Multimedia Tools and Applications (2021) 80:29849–29873
https://doi.org/10.1007/s11042-021-11154-x

Learning adaptive updating siamese network for visual tracking

Yifei Zhou[1] · Jing Li[1] · Bo Du[1] · Jun Chang[1] · Zhiquan Ding[2] · Tianqi Qin[2]

Received: 1 October 2020 / Revised: 7 January 2021 / Accepted: 3 June 2021 /
Published online: 14 July 2021

Abstract
Recently, Siamese network (Siam)-based visual tracking describes the tracking problems as the cross-correlation between convolutional features of the target template and searching regions and solves them by similarity learning, which has achieved great success in performance. However, most of the existing Siam-based tracking methods neglect to explore the feature correlations, which is very important to learn more representative features. Moreover, the first frame is used as the fixed template without updating the template, which leads to a reduction in accuracy. To address these issues, in this paper, we propose an Adaptive Updating Siamese Network (AU-Siam) for more powerful feature correlations and adaptive template updating. Specifically, a siamese feature extraction subnetwork is proposed to introduce the attention mechanism for more discriminative representations. Furthermore, an object template updating subnetwork is developed to dynamically learn object appearance changes for robust tracking. It's interesting to show that the proposed AU-Siam can effectively reduce the probability of tracking drift in the case of fast motions and heavy occlusion and improve the tracking accuracy. Experimental results on public tracking benchmarks with challenging sequences demonstrate that our AU-Siam performs favorably against other state-of-the-art methods.

Keywords Visual tracking · Attention mechanism · Template updating · Region proposal network

1 Introduction

As one of the most important research hotspots in the computer vision[52], object tracking[50] has been widely applied in fields like visual surveillance[33, 38–40] and so on.

This work was supported in part by the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170.

✉ Jing Li
leejingcn@whu.edu.cn

[1] School of Computer Science, Wuhan University, Wuhan, 430072, China

[2] Sichuan Institute of Aerospace Electronic Equipment, Chengdu, 610100, China

Although much progress[49, 55] has been made in recent years, it is still challenging to design a robust tracking algorithm due to factors such as geometric deformations, partial occlusions, fast motion and background clutters, etc.

A typical tracking algorithm consists of three components. The first is an object appearance model, which evaluates the likelihood that a candidate belongs to the tracked object. The second component is a motion model, which describes the dynamics of the object over time. The third component is the search strategy that decides where to look for the most likely position of the next object state. The tracking algorithm in the literature is optimized from one, two or all aspects of the three parts to improve the performance of the tracking algorithm. However, the appearance model contains a large number of clues to distinguish the tracking target from other video content, which is an important aspect of the tracking algorithm and largely determines the overall performance of the tracking algorithm. The tracking algorithm based on an appearance model can be divided into two groups, one based on the classical online updating network, the other based on the matching tracking approach. The former takes advantage of deep feature which can update the classifier online or target appearance model with superior tracking performances. However, the majority of the algorithms prove to be slow. Whereas the tracking method based on matching utilizes a target template to match candidate samples, requiring no online updating, which is fast but not highly accurate[29]. Among the matching-based tracking methods, the siam-based tracking has aroused wide attention from many researchers for it can balance the tracking accuracy and speed.

Siam-based tracking regards the first frame of the tracking video sequences as the template frame of the tracking object. It's usually easy to distinguish the first frame, more often than not, artificially marked or calculated by the target detection algorithm. However, the target in the video is dynamic and cannot be represented in real-time by using only the first frame as a fixed template. When the target varies greatly, the target of the current frame may share little similarity with that of the first frame, leading to tracking mistakes. Therefore, the previous frame or the continual previous frames of the current one may be taken as the target template to acquire dynamic status. Even in this way, the real-time updating template may still result in template shifting and tracking loss. Figure 1 illustrates the comparison of different frames as template methods in visual tracking.

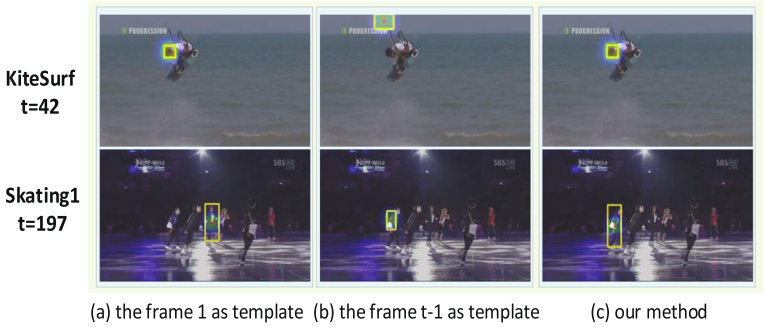

Fig. 1 Comparison of different frames as template methods in visual tracking. We use video Kitesurf and Skating1 as examples. From the figure, we can find that in some cases, the first frame and the previous frame as the template would lose target. With the proposed AU-Siam, our method can get more effective feature representations and achieve more accurate tracking

In this paper, we propose an Adaptive Updating Siamese Network (AU-Siam) for both channel and spatial correlations learning and adaptive template updating learning. Specifically, we propose a siamese feature extraction subnetwork for more powerful feature representations inspired by CBAM[46]. The subnetwork can be utilized to selectively emphasize features and suppress less useful ones, thus performing feature selection and improving the quality of representations. Moreover, an object template updating subnetwork is developed to update the template by fusing the first frame and the previous frame, thus enhancing the robustness to trackers with fast motions and heavy occlusions. Therefore, our method obtains better robustness and accuracy compared with other state-of-the-art tracking methods.

The main contributions of this work are summarized as follows:

(1) A well-designed feature extraction subnetwork is presented to explore how to drive the network to learn more effective appearance features by explicitly modeling the channel and spatial correlations. With more accurate and comprehensive appearance features, our method can enhance the robustness when the dataset is corrupted by strong noise.
(2) With the template updating subnetwork, adaptive template updating mechanism can be introduced to update the tracking models by fusing the information of the first frame and the previous frame, which helps to learn more powerful appearance features when dealing with complicated cases, especially for target with extremely fast motions and heavy occlusions.
(3) A novel algorithm called AU-Siam is developed to seamlessly integrate the feature extraction subnetwork and template updating subnetwork into an Adaptive Updating Siamese Network (AU-Siam) to handle tracking problems in-the-wild. With more robust and effective appearance features, our algorithm outperforms state-of-the-art methods on challenging benchmark datasets such as OTB2013, OTB50, OTB100, VOT2016 and VOT2018.

The rest of this paper is organized as follows: Section 2 reviews previous research in this field. Section 3 introduces the proposed adaptive updating siamese network approach. We present the experiment results and make a comparative analysis between our method and other state-of-the-art tracking methods in Section 4. A summary of this paper is presented in Section 5.

2 Related work

This section is about a brief review of tracking methods that are most related to our work.

Correlation filtering based tracking The correlation filtering trained by the exemplar image to conduct a filtering process on the image, then figuring out the position of the maximum value in the response image, namely, the corresponding target position in the image. D. S. Bolme et.al [51]proposed the correlation filtering based on the least square error for the first time. In 2015, P.Martins[21] proposed a high-speed tracker with kernelized correlation filters (KCF) and multi-channel features. Afterwards, an increasing number of improved algorithms about correlation filtering have occurred [2, 5, 31, 32, 36]. In recent

years, tracking algorithm based on deep learning has achieved remarkable performances in target tracking field. The current prevailing deep learning based tracking includes deep correlation based tracking and some other types[11–13, 22]. The deep correlation filtering based tracking[6–8] uses deep network to extract features, others adopt a siamese network structure together with feature extraction and classifier.

Siamese network based tracking Siamese Network It treats the tracking problem as a similarity learning problem that operates on target branches and searches branches to obtain response images and judge the status of the target according to the maximum value position on the response image. SINT[41] transformed the target tracking problem as a patched matching problem for the first time and realized it through a neural network. SiamFC[1], improved SINT, was the first to solve the tracking problem with a fully convolutional Siamese network structure. Then CFNet[42] integrated correlation filtering(CF) as a network layer and embedded it into a siamese network. StructSiam[54] took into account local structures during tracking, whereas SiamFC-tri[9] introduced Triplet Loss in the siamese tracking network. The dynamic siamese network(DSiam)[19] added the target appearance transformation layer and background suppression hierarchical layer to improve the discrimination ability of the siam-based network. More to be mentioned, Twofold Siamese network(SA-Siam)[20] learned respectively different features and added attention mechanisms and integration of multi-layer features in the Z branch. At the same time, SiamRPN[29] extracted candidate region from correlation feature and encoded target appearance information on the template branch into RPN[37] features, which can effectively discriminate the foreground and background of images and have improved the tracking performances. Based on that, many scholars have proposed improved algorithms [28, 30, 53, 56]. In recent years, siamese networks have been widely used not only in target tracking, but also in other fields [14–18].

