Multimedia Tools and Applications (2023) 82:15199–15222
https://doi.org/10.1007/s11042-022-13930-9

Online bionic visual siamese tracking based on mixed time-event triggering mechanism

Huanlong Zhang[1] · Zhuo Zhang[1] · Jiapeng Zhang[1] · Yanchun Zhao[2] · Miao Gao[3]

Received: 28 March 2022 / Revised: 29 June 2022 / Accepted: 12 September 2022 /
Published online: 29 September 2022

Abstract
Existing Siamese-based trackers deal with target deformation and occlusion by introducing online updates. However, these trackers still suffer from model drift due to the cumulative error in tracking results and the lack of a suitable model update strategy. To solve this problem, we propose an online bionic visual siamese tracking framework based on the mixed time-event triggering mechanism. In which, the bionic vision network introduces the receptive field block and the blurpool, which improve the quality of feature extraction while maintaining the translational invariance of the convolutional neural network. The former uses dilated convolution kernels with different dilation rates to fuse depth features, which effectively increases the receptive field of the network. The latter uses low-pass filtering to anti-alias before downsampling, reducing the negative impact of the downsampling operation on the generalization ability of the network. In addition, to enable the model to effectively capture target appearance variations, a template update strategy with the mixed time-event triggering mechanism is designed. The strategy evaluates the quality of tracking results via a quality assessment model, guided by the mixed time-event triggering mechanism to adaptively weighted fusion of fixed and mutative templates. Numerous experiments conducted on OTB100, VOT2016, VOT2018, UAV123, GOT-10k benchmarks show that the proposed tracker outperforms the baseline tracker and achieves state-of-the-art performance.

Keywords Object tracking · Siamese network · Template-updating · Deep learning

1 Introduction

Visual target tracking is one of the most fundamental and challenging tasks in computer vision. Given the bounding box of the target in the first frame of the video, the tracker locates the target in all subsequent frames. Although visual target tracking has a large number of applications in human-computer interaction and autonomous vehicle navigation, it is still challenging to design target tracking algorithms in scenarios such as deformation, occlusion.

✉ Huanlong Zhang
zhl_lit@163.com

Extended author information available on the last page of the article.

To overcome the above challenges of tracking scenarios and obtain stable tracking performance, researchers make efforts in visual tracking. However, most of the existing methods [17, 27, 35, 40, 44] mainly focus on modeling the appearance of the target. Although these siamese tracking algorithms have a balanced speed and accuracy, there is still an important problem: most of these siamese tracking algorithms are not able to update the templates during the tracking process. Their fixed target appearance model ensure high tracking speed, but sacrifice the ability of the tracker to better adapt to the target appearance deformation. These algorithms also have difficulty identifying objects in partially occluded scenes, due to the dynamic nature of occlusion that causes occluded objects to take on different appearances over time. As a result, the tracking is particularly prone to failure in the presence of occlusion and deformation, as shown in Fig. 1.

During the tracking process, it is difficult for the static features of the target to reflect the appearance change of the dynamic target in time, which is the main reason for the serious performance degradation of the tracker in the deformation and occlusion scenarios. Therefore, some researchers try to introduce a template update method or improve the tracking network, to enable the algorithm to adapt to the change of target appearance in time. Xu et al. [45] propose a template update network that consists of two independent networks: a contour network and a detection network. The maximum response of the response map is compared with a fixed threshold to determine whether to activate the template update network. If the network is activated, a new appearance model is regenerated by selecting the template with the highest confidence in the past. This method of updating templates with a fixed threshold. In the early stage of tracking, a lot of computing power will be consumed,

