Dynamic Siamese Network With Adaptive Kalman Filter for Object Tracking in Complex Scenes

YOUMING WANG\textsuperscript{1,2} AND XIAOYANG MU\textsuperscript{1}

\textsuperscript{1}School of Automation, Xi’an University of Posts and Telecommunications, Xi’an 710121, China
\textsuperscript{2}Xi’an Key Laboratory of Advanced Control and Intelligent Process, Xi’an 710121, China
Corresponding author: Youming Wang (xautroland@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 51875457, in part by the Key Research and Development Program of Shaanxi Province of China under Grant 2019GY-086, and in part by the graduate student innovation fund of Xi’an University of Post and Telecommunications China under Grant CXJJLY2018070.

\section*{ABSTRACT}
Due to the deficiency of prior information for online updating process, the tracking accuracy of fully-convolutional Siamese network (SiamFC) in complex scenes such as similar object interference, fast moving, and appearance change is not good. To solve the problem, a new object tracker based on a dynamic template updating strategy and the re-location mechanism based on the adaptive Kalman is proposed. To suppress the object interference and overcome the instability of fast-moving object tracking, an adaptive Kalman filter method is designed to change the selection method of search region and select the bounding box of the object closest to the predicted position. For the adaptation of appearance change, the high-confidence tracking results are fused with the initial template to dynamic update the template. Compared with traditional Kalman filter, the expectation of residual error for the adaptive Kalman filter method can be controlled in a low range by the adjustment of the gain online. The introduction of the adaptive Kalman based re-location mechanism improves the discriminative ability of SiamFC in interference scene. With the dynamic template updating strategy, the tracker obtains strong generalization capability to adapt to the appearance change of the tracking target. It is demonstrated that the proposed method performs real-time object tracking at the speed of 43fps and achieves competitive performance on OTB, VOT and TC128 datasets compared with other state-of-the-art trackers.

\section*{INDEX TERMS}
Object tracking, Siamese network, Kalman filter, template updating.

\section*{I. INTRODUCTION}
Object tracking is a fundamental task in computer vision, such as autonomous driving, video surveillance, human motion analysis [1]–[5]. The main task for object tracking is to track the target in the subsequent frames and obtain the bounding boxes of the target, which contain the location and size of target in each frame. The main difficulty of object tracking is to build an object tracker, which can adapt to various complex scenes such as appearance changes, fast motion, and similar object interfering.

In traditional tracking algorithms, the features are handcrafted [6]–[10], which are designed for simple scenes and can not perform well in complex scenes, such as lighting changes, occlusion, deformation, fast motion. The convolutional neural networks (CNNs) have gained great attention in computer vision due to its ability to learn the high-level semantic representation and distinguish different categories of objects. The tracking algorithms based on CNNs have demonstrated their superior accuracy over traditional tracking algorithms, such as DeepSRDCF [11] and ECO [12]. However, the online training of CNN based trackers are time-consuming and much slower than real-time tracking algorithms since they need to update a large number of the parameters in the networks.

In recent years, the Siamese network structure has drawn a lot of attention in object tracking because it is an offline training algorithm on large-scale image pairs and discards online updating of network parameters. In a Siamese network, CNNs are utilized to extract the features of inputs and a fully connected network is used to compare the similarity between the search region and the template region. There are many Siamese network structures utilized in object tracking, such as SiamFC, CFNet, SiamRPN, RASNet, SiamFC\textsubscript{tri}, and SA-Siam [13]–[18]. The fully-convolutional Siamese network (SiamFC) [13] is one of the most presentive Siamese
networks in object tracking. It has simple structure and promising features on the balance between the speed and the accuracy of tracking. However, SiamFC may not achieve sufficient performance in complex scenes such as objects interference, fast moving, and appearance change since it lacks strong generalization capability that can be enhanced from the surrounding background.

The object tracking is a process involving both time and space dimension. Hence, many researchers integrate the historical information of the target into tracking process to improve the generalization capability of tracker. Yang et al. [19] adopted the pre-trained long short-term memory (LSTM) networks as target information memorizers to update the template in Siamese networks. Du et al. [20] utilized an adaptive LSTM network to learn the appearance variations of the target. However, the introduction of the memory networks adds additional training burden to the trackers. Recently, the Kalman filter has been widely studied and used in the state estimation of linear system in the past several years. Compared with the memory network like LSTM, the Kalman filter does not need additional training process. The Kalman filter consists of the state update and the measurement update, which combines the priori prediction with current observation information to decrease state error covariance and produce a posteriori state estimate. The property of optimal estimation enables the Kalman filter to enhance the generalization ability and discriminative ability of the model and solve the problem of appearance change and fast-moving in object tracking.

