.2%, respectively, indicating a more robust tracking system in practice.

C. COMPARISON WITH STATE-OF-THE-ARTS

We compared Anti-occlusion-SiamRPN with the state-of-the-art real-time trackers on five benchmarks, including OTB2015, OTB2013, OTB50, VOT2016 and VOT2018.

1) EXPERIMENTS ON OTB BENCHMARKS

The OTB2013 dataset [3] is one of the most widely used normative datasets in visual tracking which contains 51 video sequences. The OTB2015 [3] extends OTB2013 dataset to 100 video sequences. The OTB50 contains 50 relatively difficult video sequences. All these OTB benchmark have 11 challenging attributes including illumination variation, out-of-plane rotation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out of view, background clutter, and low resolution, which are considered as different challenges in visual tracking. The evaluation criteria are the precision plots and the success plots. The precision plot shows the percentage of frames that the tracking results are within 50 pixels from the ground-truth object. The success plot shows the ratios of successful frames when the threshold varies from 0 to 1, where a successful

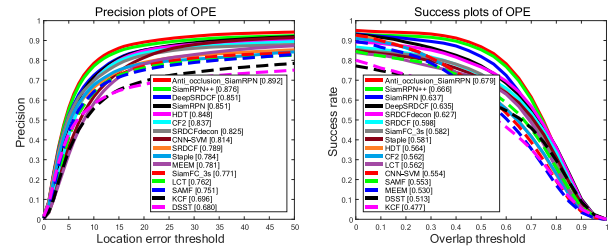


FIGURE 5. The precision plots and success plots on OTB-2015 benchmarks. The curves and numbers were generated with the visual tracker benchmark (OTB) toolkit.

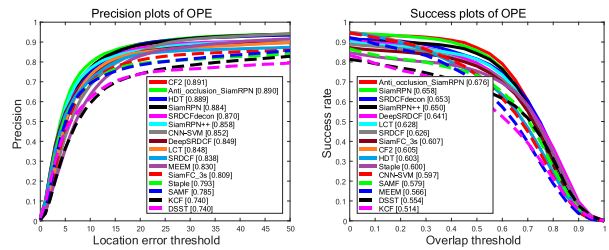


FIGURE 6. The precision plots and success plots on OTB-2013 benchmarks. The curves and numbers were generated with the visual tracker benchmark (OTB) toolkit.

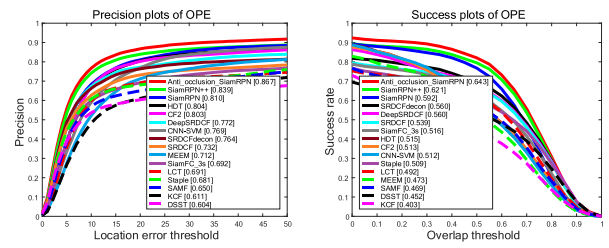


FIGURE 7. The precision plots and success plots on OTB-50 benchmarks. The curves and numbers were generated with the visual tracker benchmark (OTB) toolkit.

frame means its overlap is larger than given threshold. The area under curve (AUC) of success plot is used to rank tracking algorithm, which is also regarded as an important evaluation standard in the visual object tracking field.

Anti-occlusion-SiamRPN tracker is compared with sixteen recent state-of-the-art trackers including SiamRPN++ [1], SiamRPN [15], SiamFC [10], DeepSRDCF [28], SRDCF [8], SRDCFdecon [29], CF2 [30], CNN-SVM [31], LCT [32], Staple [33], HDT [34], MEEM [35], KCF [5], SAMF [6] and DSST [7] on OTB2015, OTB2013, and OTB50 benchmarks. These trackers are evaluated with one pass evaluation (OPE), and its corresponding precision plots and success plots are shown in Fig.5-7. The comparison shows that our method achieves the best performance at real-time speed (80 FPS) among these real-time trackers on all three OTB benchmarks. Specifically, compared with the baseline (SiamRPN++), Anti-occlusion-SiamRPN tracker improves by 2.6%, 1.3% and 2.2%, respectively.