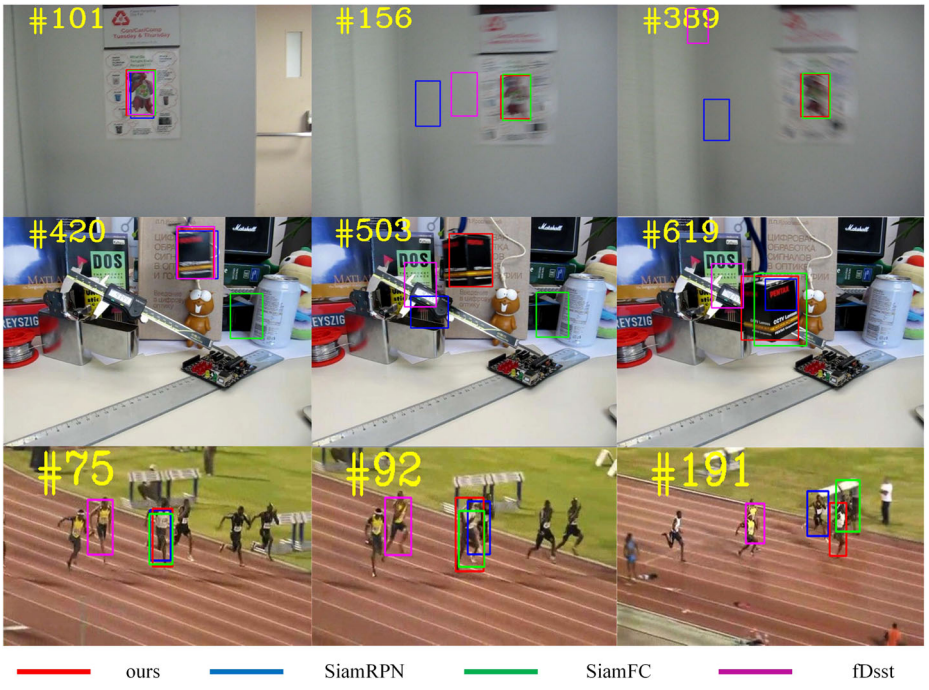

Fig. 1 Compared with fDsst [6], SiamFC and SiamRPN [28], the proposed tracking algorithm performs well in target deformation scenarios

affecting the tracking speed, and the update frequency will be significantly reduced in the later stage, affecting the tracking performance. Guo et al. [16] design a target appearance variation transformation as a single network layer to adapt to target changes. The transformation uses the first frame template and the previous frame template to solve the linear transformation matrix in the frequency domain. The calculated linear transformation matrix is used in the target feature extraction branch of the static siamese network to dynamically update the template feature. However, the algorithm is proposed under the assumption that the target changes are smooth in time, so the adaptation of the algorithm to the tracking scene is limited. Zhou et al. [52] propose an adaptive updating siamese network, which includes two sub-networks. The first sub-network extracts target features, introducing channel and spatial attention modules, which can selectively amplify valuable features. The second sub-network utilizes the first frame and the previous frame to generate two response maps summed with fixed weights for localization and regression. This method requires frequent feature extraction operations to maintain tracking performance, and the fixed weight fusion method is not well suited to various tracking scenarios. Siamese block attention network for online update object tracking is proposed by Xiao et al. [43], mainly for appearance change and occlusion problems. This method proposes a Siamese block attention module, which relies on the characteristics of max pooling and average pooling to generate channel weights, enhance key features, and suppress irrelevant features. The method also proposes a template update method that relies on the linear addition of the initial template, the process template and the final tracking result with fixed weights. In short, the efforts of a large number of researchers make these online update-based siamese algorithms achieve good performance.

In the above algorithms, we summarize two common characteristics: First, most algorithms extract target features to update templates by convolutional neural networks. Second, most algorithms update strategy are single or too complex. For the first characteristic. We analyze that the reason why these algorithms use convolutional neural networks is that the feature quality of the same target extracted by convolutional neural networks is stable. However, the quality of features extracted by convolutional neural networks for the same target at different locations can fluctuate greatly. The reason for this phenomenon is that convolutional neural networks are not translation invariant [1, 49]. In the template update algorithm relying on target features, the loss of translation invariance of the convolutional neural network will lead to the extraction of low-quality target features. The low-quality target features will gradually pollute the model and cause the tracking to drift or even fail. For the second characteristic, we believe that the update requirements of target templates need to be adjusted for different tracking situations. A single or overly complex update strategy will always be difficult to balance the performance and speed of the algorithm, either consuming a lot of computing power and affecting the tracking speed, or increasing the tracking speed and affecting the tracking performance.