Attention mechanism The Attention Mechanism can help the model give different weights to each part of the input X, extracting more critical and important information, and making the model more accurate judgments without adding more overhead to the calculation and storage of the model. The attention mechanism in deep learning is based on this concept of directing your focus, and it pays greater attention to certain factors when processing the data. In recent years, attention mechanism has been widely applied in various fields such as image classification, object segmentation and so on [24, 34, 35, 43, 44]. ACFN[4] proposed an attention mechanism correlation filtering network and used the attention networks to choose the optimal module from various feature extractors to track targets. CSR-DCF[36] took advantage of colorful histograms to restrict correlation filtering learning and establish a foreground spatial reliability map. RASNet[45] introduced three attention mechanisms to improve the discriminating ability of the model: residual attention mechanism, channel attention mechanism and residual attention mechanism. The interdependence between explicit model-establishing feature channels of SENet[23] proved the effectiveness of channel attention in the image recognition tasks. CBAM[46] combines the attention mechanism modules of spatial and channel to achieve better results than the attention mechanism of SENet focusing only on channels. The proposed method, on the one hand, is optimized from the aspect of feature extraction, using channels and spatial attention mechanism based on SiamRPN to improve feature discrimination; on the other hand, from the aspect of template update, adaptively update the tracking model to improve tracking performance and reduce the probability of template drift.

3 Adaptive updating siamese network

In this section, we give a detailed description of our tracking algorithm. We first elaborate on the proposed siamese feature extraction subnetwork in Section 3.1, and then present the object template updating subnetwork in Section 3.2. Section 3.3 illustrates the proposed Adaptive Updating Siamese Network (AU-Siam). Finally, Section 3.4 shows how to train the proposed tracker. Figure 2 depicts the overall architecture of our proposed AU-Siam.

3.1 Siamese feature extraction subnetwork

The SiameseRPN method improves the classical SiamFC method by introducing a region proposal network, which allows the tracker to estimate the bounding box of variable aspect ratio effectively. In this work, we choose SiameseRPN as our backbone due to its robustness and efficiency. In terms of feature extraction, attention mechanism is added to extract more accurate feature information through channel attention and spatial attention, as shown Fig. 3. The semantic features have strong robustness to the appearance change of the target. The channel and spatial attention module can be thought of as the process of selecting semantic properties for different contexts. Therefore, channel and spatial attention mechanism are introduced to extract target features, so that the network can selectively amplify valuable feature and suppress useless feature from the perspective of global information, so as to improve the robustness of the model.

Inspired by [46], the feature F_z firstly extracted from the template frames is used to do attention extraction on channel dimension, and then attention extraction on spatial dimension.

$$F_z' = M_c(\mathrm{F_z}) \otimes \mathrm{F_z}, \tag{1}$$

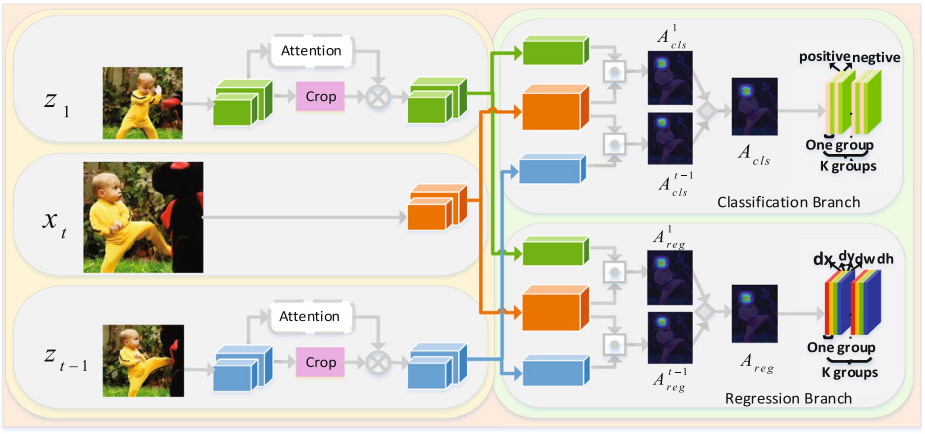

Fig. 2 The overall architecture of our proposed AU-Siam. The siamese feature extraction subnetwork can implicitly enhance feature representations and context information for more effective feature representations. The object template updating subnetwork can generate more accurate and robust response maps by incorporating the first frame with the previous frame of the current frame as the template frames. Then, by integrating the two subnetworks into an Adaptive Updating Siamese Network via a seamless formulation, our AU-Siam can outperform state-of-the-art methods. In classification branch, the output feature map has $2k$ channels corresponding to foreground and background of k anchors. In regression branch, the output feature map has $4k$ channels which corresponding to four coordinates used for proposal refinement of k anchors

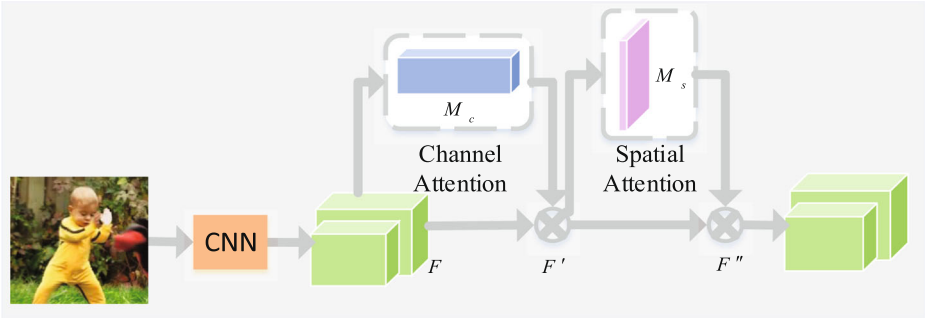

Fig. 3 The network architecture of our siamese feature extraction subnetwork. By introducing the channel and spatial attention, our proposed method can be used to capture hierarchical patterns and attain context information for robust tracking

$$F_z'' = M_s(F_z') \otimes F_z', \tag{2}$$

$\otimes$ denotes element-wise multiplication, F_z' is the output after channel attention module and F_z'' is the final refined output after the two attention modules.

Channel attention We first aggregate spatial information of the feature F_z extracted from the template frame by using both average-pooling and max-pooling, generating two different spatial context descriptors: the average-pooled features F_{avg}^c and max-pooled features F_{max}^c. Both descriptors are then forwarded to a shared network to produce our channel attention map M_c. The shared network is composed of multi-layer perceptron (MLP) with one hidden layer.

$$\begin{aligned} M_c(\mathrm{F_z}) &= \sigma\left(MLP\left(AvgPool\left(\mathrm{F_z}\right)\right) + MLP\left(MaxPool\left(\mathrm{F_z}\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F_{avg}^c\right)\right) + W_1\left(W_0\left(F_{\max}^c\right)\right)\right), \end{aligned} \tag{3}$$

σ denotes the sigmoid function, W_0 and W_1 are the MLP weights, shared for both inputs and the ReLU activation function is followed by W_0.

Spatial attention We aggregate channel information of a feature map by using two pooling operations, generating two maps: F_{avg}^s and F_{max}^s. Each denotes average-pooled features and max-pooled features across the channel.

$$\begin{aligned} M_s(\mathrm{F_z}) &= \sigma\left(f^{7\times7}\left(\left[AvgPool\left(\mathrm{F_z}\right); MaxPool\left(\mathrm{F_z}\right)\right]\right)\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F_{avg}^s; F_{\max}^s\right]\right)\right), \end{aligned} \tag{4}$$

σ denotes the sigmoid function, $f^{7\times7}$ represents a convolution operation with the filter size of 7×7.

Finally, after the attention module, the formula of response map obtained by the convolution of template frame and detection frame can be expressed as follows:

$$f(z, x) = \varphi''(z) * \varphi(x) + b, \tag{5}$$

$\varphi''(z)$ represents the feature of template frame after channel attention and spatial attention modules.