In Fig.8, we select some tracking examples to intuitively show the performances of our tracker compared

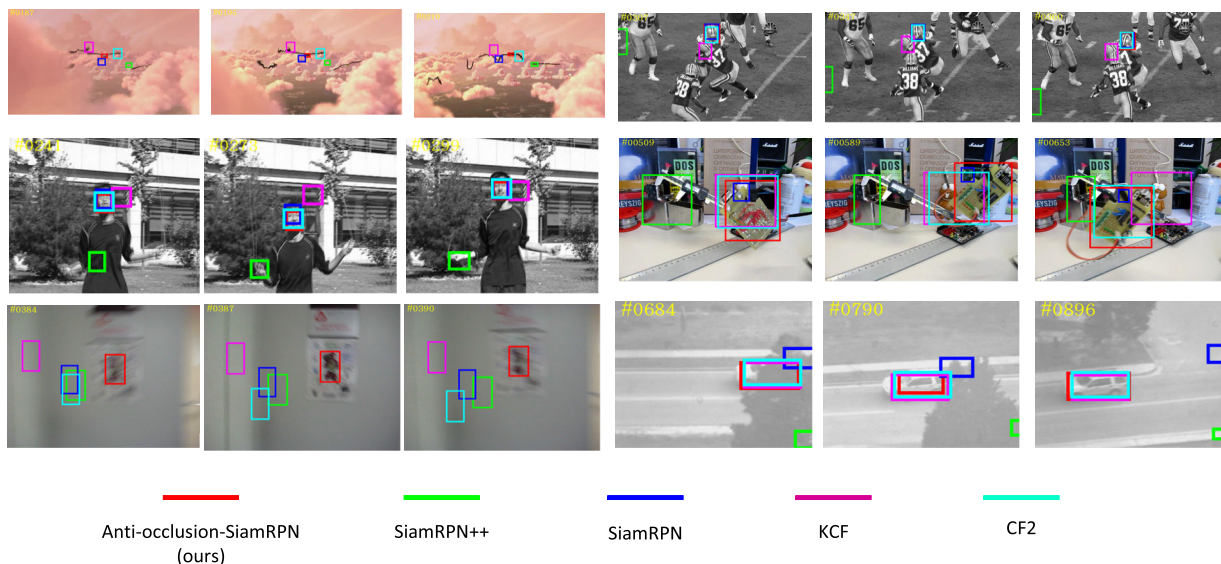


FIGURE 8. The screenshots of the tracking process on six challenging sequences including Bird1, Jumping, Football1, BlurOwl, Board and Suv. The results of these trackers (anti-occlusion-SiamRPN, SiamRPN++, SiamRPN, KCF and CF2) are respectively marked with red, green, blue, purple and blue green bounding boxes.