In recent years, researchers have paid much attention to the Kalman filter-based tracking methods. Zhou et al. [21] combined the Kalman filter with SiamFC to process complex tracking scenes and change the selection method of the search region for stably tracking fast moving targets. Li et al. [22] employed the target’s prior color information and the target’s prior trajectory information to design a robust tracking framework on the basis of SiamFC, histogram score model, and Kalman filter model. However, there are several drawbacks in the Kalman filter. The precondition for the Kalman filter to stably tracking is that the mathematical model of the target must be stable. The incorrect model and noise can lead to the decline in tracking accuracy or even the divergence of the filter. When the Kalman filter reaches the steady state, the Kalman gain tends to be stable. If the state of target changes suddenly, the Kalman filter cannot adjust the gain quickly which will lead to the divergence of the filter.

In this paper, an adaptive Kalman filter is proposed to suppress the object interference and overcome the instability of fast-moving object tracking. Compared with the Kalman filter, the adaptive Kalman filter converges quickly and the expectation of residual error retains in a low range by adjusting the Kalman gain online. A response peak detection strategy is proposed to determine whether the target is interfered by similar objects. When multiple peaks in response map are detected, it is assumed that the target is interfered by similar objects. In this case, the bounding box of the peak closest to the predicted position by the adaptive Kalman filter is selected. By the re-location mechanism, the tracker is able to track the target stably in the interference scene. In the fast-moving scene, the target will go to the edge of search region and suppressed by the cosine window of SiamFC, which leads to the instability of fast-moving object tracking. By centering the search region to the predicted position, the target can be close to the center and the instability of fast-moving object tracking is overcome. Besides, a strategy of template updating with high-confidence results is presented to dynamic update the Siamese network. By this method, the adaptation of the Siamese network to the appearance changes is improved. The proposed method achieves very competitive performance on OTB [23], [24], VOT [25] and TC128 [26] datasets.

The main contributions of this paper are summarized as follows:

1) An adaptive Kalman filter is proposed to suppress the interference and overcome the instability of fast-moving object tracking. By adjusting the Kalman gain online, the filter converges quickly and the expectation of residual error retains in a low range.

2) A re-location mechanism based on the adaptive Kalman filter is proposed to improve the discriminative ability of SiamFC in interference scenes.

3) A dynamic template updating strategy is presented to improve the adaptation of tracker to the appearance changes.

II. RELATED WORK

A. TRACKERS BASED ON DEEP LEARNING

In recent years, many researchers combined deep features with traditional correlation filter methods for the object tracking problem. In DeepSRDCF [11], CNN features are combined with Spatially Regularized Correlation Filter (SRDCF) [10] to investigate the effect of convolutional features in the object tracking. Danelljan et al. [27] designed a method of continuous convolution operators (CCOT) to interpolate discrete features and train spatial continuous convolution filters, which enabled the efficient integration of multi-resolution deep feature maps. Danelljan et al. [12] proposed a factorized efficient convolution operator (ECO) for object tracking, where CNN features were combined with handcrafted features. However, the online training of CNN-based trackers is time-consuming and much slower than real-time tracking algorithms since they need to update a large number of the parameters in the network.

For Siamese networks, visual object tracking is considered to be the process of searching for the position with the highest similarity. Bertinetto et al. [13] proposed fully-convolutional Siamese network to build a template-based tracker for object tracking. Valmadre et al. [14] constructed an object tracker based on SiamFC and correlation filter which achieved good performance at high framerates. Li et al. [15] combined region proposal network (RPN) with SiamFC. Dong et al. [17] presented a novel triplet loss to extract expressive deep
features for object tracking and added it into Siamese network instead of pairwise loss for training. He et al. [18] designed a twofold Siamese network comprised of a semantic branch and an appearance branch.

B. TRACKERS BASED ON POSTERIORI STATE ESTIMATION
To enhance the generalization ability and discriminative ability of the tracker, many researchers integrate the historical information of the target into the tracker. Yang et al. [19] adopted the pre-trained LSTM networks as target information memorizers to update the template in Siamese networks. Du et al. [20] utilized an adaptive LSTM network to learn the appearance variations of the target. However, the utilization of these memory networks brings additional computational burden in training and tracking. In recent years, the Kalman and other simple-structured filters are used to enhance the generalization ability and discriminative ability of the tracker. Dai et al. [28] combined kernel correlation filters (KCF) with a Gaussian Particle Filter (GPF) to reduce the influence of deformation. Zhou et al. [21] combined the Kalman filter with SiamFC to process complex tracking scenes and change the selection method of the search region for stably tracking fast moving targets. Li et al. [22] employed the target’s prior color information and the target’s prior trajectory information to design a robust tracking framework on the basis of SiamFC, histogram score model, and the Kalman filter model.