In this paper, we propose an online bionic visual siamese tracking framework based on the mixed time-event triggering mechanism. The novel bionic visual network, which exploits the receptive field block that mimics the human visual perceptual domain to expand the perceptual field of the network and uses the blurpool that alleviates the loss of translational invariance, improves the stability and quality of feature extraction. Secondly, we propose a template update strategy with the mixed time-event triggering mechanism, which enables the model to adapt to target appearance variations timely, and ensures its performance and speed balance. The template update strategy consists of a quality assessment model, an adaptive template fusion mechanism and a mixed time-event triggering mechanism. The quality assessment model evaluates the tracking quality and collects templates

through an online updated classification network. The adaptive template fusion mechanism adaptively weights the collected templates to obtain the latest target feature changes. The mixed time-event triggering mechanism effectively regulates the performance and speed of the tracker. Finally, to evaluate our proposed approach, we conducted extensive experiments on publicly available benchmark datasets OTB100 [42], VOT2016 [13], VOT2018 [23], UAV123 [31], GOT-10k [21]. The experimental results show that our method plays an important role in confronting the target deformation and occlusion problem.

The main contributions of this paper can be summarized as follows:

(1) We propose a novel bionic visual network that can effectively improve the quality and stability of feature extraction, which combines the receptive field block that mimics the human visual perceptual domain and the blurpool that alleviates the loss of translational invariance.
(2) We propose a template update strategy with the mixed time-event triggering mechanism to allow the tracker to adapt to the feature changes of the target in time. The strategy consists of a quality assessment model, an adaptive template fusion mechanism, and a mixed time-event triggering mechanism. The quality assessment model relies on an online learning classification network to evaluate the prediction results. The adaptive template fusion mechanism adaptively fuses the target and dynamic templates based on the evaluation values. The mixed time-event triggering mechanism combines time and threshold to jointly regulate the update frequency of the template update strategy.
(3) The method is extensively experimented on several publicly available benchmark datasets and achieves better performance than several state-of-the-art trackers.

The rest of this paper is organized as follows: Section 2 reviews previous research in the field of object tracking. Section 3 introduces the proposed online bionic visual siamese tracking framework based on the mixed time-event triggering mechanism. In Section 4, we present the experimental results and compare our method with other state-of-the-art tracking methods. The conclusion of this paper is presented in Section 5.

2 Related work

In Siamese-based tracking, the search area is cropped from the current frame according to the target tracking result of the previous frame, and the template is cropped from the initial frame. In deformation or occlusion scenarios, the semantic information of the initial object is not enough to match the current object, which is the main reason for the failure of tracking. Therefore, online update based siamese trackers are proposed. Most of these trackers generate multiple response maps by adding an update template or template pools, and linearly adds multiple response graphs to locate targets. This method not only puts forward higher requirements on the generalization ability of the network, but also requires a suitable update strategy to effectively balance speed and performance. In response to the above requirements, the first contribution of this paper is the novel bionic visual network. By introducing two special network structures, the network not only weakens the loss of translation invariance, but also expands the receptive field of the network. The network can extract target features stably and efficiently. The second contribution is the template update strategy with the mixed time-event triggering mechanism. This strategy collects high-quality templates, adaptive fusion features, and a special update trigger mechanism during the tracking process, which effectively balances the speed and performance of the tracker.

Siamese network based trackers SiamFC approaches tracking modeling as a similar learning problem, and uses a fully convolutional siamese network to extract the features of the target template and the search region, then uses a simple intercorrelation operation to perform a sliding window evaluation of the search region. SiamFC achieves speed beyond real-time tracking, benefiting from a network trained offline and no need for online updates. However, robustness and recognition capability of SiamFC are still insufficient under deformation and occlusion scenarios. SiamRPN [28] introduces a region proposal network into SiamFC, which gets rid of the traditional multi-scale testing, resulting in a certain performance improvement in deformation scenarios, and the algorithm shows high performance on multiple challenging datasets [23–25]. SiamMN [12] designs a new feature extraction method based on SiamRPN. By fusing the template with the three-layer features of the search area, SiamMN obtains deep features containing more semantic information, so that the target can obtain a higher response on the score map. SAsiam [17] constructs a twofold Siamese network, in which one branch learns semantic information and the other branch learns appearance information, and combines the two to improve the localization accuracy of the tracker. CFNet [39] introduces a correlation filter layer in the Siamese backbone network, which is capable of end-to-end training. This way of constructing the network reduces the amount of parameters without sacrificing accuracy. This method is also one of the commonly used methods in subsequent online update algorithms. In the latest research of some Siamese algorithms, various attention modules are often introduced in the feature extraction network, and the localization accuracy of the tracker is improved. Zhu et al. [54] add multiple attention mechanisms to the basic backbone network, and design a template search collaborative attention module. This module obtains the context information of the target through global average pooling and one-dimensional convolution, and uses it in combination with the template feature to improve the accuracy of target positioning. Compared with previous siamese-based trackers, we design the bionic vision network that can extract object features stably and efficiently during tracking. Moreover, since the loss of translation invariance is alleviated, the network can stably extract high-quality target features. It provides a solid foundation for the design of template update strategies and prevents tracker degradation.