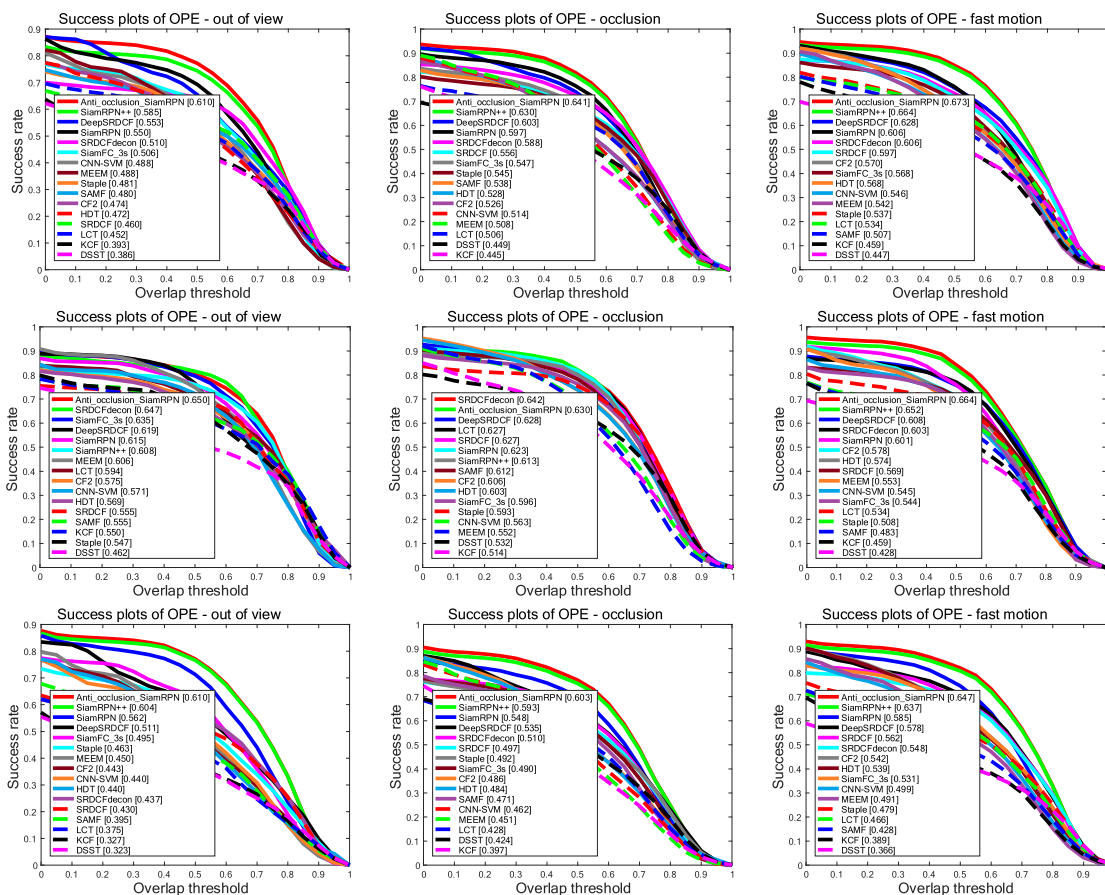


FIGURE 9. The success plots on the OTB-2015, OTB-2013 and OTB-50 dataset in the three scenarios of out-of-view, object occlusion and fast motion.

with four high accuracy and real-time trackers (representing the Siamese network and the correlation filter respectively). In order to further reflect the adaptability of our

method in dealing with three difficult scenes of out-of-view, object-occlusion and fast motion, Fig.9 shows the OTB2015, OTB2013 and OTB50 success plots in these

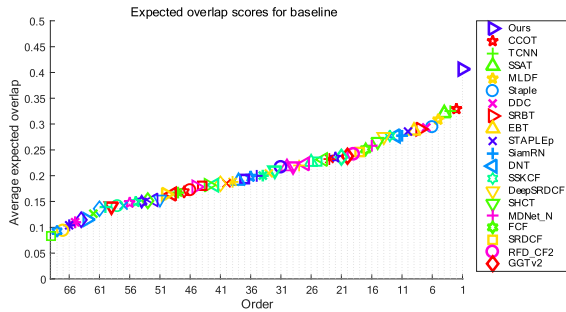


FIGURE 10. The EAO ranking with trackers in VOT2016. The legend shows the results of the top 20 tracker and anti-occlusion-SiamRPN.

TABLE 2. Performance comparison on VOT2016.

Tracker	EAO	Accuracy	Robustness
Anti-occlusion-SiamRPN	0.404	0.620	0.233
SiamRPN++	0.393	0.618	0.238
CCOT	0.331	0.539	0.238
TCNN	0.325	0.554	0.268
SSAT	0.321	0.577	0.291
MLDF	0.311	0.490	0.233
Staple	0.295	0.544	0.378
DDC	0.293	0.541	0.345
EBT	0.291	0.465	0.252
SRBT	0.290	0.496	0.350

three scenarios, respectively, which proves our method, Anti-occlusion-SiamRPN, has always maintained good performance corresponding to these difficult scenario above.

2) EXPERIMENTS ON VOT2016 AND VOT2018

The VOT competition is the tracking challenge held every year in the world. We select VOT2016 and VOT2018 to validate the trackers. VOT2016 contains 60 challenging videos, while VOT2018 includes 10 more challenging sequences. Whenever the tracking bounding box drifts way from the ground truth, the tracker re-initializes after five frames. There are three evaluation metrics in VOT benchmarks, accuracy, robustness and expected average overlap (EAO). The accuracy is computed by the total bounding box overlap ratio. The robustness represents the number of tracking failures. The EAO is the inner product of empirically estimated average overlap and typical sequence length distribution, which has become the most important metric among them.

Anti-occlusion-SiamRPN tracker is compared with all the 68 trackers on VOT2016, as shown in Fig.10, our tracker achieves the leading performance and significantly outperforms other trackers. Specifically, we compared our tracker with the baseline (SiamRPN++ [1]), CCOT [38], TCNN [39], SSAT [40], MLDF [41], Staple [33], DDC [42], EBT [43], and SRBT [42] on VOT2016 to evaluate their performance in detail including the EAO, the accuracy and the robustness, as shown in Table 2. The EAO score of the proposed Anti-occlusion-SiamRPN is 0.404, which is significantly higher than the peer trackers and outperforms the baseline (SiamRPN++) by 1.1.