In this paper, a re-location mechanism based on the adaptive Kalman filter is proposed to suppress the object interference and overcome the instability of fast-moving object tracking. Compared with the Kalman filter method, the proposed adaptive Kalman filter method converges quickly and the expectation of residual error retains in a low range by adjusting the Kalman gain online. When the target is interfered, the adaptive Kalman filter is designed to re-locate the target. By centering the search region to the predicted location, the instability of fast-moving object tracking is overcome. Besides, a dynamic template updating strategy is introduced into SiamFC to improve the adaptation of the tracker to the appearance change.

III. METHODOLOGY
The object tracking algorithm is introduced in this section.

A. SiamFC
SiamFC adopts fully convolutional layers to train an end-to-end Siamese network for learning a similarity function between the template and searching region. The architecture of SiamFC is shown in Figure 1.

In Figure 1, \(T\) is the template and \(X_n\) is the search region in \(n\)-th frame, respectively. Fully-connected layers and pre-offline learning approach are used in SiamFC to train an end-to-end Siamese network. The image template \(T\) and searching region \(X_n\) of current video frame are the inputs of SiamFC. Then, the features of \(T\) and \(X_n\) are extracted and cross-correlated in the feature space. This process can be expressed as:

\[
G(T, X_n) = \text{Corr}(\varphi(T), \varphi(X_n)) + b, \tag{1}
\]

where \(b\) denotes a signal that takes value \(b \in \mathbb{R}\) in every location, \(\varphi\) is the feature extractor in SiamFC, \(\text{Corr}(\cdot)\) is the correlation operation, \(G(T, X_n)\) is the response map, the score of which means the similarity between image template \(T\) and searching region \(S\). In SiamFC, the position with maximum value in response map \(G(T, X_n)\) corresponds to the location of the target. To realize real-time tracking, the feature extractor is usually used based on a convolutional neural network with simple structure like AlexNet [29].

B. ADAPTIVE KALMAN FILTER

The object tracking is a process involving both time and space dimension. The priori positions contain the temporal trajectory information of the target. Hence, the Kalman filter is used to integrate the temporal trajectory information of the target with the tracker. However, it has obvious drawbacks in tracking process. The precondition for Kalman filter to stably tracking is that the mathematical model of the target must be precise. The incorrect model and noise would lead to the decline in tracking accuracy or even the divergence of the filter. When the Kalman filter reach the steady state, the Kalman gain tends to be stable. If the state of target changes suddenly, the Kalman filter cannot adjust the gain quickly, which will lead to the divergence of the filter.

To overcome the shortcomings of traditional Kalman filter method, an adaptive Kalman filter method is proposed to suppress the interference and overcome the instability of fast-moving object tracking. By introducing an adjustment factor, the filter gain matrix can be adjusted online to ensure that the residual signals at different time are close to be orthogonal. Thus, the robustness of model tracking and the ability of tracking target state mutation are obtained.

There are two update processes in traditional Kalman filter, including the time and observation update. The time update is mainly used for the prediction of the system, including state and covariance prediction. The observation update includes the calculating of the Kalman gain, status update, and covariance update, which is known as the correction phase. The observation of the current state is used to correct the predicted values during the prediction of the system. The state
The Kalman filter relies on the initial value of \( x_{k\mid k-1} \) and \( B_k U_k \).

The effectiveness of the Kalman filter lie in the covariance update. The covariance update can be expressed as:

\[
P_{k\mid k-1} = \lambda_k C_k P_{k-1\mid k-1} C_k^T + Q_k,
\]

where \( P_{k\mid k-1} \) is a prediction of the covariance of state \( x_{k\mid k-1} \), and \( Q_k \) is the covariance matrix of system noise. \( \lambda_k \) is the adjustment factor for \( P_{k-1\mid k-1} \). In the Kalman filter, \( \lambda_k = 1 \).

The next step is the observation update. First, the Kalman gain for status updates and covariance updates is calculated. After that, the covariance of residual error \( \rho \) is obtained by the Kalman filter. Hence, it should satisfy the condition \( 0 < \rho \leq 0.1 \).