Convolutional neural network It is well known that convolutional neural networks revolutionize computer vision, allowing the fields of image recognition, target detection, and target tracking to advance by leaps and bounds. In the research process of network architecture, many complex deep architectures have appeared, such as VGGNet [37], AlexNet [26], ResNet [18], etc. VGGNet uses a small-scale convolution kernel, reduces network parameters, and adjusts the network structure to facilitate parallel acceleration using hardware. AlexNet adds the relu activation function, which improves the training speed, and uses the dropout operation to alleviate overfitting. ResNet alleviates model degradation and deepens the number of neural network layers through the residual structure. These networks facilitate the development of vision algorithms. The key to the success of these vision algorithms is the inductive bias of the method. In particular, the choice of convolution and pooling in the convolutional neural network is motivated by the desire to endow the networks with invariance to irrelevant cues such as image translations, scalings, and other small deformations. For the Siamese-based tracker, the position of the object changes due to both displacement and distortion, which requires the network to have translation invariance to ensure the performance of the tracker. However, according to the recent research [1, 49], modern convolutional networks are not shift-invariant, as small input shifts or translations can cause drastic changes in the output. The lost of translation invariance affects the quality of

target features extracted by the network. Low-quality template features are detrimental to the study of template update algorithms. To extract the target features stably and effectively, this paper introduces blurpool and receptive field block into the convolutional neural network. Unlike other convolutional neural networks for the task of improving performance, the bionic visual network aims to alleviate the lost of translation invariance, improve the quality of extracted features, and provide rationality for the template update strategy of the tracking algorithm.

Template update Most trackers [3, 33, 55] do not update the initial template during the tracking process. It is difficult for these trackers to perform well in deformation and occlusion scenarios, due to the difficulty in obtaining recent feature changes of objects. To track stably and continuously, some algorithms [4, 22, 36] update the target template during the tracking process to improve the tracking accuracy. The most common of these template update strategies is to update at a fixed frame interval or every frame. Zhao et al. [51] propose a template updating method via reinforcement learning, which is used to determine whether each template in the template pool should be replaced by constructing an actor and critic network. Multiple templates in the template pool and a search area obtain multiple response maps to locate the current target position. The method updates the parameters of the network at each frame during the tracking process. Wei et al. [41] introduce a squeeze-and-excitation block in the Siamese backbone network, which enhance the network's ability to perceive the target, and introduce the template update mechanism of UpdateNet [48], relying on accumulated templates and an offline trained network to predict the target feature. This method also requires frequent updating of the template feature during tracking. These algorithms are easily affected by low-quality templates during tracking, which eventually lead to tracking failure, and the fixed update method also affects the speed of tracking. Some researchers design template quality assessment methods to identify updated target templates. Yuan et al. [47] introduce average peak-to-correlation energy (APCE) to determine whether to update the template. APCE mainly reflects the fluctuation degree of the response map and the confidence level of the detection target. Yang et al. [46] employ a long and short-term memory method that takes previously collected templates to estimate the current template, but this model is a rather complex network structure and computationally expensive. A single or overly complex update method is difficult to adapt to the changing tracking environment, which requires the template update mechanism to change the update strategy according to the tracking situation, effectively balancing the performance and speed of the tracker. Therefore, we propose a template update strategy with the mixed time-event triggering mechanism, which not only reduces the possibility of template contamination, but also reduces the computational effort of frequently computing useless template features. Through extensive experiments, we demonstrate that this strategy is effective.