TABLE 3. Performance comparison on VOT2018.

Tracker	EAO	Accuracy	Robustness
LADCF	0.389	0.503	0.159
MFT	0.385	0.505	0.140
DaSiamRPN	0.383	0.586	0.276
UPDT	0.378	0.536	0.184
RCO	0.376	0.507	0.155
Anti-occlusion-SiamRPN	0.364	0.581	0.286
DRT	0.356	0.519	0.201
SiamRPN++	0.352	0.576	0.290
DeepSTRCF	0.345	0.523	0.215
CPT	0.339	0.506	0.239

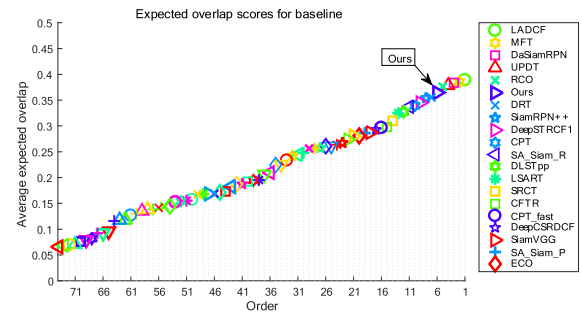


FIGURE 11. The EAO ranking with trackers in VOT2018. The legend shows the results of the top 20 tracker and anti-occlusion-SiamRPN.

Anti-occlusion-SiamRPN tracker is also compared with all the 73 trackers on VOT2018, as shown in Fig.11, our tracker ranked sixth, which also achieves good performance on VOT2018, Table 3 reports the details of the comparison with LADCF [44], MFT [45], DaSiamRPN [36], UPDT [46], RCO [47], DRT [48], SiamRPN++ [1], DeepSTRCF [47], and CPT [47] on VOT2018. The EAO score of the proposed Anti-occlusion-SiamRPN is 0.364, which is significantly outperforms the baseline (SiamRPN++) by 1.2. The above experimental results demonstrate the effectiveness and stability of our proposed method.

V. CONCLUSION

In this paper, we have presented a unified framework, referred to as Anti-occlusion-SiamRPN, to improve Siamese region proposal network with adaptive Kalman filter for stable anti-occlusion real-time visual tracking. In addition, we propose the hard example discrimination method (HEDM) to effectively judge the occlusion situation, and improve Kalman filtering mechanism for adaptively updating. In the framework, our tracker builds the object long-term moving trajectory by adaptive Kalman filter with the powerful feature-representation ability of Siamese region proposal network, which making full use of the object trajectory based on spatial information and temporal information. As a result, the proposed tracker enables the tracking process online learning to accurately predict the object position and improve the search area selection manner so as to robustly deal with complex tracking scenes such as fast motion, object occlusion, out-of-view condition and so

on. The Anti-occlusion-SiamRPN tracker was evaluated on five public datasets including OTB2015, OTB2013, OTB50, VOT2016 and VOT2018, and our method achieved outstanding gain relative to the baseline and reached state-of-the-art status with average speed of 80 FPS.

ACKNOWLEDGMENT

The authors express thank for the experiments provided by the Signal Processing Laboratory.