Equation (11) is obtained by the nature of the Kalman filter. According to equation (5) and (11):

\[
P_{k\mid k-1} H^T [I - (HP_{k\mid k-1} H^T + R_k)^{-1} V_k] = 0,
\]

To simplify (12):

\[
HP_{k\mid k-1} H^T = V_k - R_k,
\]

Combining and simplifying (4) and (13):

\[
\lambda_k H C_{k-1\mid k-1} C_k^T H^T = V_k - HQH^T - R_k.
\]

The \( \lambda_k \) can be obtained by function (14). It is in the form of:

\[
\lambda_k = \frac{tr(V_k - HQH^T - R_k)}{tr(H C_{k-1\mid k-1} C_k^T H^T)},
\]

The \( \lambda_k \) is designed to eliminate the residual error at time \( k \). Though the above \( \lambda_k \) can adjust filter quickly, it will also lead to the serious drift in the filter. Because unexpected noise or drift always exist in tracking process, the much too quick adjustment will lead to the drift in filter. Hence, we redesigned the \( \lambda_k^* \) to obtain a smooth adjustment to the filter.

\[
\lambda_k^* = \begin{cases} 
\lambda_0, & \lambda_0 \geq 1 \\
1, & \lambda_0 < 1 
\end{cases}
\]

\[
\lambda_0 = \frac{tr(M_k)}{tr(N_k)}
\]

\[
N_k = V_{0\mid k} - HQH^T - R_k
\]

\[
M_k = HC_{k-1\mid k-1} C_k^T H
\]

where \( V_{0\mid k} \) is the covariance of residual error in the form of:

\[
V_{0\mid k} = \begin{cases} 
V_k V_k^T, & k = 1 \\
\rho V_{0\mid k-1} + (1 - \rho) V_k V_k^T, & k > 1
\end{cases}
\]

\[
\rho = 0.45 + 0.4 \times \text{sigmoid}(k - 10)
\]

\[
\text{sigmoid} (t) = \frac{1}{1 + e^{-t^2}}
\]

where \( \rho \) is the fading factor for \( V_{0\mid k-1} \). It is designed as a sigmoid function. At the beginning of the tracking, the value of \( \rho \) is small. The value of \( \lambda_k^* \) is mainly affected by the covariance of residual error \( V_k V_k^T \), so the filter will converge quickly. As \( k \) becomes larger, the value of \( \lambda_k^* \) is less affected.
by the covariance of residual error. Hence, the noise has little effect to the filter and the curve of the prediction become smooth. With the adjustment of the designed $\rho$, the residual error of the filter at time $k$ retains in a low range and the expectation of residual error tends to be zero as time goes by. By the above method, the filter can be adjusted quickly and smoothly.

In object tracking process, the tracking result of target is defined as the bounding box, which contains the position and size information of target in the frame. In this tracker, the bounding box is input into the adaptive Kalman filter to predict the position of target. In the interference scene, the predicted position by the Kalman filter is used to re-locate the target. In this way, the tracker can stably track the target even in interference scene.

In SiamFC, a cosine window centered at the previous position of the target is input into the search region to suppress the response in the edge. However, when the target is fast moving, the target is away from the center of search region and the response of the target is suppressed by the cosine window. The target with low response is easy to be interfered by background, which will add the instability of tracking. To overcome the instability of fast-moving object tracking, the search region is centered at the predicted position by the adaptive Kalman filter. In this way, the target can be close to the center of cosine window and the tracking accuracy in fast-moving scene is improved.

C. RE-LOCATION MECHANISM

In SiamFC, the response map can be acquired by correlating the features of template and search region. In ideal situation, the position of target locates in the position with max value of response. However, when the similar interference exists, multiple peaks appear in response map and the target may locate in the secondary peak. As is shown in Figure 2, (a) is the 335th frame in Liquor sequence of OTB dataset and (b) is the response map of the frame in SiamFC. In the 335th frame, the response of the similar object is higher than the response of the target. Hence, a re-location mechanism based on peak detection and the adaptive Kalman filter is proposed. If multiple peaks are detected and the value of secondary peak is larger than the main peak by $\varepsilon_n$ times, it is assumed that the target is interfered by other objects. The bounding box of the peak closest to the prediction of the adaptive Kalman filter is selected. The factor $\varepsilon_n$ is in the form of:

$$\varepsilon_n = 0.7 + 0.2 * \frac{F_{\text{max}}^n}{F_{\text{max}}^1},$$

(23)

where $F_{\text{max}}^1$ and $F_{\text{max}}^n$ denote the maximum elements of the response map in the first and n-th frame. According to the Reference [30], the value of $F_{\text{max}}^n$ reflects the confidence of the tracking result in n-th frame. The factor $\varepsilon_n$ should be small and the re-location mechanism tends to be activated in the low-confidence scene. The experiment shows that the optimal coefficient should be between 0.7 to 0.9. Besides, the tracking result in the first frame have the highest confidence in the tracking process. Hence, the factor “0.7” and “0.2” is presented in the equation (24). The value of $\varepsilon_n$ becomes 0.9 in the first frame and reduces in the low-confidence scene. Figure 2(b) shows the main peak A and secondary peak B. Figure 2(c) shows the tracking results of SiamFC (green) and the proposed tracker (blue). The red bounding box is the ground truth. It can be seen that the proposed tracker with the re-location mechanism achieved better performance compared with SiamFC.