3.2 Object template updating subnetwork

During target tracking, the appearance information of the template frame is the key to the tracking results. The first frame of the tracking video sequence contains a lot of highly reliable target appearance information. However, only the first frame is used as the fixed template frame without updating the target template. When the target appearance changes greatly, it is easy to lose the target. The previous frame is very related to the current state of the target and contains more useful information about the change of the target appearance. However, if only the previous frame is used as a template frame, the tracking shift will occur in all subsequent frames when the previous frame is tracking drift. Therefore, this paper proposes an object template update subnetwork, which extracts the appearance information of the template by combining the first frame and the previous frame as the target template, to adapt to the specific update requirements of the current frame, improve the reliability of the target template, and improve the robustness of tracking.

When the first frame z_1 as the target template, the corresponding response map of it and the detection frame x is as follows:

$$f\left(z_1, x\right) = \varphi''\left(z_1\right) * \varphi\left(x\right) + b, \tag{6}$$

$\varphi''\left(z_1\right)$ refers to the feature acquired via template frame z_1 passing through attention module. $*$ denotes the convolution operation.

When the previous frame z_{t-1} as the target template, the corresponding response map of it and the detection frame x is as follows:

$$f\left(z_{t-1}, x\right) = \varphi''\left(z_{t-1}\right) * \varphi\left(x\right) + b, \tag{7}$$

$\varphi''\left(z_{t-1}\right)$ represents the feature obtained by template frame z_{t-1} going through attention module.

When combine the first frame z_1 and the previous frame z_{t-1} as the target template, the corresponding response map of it and the detection frame x is as follows:

$$f\left(z_1, z_{t-1}, x\right) = \eta\left(\varphi''\left(z_1\right) * \varphi\left(x\right)\right) + (1-\eta)\left(\varphi''\left(z_{t-1}\right) * \varphi\left(x\right)\right) + b, \tag{8}$$

where η is the weighted parameter of the two templates.

In the case that the frame before the current frame loses the target, we only use the first frame as the matching template. We use a condition to determine if the target was lost in the previous frame before each frame is calculated. As shown in Fig. 4, we use S_t to represent the target position of frame t when the first frame as the template, and use S'_t to represent the target position of frame t when the previous frame as the template. C_{error} is the Euclidean distance of S_t and S'_t. If C_{error} is greater than threshold θ, it means that the previous frame of the current frame has been lost target, then we use the first frame as the template. If the error is less than threshold θ, it means that the previous frame of the current frame has been not lost target, then we use the first frame and the previous frame as the template.

$$C_{error} = \left\| S_t - S'_t \right\|, \tag{9}$$

$$\begin{cases} \eta = 1, C_{error} \geq \theta \\ \eta \neq 1, C_{error} < \theta \end{cases}. \tag{10}$$

θ is the threshold, which determines whether the target was lost in the previous frame.

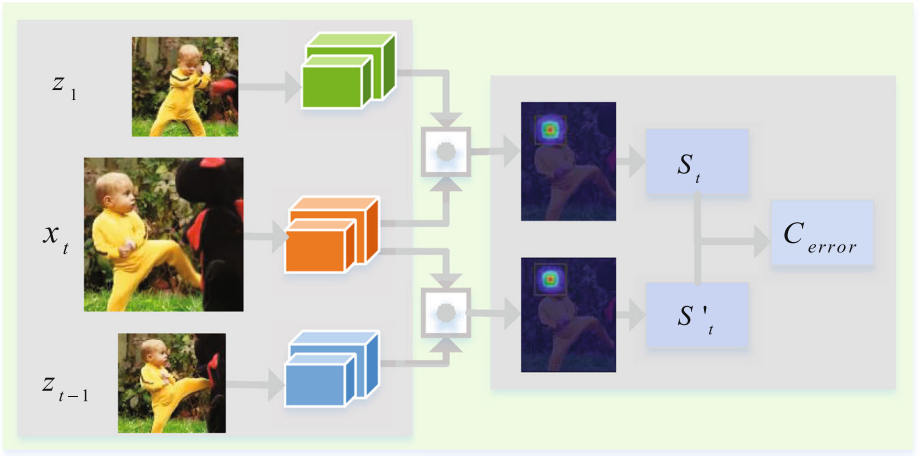

Fig. 4 The illustration of error calculation in the Object template updating subnetwork. The C_{error} is used to determine if the target was lost in the previous frame

3.3 Adaptive updating siamese network

The region proposal network consists of two parts: a classification branch to distinguish between target and background, and a regression branch to fine-tune candidate areas. As shown in Fig. 2, our adaptive updating siamese network integrates the siamese feature extraction subnetwork and the object template updating subnetwork.

The siamese feature extraction subnetwork outputs the channel and spatial attention feature of the template frame(the first frame $\varphi(z_1'')$ and the previous frame $\varphi(z_{t-1}'')$) and the feature $\varphi(x)$ of the search frame. The three features are divided into classification branch feature and regression branch feature. The feature $\varphi\left(z_1''\right)$ of z_1 template frame split into two branches $\left[\varphi(z_1'')\right]_{cls}$ and $\left[\varphi(z_1'')\right]_{reg}$ which have $2k$ and $4k$ times in channel respectively by two convolution layers. Same as z_1 template frame, the feature $\varphi\left(z_{t-1}''\right)$ of z_{t-1} template frame split into two branches $\left[\varphi(z_{t-1}'')\right]_{cls}$ and $\left[\varphi(z_{t-1}'')\right]_{reg}$. The feature $\varphi(x)$ of search frame x is also split into two branches $[\varphi(x)]_{cls}$ and $[\varphi(x)]_{reg}$ by two convolution layers but keeping the channels unchanged.

Then the two branch features generate the response maps of the classification branch and regression branch after the object template updating subnetwork.

$$A_{cls}^{w\times h\times 2k} = \left[f\left(z_1, z_{t-1}, x\right)\right]_{cls}, \tag{11}$$

$$A_{reg}^{w\times h\times 4k} = \left[f\left(z_1, z_{t-1}, x\right)\right]_{reg}. \tag{12}$$

$A_{cls}^{w\times h\times 2k}$ contains a $2k$ channel vector, which represents for negative and positive activation of each anchor at corresponding location on original map. $A_{reg}^{w\times h\times 4k}$ contains a $4k$ channel vector, which represents for dx, dy, dw, dh measuring the distance between anchor and corresponding groundtruth.

Finally, the predicted proposal with the highest classification score is selected as the output tracking region. The process of selecting the output tracking region from the response map is similar to SiamRPN. We collect the top K points in all $A_{cls}^{w\times h\times 2k}$ where l is odd number and denote the point set as $CLS^* = \left\{\left(x_i^{cls}, y_j^{cls}, c_l^{cls}\right) i \in I, j \in J, l \in L\right\}$

where I, J, L are some index set. Variables i and j encode the location of corresponding anchor respectively, and l encode the ratio of corresponding anchor, the corresponding anchor set as $ANC^* = \left\{\left(x_i^{an}, y_j^{an}, w_l^{an}, h_l^{an}\right) i \in I, j \in J, l \in L\right\}$. The activation of ANC^* on $A_{reg}^{w\times h\times 2k}$ to get the corresponding refinement coordinates as $REG^* = \left\{\left(x_i^{reg}, y_j^{reg}, dx_l^{reg}, dy_l^{reg}, dw_l^{reg}, dh_l^{reg}\right) i \in I, j \in J, l \in L\right\}$. Afterwards, the refined top K proposals set $PRO^* = \left\{\left(x_i^{pro}, y_j^{pro}, w_l^{pro}, h_l^{pro}\right)\right\}$ can be obtained by following (13) :

$$\begin{aligned} x_i^{pro} &= x_i^{an} + dx_l^{reg} * w_l^{an} \\ y_j^{pro} &= y_j^{an} + dy_l^{reg} * h_l^{an} \\ w_l^{pro} &= w_l^{an} * e^{dwl} \\ h_l^{pro} &= h_l^{an} * e^{dhl} \end{aligned} \tag{13}$$

3.4 Training

This paper trains the model with an offline training method, cropping the optimal model by minimizing a loss function. The whole network loss function is made up of classification loss and regression loss, employing backpropagation through time(BPTT) and stochastic gradient descent algorithm(SGD) to propagate in gradient and update parameters, which can be illustrated as:

$$loss = L_{cls} + \lambda L_{reg}, \tag{14}$$

where L_{cls} elaborates classification loss, λL_{reg} illustrates regression loss, λ is hyper-parameter to balance the two parts.