REFERENCES

- [1] 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.
- [2] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.
- [3] Y. Wu, J. Lim, and M. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015, doi: 10.1109/TPAMI.2014.2388226.
- [4] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.
- [5] 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.
- [6] Z. Liu, Z. Lian, and Y. Li, "A novel adaptive kernel correlation filter tracker with multiple feature integration," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 254–265.
- [7] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf. Nottingham, U.K.: BMVA Press*, Sep. 2014.
- [8] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4310–4318.
- [9] 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.
- [10] 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. Cham, Switzerland: Springer*, 2016, pp. 850–865.
- [11] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, "Learning attentions: Residual attentional siamese network for high performance online visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4854–4863.
- [12] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 FPS with deep regression networks," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2016, pp. 749–765.
- [13] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2805–2813.
- [14] Z. Zhu, W. Wu, W. Zou, and J. Yan, "End-to-end flow correlation tracking with spatial-temporal attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 548–557.
- [15] 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.
- [16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 2012, pp. 1097–1105.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [18] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, Mar. 1960.
- [19] Y. Fan, F. Lu, W. Zhu, G. Bai, and L. Yan, "A hybrid model algorithm for hypersonic glide vehicle maneuver tracking based on the aerodynamic model," *Appl. Sci.*, vol. 7, no. 2, p. 159, Feb. 2017.
- [20] A. Kumar and A. Atiika, "Speed and hardware optimization of ORDP algorithm based Kalman filter for 2D object tracking," *Int. J. Current Eng. Technol.*, vol. 4, no. 3, pp. 1–9, 2014.
- [21] S.-K. Weng, C.-M. Kuo, and S.-K. Tu, "Video object tracking using adaptive Kalman filter," *J. Vis. Commun. Image Represent.*, vol. 17, no. 6, pp. 1190–1208, Dec. 2006.
- [22] G. F. Basso, T. G. S. De Amorim, A. V. Brito, and T. P. Nascimento, "Kalman filter with dynamical setting of optimal process noise covariance," *IEEE Access*, vol. 5, pp. 8385–8393, 2017.
- [23] X. Li, "Object tracking using an adaptive Kalman filter combined with mean shift," *Opt. Eng.*, vol. 49, no. 2, Feb. 2010, Art. no. 020503.
- [24] Y. Zhang, J. Wang, and X. Yang, "Real-time vehicle detection and tracking in video based on faster R-CNN," *J. Phys., Conf. Ser.*, vol. 887, Aug. 2017, Art. no. 012068.
- [25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2014, pp. 740–755.
- [26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
- [27] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "YouTube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5296–5305.
- [28] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 58–66.
- [29] 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.
- [30] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3074–3082.
- [31] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 597–606.
- [32] 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.
- [33] 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.
- [34] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, "Hedged deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4303–4311.
- [35] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2014, pp. 188–203.
- [36] 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.*, 2018, pp. 101–117.
- [37] 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.
- [38] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2016, pp. 472–488.
- [39] H. Nam, M. Baek, and B. Han, "Modeling and propagating CNNs in a tree structure for visual tracking," 2016, *arXiv:1608.07242*. [Online]. Available: <http://arxiv.org/abs/1608.07242>
- [40] 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.
- [41] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3119–3127.
- [42] S. J. Hadfield, R. Bowden, and K. Lebeda, "The visual object tracking VOT2016 challenge results," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2016, pp. 777–823.

- [43] 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.
- [44] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5596–5609, Nov. 2019.
- [45] S. Bai, Z. He, T.-B. Xu, Z. Zhu, Y. Dong, and H. Bai, "Multi-hierarchical independent correlation filters for visual tracking," 2018, *arXiv:1811.10302*. [Online]. Available: <http://arxiv.org/abs/1811.10302>
- [46] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg, "Unveiling the power of deep tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 483–498.
- [47] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Plugfelder, C. Zajc, T. Vojr, G. Bhat, A. Lukezic, A. Eldesokey, and G. Fernandez, "The sixth visual object tracking VOT2018 challenge results," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2018.
- [48] J. Gao, T. Zhang, X. Yang, and C. Xu, "Deep relative tracking," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1845–1858, Apr. 2017.
- [49] 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.



computer vision based on deep learning.

FEI WU received the B.S. degree in electronic and information engineering from Central South University, Changsha, China, in 2017. He is currently pursuing the Ph.D. degree in signal and information processing with the University of Chinese Academy of Sciences, Beijing, China. He is currently pursuing the Ph.D. degree with the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China. His research interests include object tracking, object detection, and computer vision based on deep learning.



JIANLIN ZHANG received the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Chengdu, China, in 2008. He is currently a Full Professor with the Institute of Optics and Electronics, Chinese Academy of Sciences. He has published more than 20 articles and conference papers in his research areas. His research interests include image processing and understanding, computer vision, machine learning, and artificial intelligence.



ZHIYONG XU received the master's degree in computer science from the University of Electronic Science and Technology of China, in 1998. He has supervised more than 30 post-graduate students. He is currently a Professor with the Institute of Optics and Electronics, University of Chinese Academy of Sciences. His research interests include video analysis, image and signal processing, and computer vision.

• • •