D. DYNAMIC UPDATING STRATEGY

In SiamFC, the template is cropped in the first frame of video sequences, which is the target’s initial appearance not being occluded or deformed. When the target’s appearance changes greatly, the similarity between the template and target in present frame becomes low and the tracker will fail to track the target. The traditional approach to solve this problem is to update the template in each frame. These approaches may bring a severe computational load and the real-time performance will be damaged. Besides, updating the target with low confidence tracked results will bring the noise into the framework, which can severely affect the tracking accuracy.

To improve the tracker’s adaptation to the change in appearance and keep a real-time tracking speed, a dynamic template updating method with high confidence tracking results is introduced into SiamFC. If the tracking result has high confidence in this frame, it will be cropped and input into updating process. Average peak-to-correlation energy (APCE) [30] and the maximum response score $F_{\text{max}}^n$ of the response map in n-th frame are utilized to measure the confidence of tracking result. The $F_{\text{max}}^n$ is shown in equation (25). The APCE is expressed as:

$$\text{APCE} = \frac{|F_{\text{max}}^n - F_{\text{min}}^n|^2}{\text{mean}(\sum_{w,h} (F_{w,h}^n - F_{\text{min}}^n))^2)},$$

(25)

where $F_{\text{max}}^n$, $F_{\text{min}}^n$ and $F_{w,h}^n$ denote the maximum, minimum and the w-th row h-th column elements of the response map in n-th frame. APCE indicates the fluctuation degree of response map and $F_{\text{max}}^n$ indicates the confidence level of the detected target. When the target is occluded or lost in search region, APCE and $F_{\text{max}}^n$ will significantly decrease. When $F_{\text{max}}^n$ are larger than its historical average values ($F_{\text{max}}^n_{\text{average}}$) with certain
ratios $\beta_1$ and $APCE$ is larger than certain threshold $T_A$, it is assumed that the tracking result has high confidence in this frame. The template will be updated and the updating process can be expressed as:

$$G(T_n, X_n) = (1 - \alpha) G(T_0, X_n) + G(T_{n-1}, X_n).$$ (26)

where $T_0$ is the template of SiamFC in first frame, $T_{n-1}$ is the tracked target in ($n$-1)-th frame, $\alpha$ is the updating rate of template. The original template $T_0$ contains abundant appearance information of the target without occlusion or deformation, while the template $T_{n-1}$ contains the recent appearance information of target which is different from original template. To utilize the recent and the original appearance information of target, the corresponding response maps of $T_0$ and $T_{n-1}$ are fused in rate $\alpha$. In tracking process, $\alpha$ is set between 0 and 0.5. In this method, the Siamese network can be dynamic updated to adapt to the appearance change of target.

The architecture of the whole tracker is shown in Figure 3 and the pseudo-code of the proposed method is shown in Table 1.

### IV. EXPERIMENTS

#### A. IMPLEMENTATION DETAILS

The SiamFC is trained end-to-end offline to obtain the response map in a search area. In the process of training SiamFC, we use ILSVRC-2015 [31] video dataset to train the network. The initial values of the parameters fit a Gaussian distribution. Stochastic gradient descent method (SGD) is used to optimize the network training. Training is performed over 50 epochs, each consisting of 50,000 sampled pairs. The gradients for each iteration are estimated using mini-batches of size 8, and the learning rate is annealed geometrically at each epoch from $10^{-2}$ to $10^{-5}$. The loss function is similar to the function in reference [13].

The thresholds of the ratio $\beta_1$ and $T_A$ are 0.6 and 9, respectively. The $\alpha$ is set as 0.3. In experiment, the model is implemented in PyTorch 0.4.1 and performed in PC equipped with Intel (R) Core(TM) I7-9750 2.6GHz CPU and NVIDIA GeForce RTX 2080 GPU. The PyTorch implementation ran at 43 fps. The evaluation of the proposed method in different datasets only contain inference without re-training.

#### B. ABLATION ANALYSIS

To demonstrate the effectiveness of the proposed method, additional ablation study is performed in this section. First, an experiment to verify the effectiveness of the adaptive Kalman filter is held. The initial parameters of the filter algorithm in simulation process are given in Table 2.
The estimation of the adaptive Kalman and Kalman filter is shown in Figure 4. The orange curve presents the real value of input. Random noise is added to the input. Compared with the Kalman filter, the proposed adaptive Kalman filter will fast converge even when the value of input changes seriously. The output curve of the adaptive Kalman filter keeps smooth when the input drifts.