The elaborates classification loss L_{cls} is shown as follows:

$$l\,(y, v) = \log\left(1 + \exp(-yv)\right), \tag{15}$$

$$L\,(y, v) = \frac{1}{|D|}\sum_{u\in D} l\,(y\,[u]\,, v\,[u]). \tag{16}$$

Let A_x, A_y , A_w, A_h denote center point and shape of the anchor boxes and let T_x, T_y , T_w, T_h denote those of the ground truth boxes, the normalized distance is:

$$\delta\,[0] = \frac{T_x - A_x}{A_w}, \delta\,[1] = \frac{T_y - A_y}{A_h}, \delta\,[2] = In\frac{T_w}{A_w}, \delta\,[3] = In\frac{T_h}{A_h} \tag{17}$$

And the regression loss employs smooth L_1 loss is:

$$L_{reg} = \sum_{i=0}^{3} smooth_{L1}\left(\delta\,[i]\,, \sigma\right), \tag{18}$$

$$smooth_{L1}\,(x, \sigma) = \begin{cases} 0.5\sigma^2 x^2, |x| < \frac{1}{\sigma^2} \\ |x| - \frac{1}{2\sigma^2}, |x| \geq \frac{1}{\sigma^2} \end{cases} . \tag{19}$$

3.5 Algorithm Implementation

Algorithm 1 shows the proposed tracking method for our work.

Algorithm 1

Input: video sequence frame:[t], the target's initial ground-truth position p_1.
Output:target positions p_2, ...p_n.

for t=1:n do
  if t==1 then
    Given p_1, learn the feature of the target with (1) and (2)
  if t==2 then
    Calculate target positions p_t with (8), $\eta = 1$
  else
    According to p_1, calculate the target positon s_t with (6)
    According to p_{t-1}, calculate the target positon s_t' with (7)
    According to s_t and s_t', calculate the error C_{error} with (9)
    if $C_{error} \geq \theta$ then
      Calculate target positions p_t with (8), $\eta = 1$
    else
      Calculate target positions p_t with (8), $\eta \neq 1$
    end if
  end if
end for

4 Experiment results and analysis

We conduct extensive experiments to analyze and evaluate the proposed tracking method. In the following, we first introduce some experimental settings. Then we evaluate our model on five of the largest benchmarks, OTB2013[47, 48], OTB50[47, 48], OTB100[47, 48], VOT2016[25] and VOT2018[26].

4.1 Implementation Details

The proposed AU-Siam is trained with Pytorch on GeForce GTX 1080 Ti GPU.

Same with SiamRPN, but adds the previous frame as the template frame for template matching. Use AlexNet[27] and CBAM[46] as the base network. The parameters of the network are initialized with the ImageNet pre-trained models.

To increase the generalization capability and discriminative power of our feature representation, and avoid over-fitting to the scarce tracking data, our tracker is pre-trained offline from scratch on the video object detection dataset of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC15). This dataset includes more than 4000 sequences with about 1.3 million labeled frames. It is widely utilized in tracking methods recently as it depicts scenes and objects distinct to those in the traditional tracking benchmarks.

We apply stochastic gradient descent (SGD) with the momentum of 0.9 to train the network from scratch and set the weight decay to 0.0005. The learning rate exponentially decays from 10^{-2} to 10^{-5}. The model is trained for 50 epochs with a mini-batch size of 32. θ is 6, λ=1 is same as the SiamRPN, the weights of W_0 and W_1 in (3) are trained by offline training, and when the weighting parameter η in (8) is 0.9, the result is the best.

4.2 Evaluation metrics

We evaluate our approach on two popular challenging datasets, online tracking benchmark (OTB2013[47, 48], OTB50[47, 48], OTB100[47, 48]) and visual object tracking benchmark (VOT2016[25], VOT2018[26]).

OTB The object tracking benchmarks (OTB) consist of three datasets, namely OTB-2013, OTB-50 and OTB-100. They have 51, 50 and 100 real-world targets for tracking, respectively. All sequences have eleven interference properties.

The two standard evaluation metrics on OTB are success rate and precision. For each frame, we compute the IoU (intersection over union) between the tracked and ground-truth bounding boxes, as well as the distance of their central locations. A success plot can be obtained by evaluating the success rate at different IoU thresholds. Conventionally, the area-under-curve (AUC) of the success plot is reported. The precision plot can be acquired similarly, but usually, the representative precision at the threshold of 20 pixels is reported. We use the standard OTB toolkit to obtain all the numbers.

VOT The VOT2016 dataset contains 60 video sequences showing various objects in challenging scenarios. The VOT2018 dataset consists of 60 challenging video sequences, which is annotated with the same standard as VOT2016. According to the evaluation criterion, a tracker is re-initialized with the ground truth location whenever tracking fails (the overlap between the estimated location and ground truth location equals zero).

The metrics used for the evaluation of the VOT dataset include accuracy(A), robustness(R) and expected average overlap (EAO). Accuracy is defined as the overlap ratio of the estimated location and ground truth while robustness is defined as the number of tracking failures. EAO is a function of sequence length, computed by the average accuracy for a certain number of frames after tracker initialization. A good tracker has high A and EAO scores but low R scores. More details about the evaluation protocol can be found in [25, 26].

4.3 Comparison with the State of the Arts

Five benchmarks including OTB-2013, OTB-50, OTB-100, VOT2016 and VOT2018 are adopted to demonstrate the performance of our tracker against some state-of-the-art. Traditionally, tracking speeds in excess of 25(FPS) is considered real-time. Our tracker runs at 77(FPS). All results in this section are obtained by using the OTB toolkit[47, 48], and VOT toolkit[25, 26].

4.3.1 Experiments on OTB-2013, OTB-50 and OTB-100

We evaluate the proposed algorithms with comparisons to numerous state-of-the-art trackers including SiamRPN[29], DeepSRDCF[7], SRDCF[5], SiamFC[54], CFNet[9], ACFN[4], CSR-DCF[36], Staple[2] and DSiamM[19]. Note that SiamRPN, CFNet, SiamFC, and DSiamM are latest siam-based trackers, and CSR-DCF and ACFN employ attention mechanisms, and DSiamM uses a template update policy, and SimRPN and DeepSRDCF are recent fast deep trackers. All the trackers are initialized with the ground-truth object state in the first frame. Figure 5 shows the overall performance of our method and other state-of-the-art tracking algorithms in terms of the success and precision plots for OPE on OTB-2013, OTB-50, and OTB-100. Figure 6 shows the accuracy and success rate of attributes such as

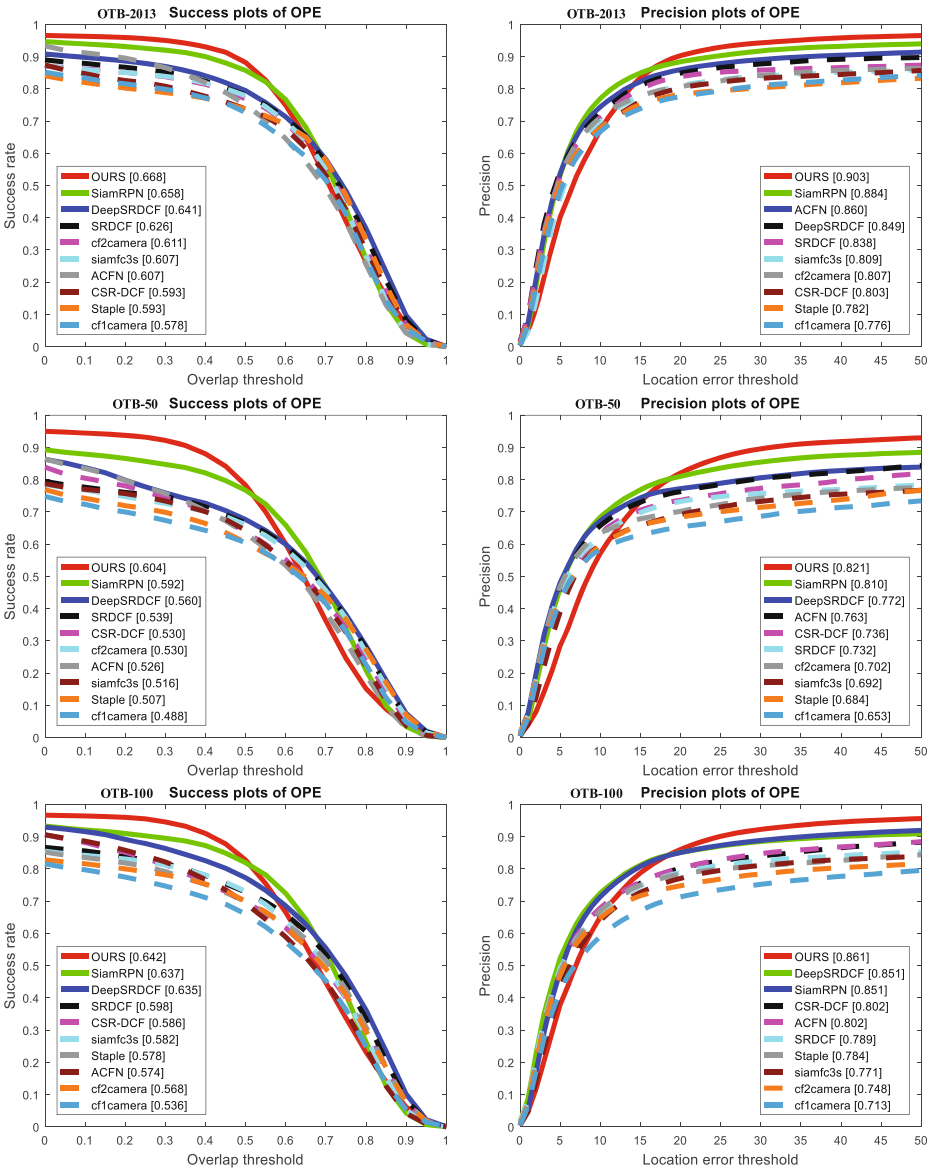

Fig. 5 Precision plots and success plots (AUC) over overall of 9 tracking algorithms on OTB2013,OTB50 and OTB100 datasets

fast motion (FM), occlusion (OCC) and scale variation(SV) on OTB-100. Table 1 summarizes more results, including the running speed in frames per second (FPS). The comparison shows that our algorithm has the best real-time tracking performance on all three OTB benchmarks.

On the results of OTB-2013 , our proposed algorithm performs the best against the other trackers with AUC and precision score of 66.8% and 90.3%. In the trackers using the

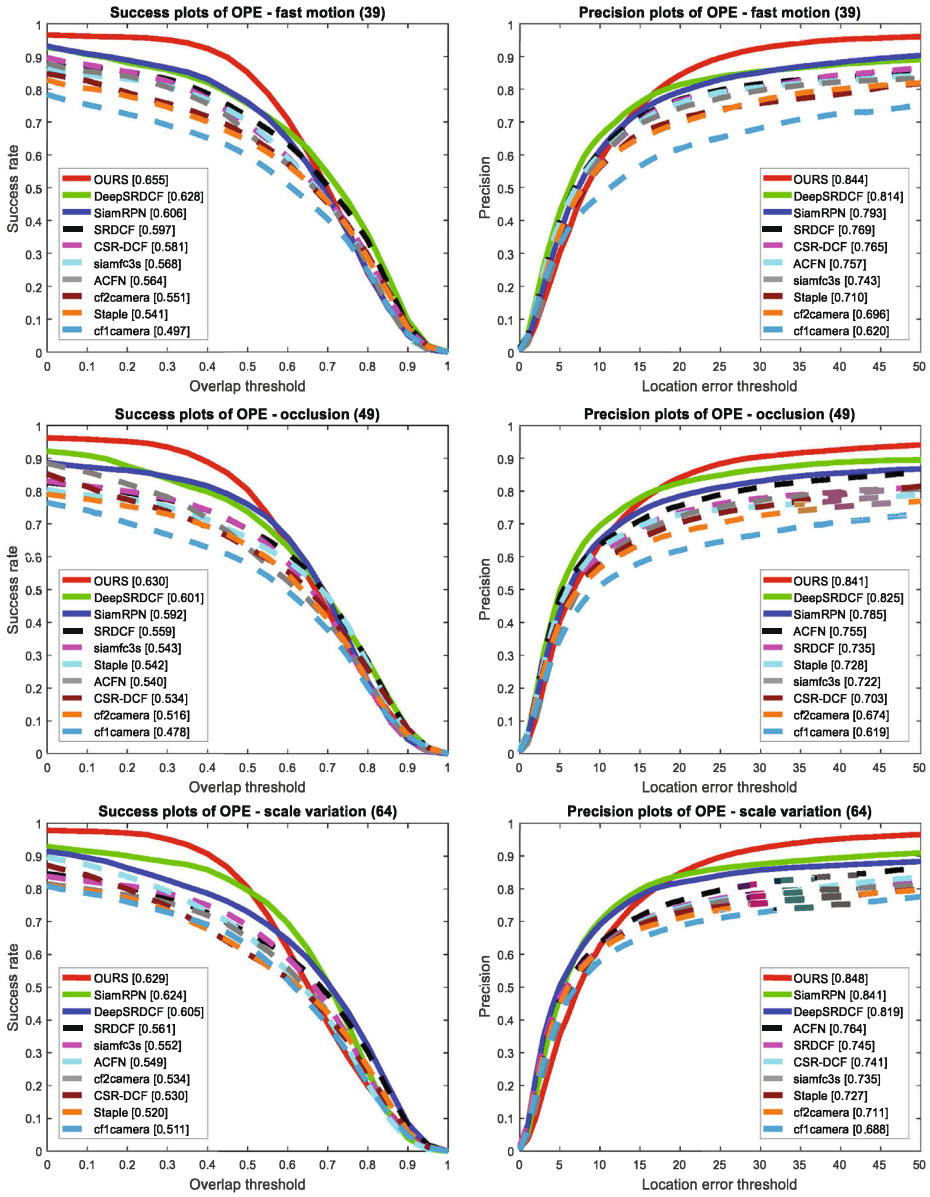

Fig. 6 Precision plots and success plots(AUC) over FM, OCC and SV attributes of 9 tracking algorithms on OTB100 dataset

Siamese network, our performance is better than that of SiamRPN, CFNet, and SiamFC, with relative improvements of 1%,5.7% and 6.1% in AUC, and improvements of 1.9%, 9.6% and 9.4% in precision score, respectively. Moreover, We perform better than CSR-DCF and ACFN, which adopt an attention mechanism, with relative improvements of 7.5% and 6.1% in AUC and 10% and 4.3% in precision score. Compared with DSiamM, which is

Table 1 Average Speed of 10 tracking algorithms on OTB2013, OTB50 and OTB100 datasets

Tracker	OTB-2013		OTB-50		OTB-100		
	AUC	Prec.	AUC	Prec.	AUC	Prec.	FPS
OURS	0.668	0.903	0.604	0.821	0.642	0.861	77
SiamRPN	0.658	0.884	0.592	0.810	0.637	0.851	200
DeepSRDCF	0.641	0.849	0.560	0.772	0.635	0.851	<1
SRDCF	0.626	0.838	0.539	0.732	0.598	0.789	5
cf2camera	0.611	0.807	0.530	0.702	0.568	0.748	70
Siamfc3s	0.607	0.809	0.516	0.692	0.582	0.771	86
ACFN	0.607	0.86	0.526	0.763	0.574	0.802	15
CSR-DCF	0.593	0.803	0.53	0.736	0.586	0.802	13
Staple	0.593	0.782	0.507	0.684	0.578	0.784	80
cf1camera	0.578	0.776	0.488	0.653	0.536	0.713	75
DsiamM	0.656	0.891					25

The first, second and third highest rates are highlighted in color

updated based on dynamic templates, our algorithm improves AUC by 1.2% and precision score by 1.2% and is three times faster than DSiamM in speed.

On the results of OTB-50 , the proposed algorithm achieves the best performance in both success and precision plots. The proposed algorithm also outperforms other online updated deep tracker, DeepSRDCF, on AUC of success plots, with relative improvements of 4.4% and 4.9%. The proposed algorithm performs better than recent siamese network-based trackers, SiamRPN, CFNet, and SiamFC. Although the proposed algorithm slower than SiamRPN, they get 1.2% and 1.1% relative improvement over SiamRPN, respectively, and both have real-time speed too. Other real-time trackers, Staple, is more likely to track the target with lower accuracy and robustness or may even lose the targets within longer sequences. Specifically, the proposed algorithm has achieved a relative improvement of 9.7% and 13.7% over Staple. These results verify the superior tracking effectiveness and efficiency of our approach.