The performance of the adaptive Kalman and Kalman filter applied in SiamFC is show in Table 3. The area under curve (AUC) scores of OTB100, OTB2013, and OTB50 are illustrated in Table 3. From Table 3, it can be seen that the Kalman-Siam obtains AUC gain of 1.6%, 2.1%, and 2.2% in OTB2013, OTB50, and OTB100 compared with SiamFC. Compared with the Kalman-Siam, the adaptive Kalman-Siam obtains AUC gain of 1.1%, 1.6%, and 1.1% in OTB2013,
OTB50, and OTB100. Our tracker achieves AUC gain of 5.7%, 7.8%, 6.8% compared with SiamFC. This ablation study proves the efficiency of our method. The $X_0$ and $H$ in experiment were set as:

$$X_0 = \begin{bmatrix} x_0 & y_0 & v_x & v_y \end{bmatrix}^T,$$

$$H = \begin{bmatrix} 1 & 0 & dt & 0 \\ 0 & 1 & 0 & dt \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$  

where $[x_0, y_0]$ is the initial coordinate of the target, $[v_x, v_y]$ is the initial speed of target and it is usually set as zero.

### C. RESULTS ON OTB

In this section, the proposed algorithm is evaluated on three benchmarks, OTB-2013 [23], OTB-50 and OTB-2015 [24]. Center position error of the bounding box and overlap rate in the one-pass evaluation (OPE) are employed for evaluation.
The area under the curve (AUC) of each plot is used to rank the tracking algorithm. The proposed tracker is compared with 15 state-of-the-art trackers, including SiamFC3s [13], long-term correlation tracking (LCT) [32], large margin object tracking with circulant feature maps (LMCF) [30], convolutional features for correlation filter based visual tracking (SRDCF) [10], complementary learners for real-time tracking (Staple) [33], end-to-end representation learning for correlation filter based tracking (CFNet) [14], scale adaptive kernel correlation filter tracker with feature integration (SAMF) [34], accurate scale estimation for robust visual tracking (DSST) [8], high-speed tracking with kernelized correlation filters (KCF) [9], efficient convolution operators (ECo-HC) [12], correlation filter network using triple loss (CFNet2_tri) [17], cropping-inside residual fully-convolutional network (CIResNet-22FC) [35], Siamese network using triple loss (SiamFC_tri) [17], background-aware correlation-filter (BACF) [36], and adaptive decontamination of SRDCF (SRDCFdecon) [37]. These trackers are all representative, and have been cited by many visual-tracking researchers.

The results of the precision plots and the success plots in one path evaluation (OPE) are shown in Figure 5-7. For the readability of the Figures, only the top ten trackers are shown. The complete results of these trackers can be seen in appendix. As can be seen from the figures, our tracker performs best in three datasets. The success rate of the proposed method can reach 66.4%, 59.4% and 65.0% on OTB2013, OTB50, and OTB100 at real-time speed (43 fps). Compared to the SiamFC, our tracker obtains significant AUC gains of 0.057, 0.078, and 0.068 on OTB2013, OTB50, and OTB100, respectively.

The OTB sequences also present many challenging problems like deformation, occlusion, fast motion, background clutter, scale variation, etc. In this paper, the comparison results on OTB100 success plot in the six scenarios of occlusion, fast motion, background clutter, illumination variation, deformation, and scale variation are chosen to depict the performance of the proposed tracker in complex scene. The success plots of OPE with other state-of-the-art tracking algorithms on different attributes are depicted in Figure 8.

Occlusion is the major challenge in visual object tracking. In occlusion scenes, the response of the target decreases and the re-location mechanism is activated. With the re-location mechanism, the tracking process keeps stably. Besides, the template will not be updated in the scenes. So that the tracker achieved good performance even in occlusion scenario.
Fast moving is another common complex scene in visual object tracking. In SiamFC, search region is centred in the position of target in last frame. When target is fast moving, it will move to the edge of search region and suppressed in response map by cosine window. In the proposed tracker, search region is centred at the predicted position by the adaptive Kalman filter. The proposed tracker performs well in fast moving scenario.

Background clutter means that the color or the texture of the background near the target is similar to the target, which requires a tracker with high discriminating ability. When it is combined with the illumination variation, the tracking results will easily drift to the similar interference near the target. In the two scenarios, the proposed tracker ranks second and first among the state-of-the-art trackers. Compared with SiamFC, the proposed tracker achieves AUC gain of 0.105 and 0.088, which proves the effectiveness of the proposed re-location mechanism.

Deformation and scale variation present the appearance change of target in tracking process. In the two scenarios, the proposed tracker ranks first among the state-of-the-art trackers. Compared with SiamFC, the proposed tracker achieves AUC gain of 0.111 and 0.076, which proves the effectiveness of the proposed dynamic updating strategy.