On the results of OTB-100 , our proposed method, occupies the best one, outperforming the second-best tracker SiamRPN by a gain of 0.5% in AUC and 1% in precision score. Among the trackers using the siamese network, ours outperforms SiamRPN, CFNet, and SiamFC. SiamFC is a seminal tracking framework, but the performance is still left behind by the recent state-of-the-art methods. Even though CFNet and SiamRPN add a performance gain. Incorporating our attention mechanisms to the proposed tracker elevates to an AUC of 64.2% and precision score of 86.1%, leading to a consistent gain of 6%(9%), 7.4%(11.3%) and 0.5%(1%), compared to SiamFC, CFNet, and SiamRPN. Compared with attention-based CSR-DCF and ACFN, our algorithm not only scored higher in AUC and precision score but also faster in speed, which demonstrates that our method achieves robustness. What's more, the accuracy and success rate of attributes such as fast motion (FM), occlusion (OCC) and scale variation(SV) on OTB-100, our method also win the best.

As for the average speed illustrated in the Table 1, the top three algorithms are SiamRPN, Siamfc3s, and Staple with a respective speed at 200(FPS), 86 (FPS) and 80(FPS). Among

the three algorithms, the first two are based on the siamese network, while Staple is based on correlation filtering. The speed of the algorithm in this paper is 77(FPS), which is lower than the first two algorithms because the algorithm in this paper adopts an attention mechanism and template adaptive update strategy to increase the computational load. In spite of this, the algorithm in this paper is close to that of Staple in speed, but the algorithm in this paper has a big gap in AUC and precision score compared with Staple.

4.3.2 Experiments on VOT2016 and VOT2018

VOT2016 For the assessment on VOT2016, we report the performance of some of the best non-siam-based trackers for reference, including CCOT[8], DeepSRDCF[7], SRDCF[5], Staple[2] and CSRDCF[36]. And compared with other siam-based trackers, such as SiamRPN[29] and SiamFC[1]. The EAO curve evaluated on VOT2016 is presented in Figs. 7 and 6 other state-of-the-art trackers are compared. Table 2 and Fig. 7 shows the results of the proposed tracker are on par with that of the state-of-the-art algorithms and are the best with an EAO score of 0.353. The second best tracker, SiamRPN, is much faster than our tracker, while much lower in terms of EAO and Robustness, suggesting that the attention and template update mechanism introduced improves tracking performance. What's more, our tracker is much faster than CCOT, DeepSRDCF, and CSRDCF, which verifies that our tracker achieves a fast processing speed as well as excellent performance and shows a potential to the practical tracking application.

VOT2018 We compare the proposed tracker with 7 state-of-the-art tracking algorithms on VOT2018 dataset. These trackers are: SiamRPN[29], UPDT[3], RCO[26], ECO[6], CCOT[8], Dsiam[19] and SiamFC[1]. We evaluate the proposed method on VOT2018, and report the results in Table 3. As shown in Table 3 and Fig. 7, our method achieves the best EAO score 0.396 and the best accuracy score 0.588. Notably, our method sets a new state-of-the-art by improving 0.013 absolute value, i.e., 1.3% relative improvement, compared to SiamRPN, indicating that the attention and template update mechanism can significantly decrease the tracking failure.

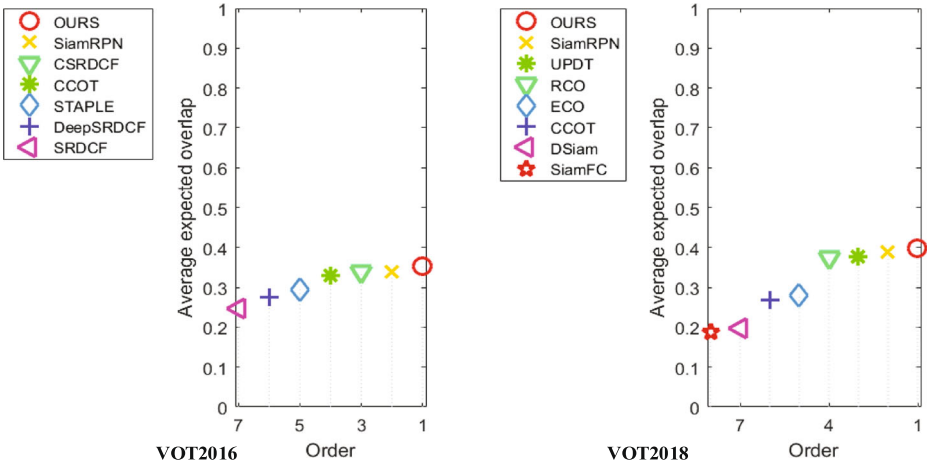

Fig. 7 EAO ranking with trackers in VOT2016 and VOT2018. The better trackers are located at the right. Best viewed on color display

Table 2 VOT2016 performance results

Attributes	Accuracy	Robustness	EAO
OURS	0.53	0.82	0.353
SiamRPN	0.56	1.12	0.340
CCOT	0.52	0.85	0.331
CSRDCF	0.51	0.85	0.338
Staple	0.54	1.35	0.295
DeepSRDCF	0.51	1.17	0.276
SRDCF	0.52	1.50	0.247
SiamFC	0.53	0.46	0.235

Red, green and blue fonts indicate 1st, 2nd, 3rd performance, respectively. Best viewed in color

4.4 Qualitative results

To visualize the superiority of the proposed algorithm, we show examples of our tracking results compared to recent trackers (SiamRPN[29], CSR-DCF [36], and SRDCF[5]) on challenging sample video sequences, as shown in Fig. 8.

(1)Fast Motion (FM) Figure 8 shows the tracking results of the 4 algorithms in Tiger1, Jumping and Coke video sequences, which under the condition of FM. In the Tiger1 video sequences, due to the target's fast walking, rotating and other factors, the algorithms SiamRPN, CSR-DCF, and SRDCF lose the target, but our method can accurately locate the target. In the Jumping video sequences, due to the target's jumping and other factors, the algorithms SiamRPN and CSR-DCF lose the target, but SRDCF and our method can accurately locate the target. In the Coke video sequences, due to the target's motion and other factors, the algorithms SRDCF and CSR-DCF have some drift, but SiamRPN and our method can accurately locate the target.

(2)Scale Variation (SV) Figure 8 shows the tracking results of the 4 algorithms in FaceOcc2, Ironman and Liquor video sequences, which have been through scale changes. In the

Table 3 VOT2018 performance results

Attributes	Accuracy	Robustness	EAO
OURS	0.588	0.223	0.396
SiamRPN	0.586	0.276	0.383
UPDT	0.536	0.184	0.378
RCO	0.507	0.155	0.376
ECO	0.484	0.276	0.280
CCOT	0.494	0.318	0.267
DSiam	0.215	0.646	0.196
SiamFC	0.503	0.585	0.188

Red, green and blue fonts indicate 1st, 2nd, 3rd performance, respectively. Best viewed in color

Fig. 8 Qualitative comparison of our method with state-of-the-art trackers on the Tiger1, Jumping, Coke, FaceOcc2, Ironman, Liquor, Matrix, Lemming and Bolt videos, under fast motion, scale variation and occlusion

FaceOcc2 video sequences, SiamRPN, CSR-DCF and SRDCF drift the target due to scale variation and illumination effect, while our method can save the stable positioning of the target. In the Ironman video sequences, SiamRPN, CSR-DCF and SRDCF lose the target due to scale variation and illumination effect, while our method can save the stable positioning of the target. In the Liquor video sequences, SiamRPN lose the target due to scale variation and similar object interference, while CSR-DCF, SRDCF and our method can save the stable positioning of the target.

(3)Occlusion (OCC) Figure 8 shows the tracking results of the 4 algorithms in Matrix, Lemming and Bolt video sequences, which have been partially or severely occluded by the target in respective. In the Matrix video sequences, when the target is rotated, the algorithms SiamRPN, CSR-DCF and SRDCF all lose the target, but our method can accurately track the target. In the Lemming video sequences, when the target is occluded, the algorithms CSR-DCF and SRDCF all lose the target, but SiamRPN and our method can accurately track the target. In the Bolt video sequences, when the target is occluded, the algorithm SRDCF lose the target, but CSR-DCF, SiamRPN and our method can accurately track the target.