D. RESULTS ON VOT
The proposed tracker is compared with other state-of-the-art trackers in VOT2016. The visual object tracking (VOT) challenge is a competition between short-term model-free visual tracking algorithms containing 60 sequences. The tracker is evaluated by the rectangular initialization of the target in the first frame for each sequence in the data set. The latest Visual Object Tracking toolkit is used for this experiment. The trackers can be divided into various classes including correlation filter methods such as CCOT [27], ECO [12], SRDCF [10], Staple [33], deep convolutional network methods such as SiamFC [13], DeepSRDCF [11]. In Table 4, the proposed tracker achieved the second-best performance among the seven trackers, with an EAO score of 0.338. Compared to SiamFC, the proposed method has a gain about 0.054 in EAO. Besides, the speed of these tracker is listed in Table 4. Though ECO tracker performs better than the proposed tracker, it can only run 3.1 fps, which is far from real time.

E. RESULTS ON TC128
We also conducted comparison experiments on the TC128 dataset. TC128 consists of 128 sequences of color images which contain all kinds of complicated tracking environments. We compared the proposed tracker with many
FIGURE 11. Evaluation on OTB2013 benchmark.

FIGURE 12. Evaluation on OTB50 benchmark.

FIGURE 13. Evaluation on OTB100 benchmark.

other classic trackers whose tracking results could be downloaded from the homepage. Figure 9 shows the precision plot and success plot of their tracking results. The proposed tracker achieved best performance (72.5, 57.8) among these trackers.

There are also some other state-of-the-art trackers with published average precision and success values on TC128 dataset, including: SRDCFdecon [37], MEEM [38], BACF [36], SiamFC [13], SRDCF [10], DeepSRDCF [11], HDT [39], CNT [40]. The comparison of these trackers is shown in Table 5. The best scores are highlighted with bold. The proposed tracker ranks fourth and first in all ten trackers. This experiment proves the performance of the proposed tracker is still competitive on TC128 dataset.

F. QUALITATIVE EVALUATION

Figure 10 shows the frames from five challenging sequences (Bolt2, Girl2, Liquor, MotorRolling and Skating2). The legend shows the relationship between the different colors and the corresponding trackers. These sequences contain similar objects interference, fast moving, and change in appearance. In Bolt2, Liquor and Skating2, the targets are interfered by similar objects seriously. In Girl2 and MotorRolling, the targets are fast moving and facing appearance change. The results show the proposed tracker can track stably in these complex scenes. It proves the effectiveness of our tracker.

V. CONCLUSION

In this paper, we proposed a dynamic Siamese network with adaptive Kalman filter. In the proposed algorithm, the adaptive Kalman filter is presented to predict the position of the target. Compared with traditional Kalman filter, the adaptive Kalman filter converges fast and the expectation of
residual error retains in a low range by adjusting the Kalman gain online. When the target is interfered by similar objects, the prediction by the adaptive Kalman filter is used to relocate the position of target. Hence, the proposed tracker is able to stably track the target in interference scene. Then, the center of search region is the predicted position by the adaptive Kalman filter. Finally, a dynamic updating strategy is introduced into SiamFC which improves the adaptation ability of tracker to appearance change. The proposed method enables the tracking network to robustly deal with complex tracking scenes such as similar interfering, fast moving, and appearance change. Compared with other representative trackers, the proposed tracker can perform real-time tracking and obtain satisfactory results on OTB, VOT, and TC128 datasets.

However, the tracked bounding boxes still contain too much background information. The high-level CNN feature lacks the detail of the images. Hence, we think the incorporation of low-level CNN feature or region proposal network will be beneficial to the proposed tracker. In the future, we will perform more statistical significance tests to assess our tracker with other state-of-the-art trackers. Besides, we will also attempt to reduce the computation burden.

APPENDIX
In this appendix we provide the figures and tables that shows comparison of the proposed trackers with the other 15 trackers in OTB dataset.

REFERENCES
[1] A. Petrovskaya and S. Thrun, “Model based vehicle detection and tracking for autonomous urban driving,” Auto. Robots, vol. 26, nos. 2–3, pp. 123–139, Apr. 2009.
[2] B. Benfold and I. Reid, “Stable multi-target tracking in real-time surveillance video,” in Proc. CVPR, Jun. 2011, pp. 3457–3464.
[3] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras, “People tracking with human motion predictions from social forces,” in Proc. IEEE Int. Conf. Robot. Autom., May 2010, pp. 464–469.