In these 9 video sequences, other algorithms all have a certain degree of loss or drift, and the algorithm in this paper can accurately locate the target. The reasons that the proposed algorithm performs well can be explained by two main aspects. First, our algorithm contains attention mechanisms and fine-grained details that explain the appearance changes caused by deformation, rotation, and background clutter. Second, for template update, we focus on the previous frame of the target and update it appropriately to take into account the appearance changes.

4.5 Ablation Study

The proposed tracker has two important components, the siamese feature extraction subnetwork(ATT) and the object template updating subnetwork(UPT). We evaluate their concrete contributions in our method by removing each one and checking the performance of degraded trackers on OTB2013. The onlyATT module means that we use the siamese feature extraction subnetwork(ATT) to extract the feature information with only the first frame as the template frame. The onlyUPT module means that we use the first and previous frame as the template frame, and extract the template feature with the backbone network (AlexNet) without using the attention mechanisms. As shown in Table 4, the tracking accuracy decreases if we remove any component from the proposed method. Hence, all two components make positive contributions. Specifically, the second component 'UPT' contributes the most. The first component 'ATT' also plays an important role.

Table 4 Effectiveness of different components in the proposed method based on OTB-2013

	Our-onlyATT	Our-onlyUPT	Our
AUC	0.661	0.665	0.668
Prec.	0.889	0.896	0.903

Our proposed tracker contains two modules:the siamese feature extraction subnetwork(ATT) and the object template updating subnetwork(UPT). Best viewed in color

Table 5 Parameters setting of the proposed tracking algorithm

Parameter	Value	Description
η		parameter to balance the two tracking branches
θ		threshold of the target was lost in the previous frame
λ	1	hyper-parameter of the classification and regression parts
W_0	train by offline	parameter of channel attention
W_1	train by offline	parameter of channel attention

We listed 5 parameters and the corresponding values, some of which are set as the same as the SiamRPN tracker

Table 6 Tracking results on OTB-2013 of different η in the proposed tracker

value of η	0.3	0.5	0.7	0.9	0.95
AUC	0.629	0.642	0.660	0.668	0.663
Prec.	0.861	0.874	0.883	0.903	0.889

Best viewed in color

Table 7 Tracking results on OTB-2013 of different θ in the proposed tracker

value of θ	2	4	6	8	10
AUC	0.661	0.664	0.668	0.653	0.634
Prec.	0.886	0.889	0.903	0.879	0.865

Best viewed in color

Fig. 9 Failure examples of the proposed tracker on some representative sequences. Green bounding-boxes indicate ground-truth and yellow ones means our results. Left to right: Bird1 and Soccer

4.6 Parameter Analysis

In the proposed method, we have 5 parameters as shown in Table 5. λ is set as 1, followed by SiamRPN trackers, which is the base of the proposed tracking method. W_0 and W_1 are trained offline.

Among the other 2 parameters, η is the parameter to balance the two tracking branches and θ is the threshold of the target that was lost in the previous frame. Inspired by [10], we have conducted parameter analysis for η and θ, Table 6 and Table 7 list the result of setting different values of the two parameters.

For the parameter η, we set the value of η to 0.3, 0.5, 0.7, 0.9 and 0.95, report the tracking precision score and success score (AUC) for each setting, we observe that the proposed method achieves its best performance when η=0.9.

For the parameter θ, we set the value of θ to 2, 4, 6, 8 and 10, report the tracking precision score and success score (AUC) for each setting, we observe that the proposed method achieves its best performance when θ=6.

4.7 Failure case

Although the proposed tracker performs favorably against several state-of-the-art trackers in the benchmark datasets, the proposed method still has some limitations in some complicated scenes. Figure 9 illustrates two failure examples.

Firstly, let's start with the analysis of the video of Bird1. As a video with DEF (deformation), FM(Fast Motion) and OV(Out-of-View) at the same time, the target is completely occluded in multiple consecutive frames, the proposed method can hardly obtain the appearance model of the target. Hence the proposed method fails to track the target.

Now it is the discussion of the video of Soccer. As a video with IV(illumination change), SV(Scale Variation), OCC(Occlusion), MB(Motion Blur), FM(Fast Motion), IPR(In-Plane Rotation), OPR(Out-of-Plane Rotation), and BC(background clutters), the target is surrounded with some other similar objects (mainly come from the color of the video), which poses great challenges on the proposed method. Since the proposed method does not learn effective information to discriminate the target and background, it finally fails to locate the target.

5 Conclusion

In the process of visual tracking, the matching of templates and learning of features are crucial to the final tracking algorithm results. In this paper, we present an Adaptive Updating Siamese Network (AU-Siam) for real-time visual tracking. By fusing the siamese feature extraction subnetwork and the object template updating subnetwork with a seamless formulation, the robustness and accuracy of our method have been improved significantly. It is shown that the feature extraction model can be used to obtain more efficient and diverse features by attention mechanisms, which further enhances the robustness of our method. Furthermore, the template updating module can help update the target template in real-time, avoid template shifting problem and improve the final experiment result. It can be found from the experimental results that the proposed AU-Siam approach can outperform the state-of-the-art tracking methods. In the future, we plan to continue exploring the effective fusion of deep networks in tracking task.

References

1. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PH (2016) Fully-convolutional siamese networks for object tracking. In European Conference on Computer Vision Workshop, pages 850-865. Springer. 1, 2, 3, 4, 5, 7, 8
2. Bertinetto L, Valmadre J, Golodetz S, Miksik O, Torr PHS (2016) Staple: Complementary learners for real-time tracking. In The IEEE Conference on Computer Vision and Pattern Recognition, CVPR, June
3. Bhat G, Johnander J, Danelljan M, Khan FS, Felsberg M (2018) Unveiling the power of deep tracking. In: ECCV, pp 493–509
4. Choi J, Chang HJ, Yun S, Fischer T, Demiris Y (2017) Attentional correlation filter network for adaptive visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pages 4807-4816. 2 7
5. Danelljan M., Ager G. H., Khan F. S., Felsberg M. (2016) Adaptive decontamination of the training set: a unified formulation for discriminative visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pages 1430-1438. 4 7
6. Danelljan M, Bhat G, Khan FS et al (2017) ECO: Efficient convolution operators for tracking. Proc IEEE Conf Comput Vis Pattern Recognit, page 6
7. Danelljan M, Häger G et al (2015) Convolutional Features for Correlation Filter Based Visual Tracking. ICCV workshop
8. Danelljan M, Robinson A, Khan FS, Felsberg M (2016) Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: ECCV, pages 472-488. 1 2 5
9. Dong X, Shen J (2018) Triplet loss in siamese network for object tracking[C]. Proceedings of the European Conference on Computer Vision (ECCV). 459-474
10. Dong X, Shen J, Wang W et al (2018) Hyperparameter optimization for tracking with continuous deep q-learning[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 518-527
11. Dong X, Shen J, Wang W et al (2019) Dynamical Hyperparameter Optimization via Deep Reinforcement Learning in Tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence PP(99):1–1
12. Dong X, Shen J, Wu D et al (2019) Quadruplet network with One-Shot learning for fast visual object Tracking[J]. IEEE Trans Image Process 28(7):3516–3527
13. Dong X, Shen J, Yu D et al (2017) Occlusion-aware real-time object tracking[J], vol 19
14. Fan DP, Cheng MM, Liu JJ et al (2018) Salient objects in clutter: Bringing salient object detection to the foreground[C]. Proceedings of the European conference on computer vision (ECCV). 186-202
15. Fu K, Fan DP, Ji GP et al (2020) Siamese network for rgb-d salient object detection and beyond[J]. arXiv preprint arXiv:2008.12134
16. Fu K, Fan DP, Ji GP et al (2020) Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection[C].Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3052-3062
17. Fu K, Zhao Q, Gu IYH et al (2019) Deepside: A general deep framework for salient object detection[J]. Neurocomputing 356:69–82
18. Gong C, Tao D, Liu W et al (2015) Saliency propagation from simple to difficult[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2531-2539
19. Guo Q, Feng W, Zhou C, Huang R, Wan L, Wang S (2017) Learning dynamic siamese network for visual object tracking. In: ICCV. 1
20. He A, Luo C et al (2018) A twofold siamese network for real-time object tracking[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4834–4843
21. Henriques J, Caseiro R, Martins P, Batista J (2015) Highspeed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(3):583–596. 1 3 7
22. Hu H, Ma B, Shen J et al (2018) Robust object tracking using manifold regularized convolutional neural networks[J]. IEEE Transactions on Multimedia 21(2):510–521
23. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 7132-7141
24. Jianbing S, Xin et al (2019) Visual Object Tracking by Hierarchical Attention Siamese Network.[J] IEEE transactions on cybernetics
25. Kristan M, Leonardis A et al (2016) The visual object tracking vot2016 challenge results. In: ECCV, pp 777–823
26. Kristan M, Leonardis A et al (2018) The sixth visual object tracking VOT2018 challenge results. In: ECCV, pp 3–53
27. Krizhevsky A, Sutskever I et al (2012) Imagenet Classification with Deep Convolutional Neural Networks[C]. NIPS Curran Associates Inc

28. Li B, Wu W et al (2019) Siamrpn++:, Evolution of siamese visual tracking with very deep networks[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4282–4291
29. Li B, Yan J, Wu W, Zhu Z, Hu X (2018) High performance visual tracking with siamese region proposal network. In: CVPR. 1, 2, 3, 4, 5, 8
30. Liang Z, Shen J (2019) Local semantic siamese networks for fast tracking[J]. IEEE Trans Image Process 29:3351–3364
31. Liu F, Gong C, Huang X et al (2018) Robust visual tracking revisited: From correlation filter to template matching[J]. IEEE Trans Image Process 27(6):2777–2790
32. Liu F, Gong C, Huang X et al (2018) Robust visual tracking revisited: From correlation filter to template matching[J]. IEEE Trans Image Process 27(6):2777–2790
33. Liu P, Yu H, Cang S (2019) Adaptive neural network tracking control for underactuated systems with matched and mismatched disturbances[J]. Nonlinear Dynamics 98(2):1447–1464
34. Lu X, Ni B, Ma C et al (2019) Learning transform-aware attentive network for object tracking[J]. Neurocomputing 349:133–144
35. Lu X, Wang W, Ma C et al (2019) See more, know more: Unsupervised video object segmentation with co-attention siamese networks[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 3623-3632
36. Lukezic A, Vojir T, Cehovin Zajc L, Matas J, Kristan M (2017) Discriminative correlation filter with channel and spatial reliability
37. Ren S, Girshick R et al (2017) Faster r-CNN: towards Real-Time object detection with region proposal Networks[J]. IEEE Transactions on Pattern Analysis Machine Intelligence 39(6):1137–1149
38. Sun L, Zhao C, Yan Z et al (2018) A novel weakly-supervised approach for RGB-D-based nuclear waste object detection[J]. IEEE Sensors J 19(9):3487–3500
39. Tang Z, Li C, Wu J et al (2019) Classification of EEG-based single-trial motor imagery tasks using a B-CSP method for BCI[J]. Frontiers of Information Technology & Electronic Engineering 20(8):1087–1098
40. Tang Z, Yu H, Lu C et al (2019) Single-Trial Classification of Different Movements on One Arm Based on ERD/ERS and Corticomuscular Coherence[J]. IEEE Access 7:128185–128197
41. Tao R, Gavves E, Smeulders AWM (2016) Siamese instance search for tracking. In IEEE Conference on Computer Vision and Pattern Recognition. 1, 2, 3, 7
42. Valmadre J, Bertinetto L, Henriques JF, Vedaldi A, Torr PH (2017) End-to-end representation learning for correlation filter based tracking. In: CVPR. 1, 2, 3, 4, 8
43. Wang W, Lu X, Shen J et al (2019) Zero-shot video object segmentation via attentive graph neural networks[C].Proceedings of the IEEE international conference on computer vision. 9236–9245
44. Wang W, Shen J, Ling H (2018) A deep network solution for attention and aesthetics aware photo cropping[J]. IEEE transactions on pattern analysis and machine intelligence 41(7):1531–1544
45. Wang Q, Teng Z, Xing J, Gao J, Hu WS (2018) Maybank.Learning attentions: Residual attentional siamese network for high performance online visual tracking. In: CVPR. 1 2
46. Woo S, Park J, Lee JY et al (2018) CBAM : Convolutional Block Attention Module[J]
47. Wu Y, Lim J, Yang M-H (2013) Online object tracking: a benchmark. In: CVPR. 2
48. Wu Y, Lim J, Yang M-H (2015) Object tracking benchmark.TPAMI, 1, 2, 5, 6, 7
49. Xiao Y, Li J, Du B, Wu J, Chang J, Zhang W (2020) Memu: Metric Correlation Siamese Network and Multi-class Negative Sampling for Visual Tracking. Pattern Recognition, Volume 100. https://doi.org/10.1016/j.patcog.2019.107170
50. Yilmaz A., Javed O., Shah M. (2006) Object tracking: a survey. ACM Comput Surv 38(4):1–45
51. Z T, Ghanem B, Liu S, Ahuja N (2012) Robust visual tracking via multi-task sparse learning. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR
52. Zhang Z, Lai Z, Huang Z et al (2019) Scalable supervised asymmetric hashing with semantic and latent factor Embedding[J] IEEE transactions on image processing
53. Zhang Z, Peng H (2019) Deeper and wider siamese networks for real-time visual tracking[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4591-4600
54. Zhang Y, Wang L et al (2018) Structured siamese network for real-time visual tracking[C].Proceedings of the European conference on computer vision (ECCV). 351-366
55. Zhou Y, Li J, Du B, Chang J, Xiao Y (2020) A Target Response Adaptive Correlation Filter Tracker with Spatial Attention. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-020-08839-0
56. Zhu Z, Wang Q, Li B, Wu W, Yan J, Hu W (2018) Distractor-aware siamese networks for visual object tracking. In: ECCV. 1, 2, 3, 6, 7, 8

Yifei Zhou received the M.S degree in Software Engineering from Wuhan University in 2013, Wuhan, China. She is currently working for the Ph.D. degree also at Wuhan University. Her main research interests include data mining and pattern recognition.

Jing Li received the Ph.D. degree from Wuhan University, Wuhan, China, in 2006. He is currently a Professor in Computer School of Wuhan University, Wuhan, China. His research interests include data mining and multimedia technology.

Bo Du (Senior Member, IEEE) received the B.S. degree and the Ph.D. degree in photogrammetry and remote sensing from State Key Lab of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2005 and 2010, respectively. He is currently a "Luojia Talented Young Scholar" Professor appointed by the Wuhan University, China, which is the most prestigious chair professor title for young staff in the university. He has more than 40 research papers published in the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING(TGRS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE JOURNAL OF SELECTED TOPICS IN EARTH OBSERVATIONS AND APPLLED REMOTE SENSING (JSTARS), and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL). Five of them are ESI hot papers or highly cited papers. His major research interests include data mining, pattern recognition, hyperspectral image processing, and signal processing. Dr. Du received the best reviewer awards from the IEEE GRSS for his service to the IEEE JSTARS in 2011 and the ACM rising star awards for his academic progress in 2015. He was the Session Chair of the International Geoscience and Remote Sensing Symposium 2016 and the 4th IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing. He also serves as a Reviewer of 20 Science Citation Index magazines, including the IEEE TGRS, TIP, JSTARS, and GRSL.

Jun Chang received the Ph.D. degree from Wuhan University in 2011, Wuhan, China. He is currently an Assistant Professor in School of Computer Science, Wuhan University, Wuhan, China. His current research interests include computer vision, large-scale machine learning and stream data mining.

Zhiquan Ding received the B.S. degree in Information material and the M.S. degree in Signal and Information Processing from University of Electronic Science and Technology of China, Chengdu, China, in 1998 and in 2007, respectively. He is a Senior engineer at Sichuan institute of aerospace electronic equipment and at R&D Center of Intelligent Detection and Recognition Technology of Multisensor. His research interests are in target detection and tracking, and signal processing and Information fusion.

Tianqi Qin received the B.S. degree in Mathematics and Applied Mathematics (Pure) and the Ph.D. degree in Mathematics in Uncertainty Processing from Sichuan University, Chengdu, China, in 2011 and in 2016, respectively. She is an engineer at Sichuan institute of aerospace electronic equipment and at R&D Center of Intelligent Detection and Recognition Technology of Multisensor. Her research interests are in target detection and tracking, and signal processing, and Information fusion.