[4] S. S. Rautaray and A. Agrawal, “Vision based hand gesture recognition for human computer interaction: A survey,” Artif. Intell. Rev., vol. 43, no. 1, pp. 1–54, Jan. 2015.

[5] R. G. Lupu, R. G. Bozumitu, F. Ungureanu, and V. Cehan, “Eye tracking based communication system for patient with major neuro-locomotor disabilities,” in Proc. 15th Int. Conf. Syst. Theory, Control Comput., Sinaia, Romania, Oct. 2015, pp. 1–5.

[6] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2005, pp. 886–893.

[7] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, “Visual object tracking using adaptive correlation filters,” in Proc. IEEE Int. Conf. Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2544–2550.

[8] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, “Accurate scale estimation for robust visual tracking,” in Proc. Brit. Mach. Vis. Conf., 2014, pp. 65.

[9] 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.

[10] M. Danelljan, G. Häger, 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.

[11] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, “Beyond correlation filters: Learning continuous convolution operators for visual tracking,” in Proc. ECV, Amsterdam, The Netherlands, Oct. 2016, pp. 472–488.

[12] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, “ECO: Efficient convolution operators for tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5800–5808.

[13] 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., Nov. 2016, pp. 850–865.

[14] 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. 5800–5808.

[15] B. Li, J. Yan, W. Wu, Z. Zha, 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, doi: 10.1109/CVPR.2018.00935.

[16] 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. (CVPR), Jun. 2018, pp. 4854–4863, doi: 10.1109/CVPR.2018.00510.

[17] X. Dong and J. Shen, “Triplet loss in siamese network for real-time visual tracking,” Sensors, vol. 19, no. 9, p. 2201, May 2019.

[18] C. Li, Q. Xing, and Z. Ma, “HKSiamFC: Visual-tracking framework using prior information provided by staple and Kalman filter,” Sensors, vol. 20, no. 7, p. 2173, 2020.

[19] 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.

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

[21] M. Kristan et al., “The visual object tracking VOT2016 challenge results,” in Proc. ECCV, Amsterdam, The Netherlands, Oct. 2016, pp. 8–10.

[22] P. Liang, E. Blasch, and H. Ling, “Encoding color information for visual tracking: Algorithms and benchmark,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 5630–5644, Dec. 2015.

[23] M. Danelljan, G. Häger, 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. 621–629.

[24] M. Dai, G. Xiao, S. Cheng, D. Wang, and X. He, “Structural correlation filters combined with a Gaussian particle filter for hierarchical visual tracking,” Neurocomputing, vol. 398, pp. 235–246, Jul. 2020.

[25] K. ALEX, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 60, no. 6, 2012, pp. 84–90.

[26] M. Wang, Y. Liu, and Z. Huang, “Large margin object tracking with circulant feature maps,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4880–4888.

[27] 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.

[28] 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.

[29] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksis, 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.

[30] Y. Li, and J. A. Zhu, “Scale adaptive kernel correlation filter tracker with feature integration,” in Proc. ECCV, Zurich, Switzerland, Sep. 2014, pp. 254–265.

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

[32] H. K. Galoogahi, A. Fagg, and S. Lucey, “Learning background-aware correlation filters for visual tracking,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1144–1152, doi: 10.1109/ICCV.2017.129.

[33] M. Danelljan, G. Häger, 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, doi: 10.1109/CVPR.2016.159.

[34] J. Zhang, S. Ma, and S. Sclaroff, “MEEM: Robust tracking via multiple experts using entropy minimization,” in Proc. ECCV, Zurich, Switzerland, Sep. 2014, pp. 188–203.

[35] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, “Learning dynamic siamese network for visual object tracking,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1781–1789, doi: 10.1109/ICCV.2017.196.

[36] K. Zhang, Q. Liu, Y. Wu, and M.-H. Yang, “Robust visual tracking via convolutional networks without training,” IEEE Trans. Image Process., vol. 25, no. 4, pp. 1779–1792, Apr. 2016, doi: 10.1109/TIP.2016.2531283.

YOUMING WANG received the M.S. and Ph.D. degrees in mechanical engineering from the Xi’an University of Technology and Xi’an Jiaotong University, Xi’an, China, in 2006 and 2010, respectively. He is currently a Professor with the Xi’an University of Posts and Telecommunications. His current interests include fault diagnosis, vibration control, and structural health management. He is a Fellow of the Chinese Communications Society and a Director of the fault diagnosis branch of the Chinese Society of Vibration Engineering.

XIAOYANG MU received the B.Sc. degree from Xi’an Jiaotong University, Xi’an, China. He is currently pursuing the M.S. degree with the Xi’an University of Posts and Telecommunications, Xi’an. His research interests include computer vision, object tracking, pattern recognition, and machine learning.