3 The proposed approach

The goal of our approach is to address the problem that the tracker has difficulty adapting to target appearance variations during target tracking. The flow chart of our proposed approach is shown in Fig. 2, which illustrates the whole tracking process. It can be observed from Fig. 2 that we first extract the initial template feature, search region feature, and X-frame template feature with the bionic visual network. The initial template feature is convolved with the search area features to obtain response map I. X-frame template feature is convolved with the search area features to obtain response map X. The response map I and

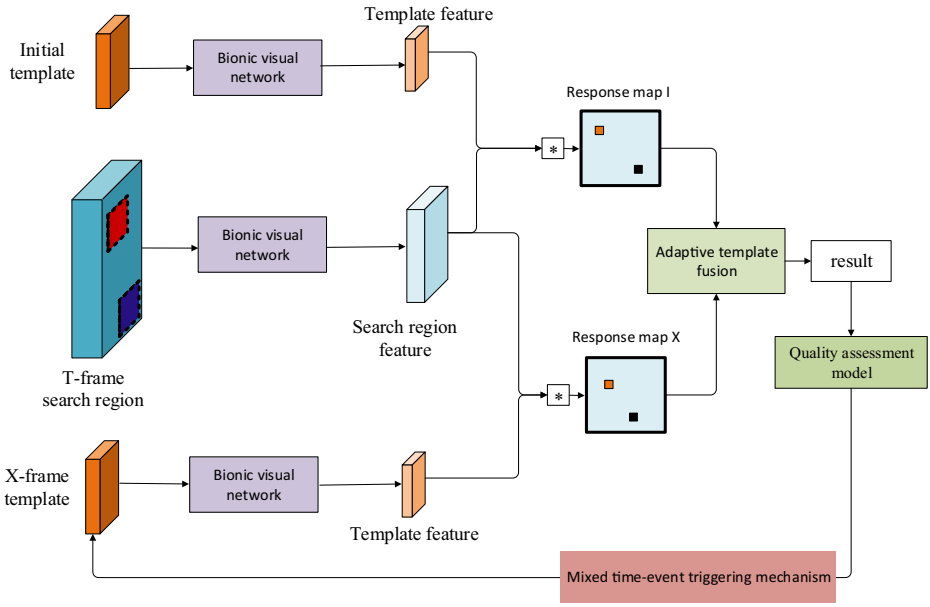

Fig. 2 Flowchart of the proposed approach

response map X are fused to obtain the prediction result based on the fused response map. Finally, the prediction result is input into the quality assessment model for quality evaluation, and the mixed time-event triggering mechanism combines the evaluation value and time to decide whether to update the template.

3.1 Basic tracker SiamFC

The Siamese-based trackers essentially find the region in the image which is most similar to the given bounding box. SiamFC propose to exploit the Siamese network [38] to learn the similarity measure function f. SiamFC applies an identical transformation ϕ to both inputs respectively denoted by z and x and then compares the similarity of representations using predefined metric function g according to $f(z, x) = g(\phi(x), \phi(z))$. The function g is fully convolutional with respect to the search image x , which enables unequal size between z and x and allows the similarity function to be computed for all translated sub-windows within x in one evaluation. The embedding function ϕ resembles the convolutional layers of AlexNet. After training, for each new frame, SiamFC searches for the target in a region two times the size of the initial annotated bounding box centered on the target position of the previous frame and obtains the response map by intercorrelation. The maximum position on the response map is identified as the new target position. This idea of treating tracking as similar learning enables SiamFC to achieve good performance and speed. However, this idea also leads to the template adaptability and feature expression of the target are key for siamese-based trackers to localize targets by similarity evaluation.

3.2 Architecture of the designed bionic visual network

In this paper, we introduce receptive field block and blurpool into AlexNet to construct a bionic vision network. The network enables to provide a more comprehensive description

for target feature extraction and alleviates the lost of translation invariance of convolutional neural networks. The property that the network maintains the translational invariance of the convolutional neural network also provides rationality for the subsequent template update strategy.

Figure 3(a) depicts the network structure of SiamFC. In SiamFC, the feature maps extracted by AlexNet are used for object location. However, the target features extracted by AlexNet are not detailed and comprehensive enough. In the target deformation scenario, it will cause a large difference between the target candidate frame and the target template. Unlike AlexNet, the proposed bionic visual network adopts two novel modules: the receptive field block and the blurpool, to improve location accuracy. The former exploits dilated convolution of different sizes to capture information over larger regions and in more contexts. This structure improves the distinguishability and robustness of the features. The latter introduces classic anti-aliasing, which improves the translation invariance of the convolutional neural network. The structure of the bionic visual network is shown in Fig. 3(b).

3.2.1 The receptive field block

SiamFC uses the first five convolutional layers of AlexNet to extract the target features, the deep low-resolution feature undergoes multiple convolutional operations to extract

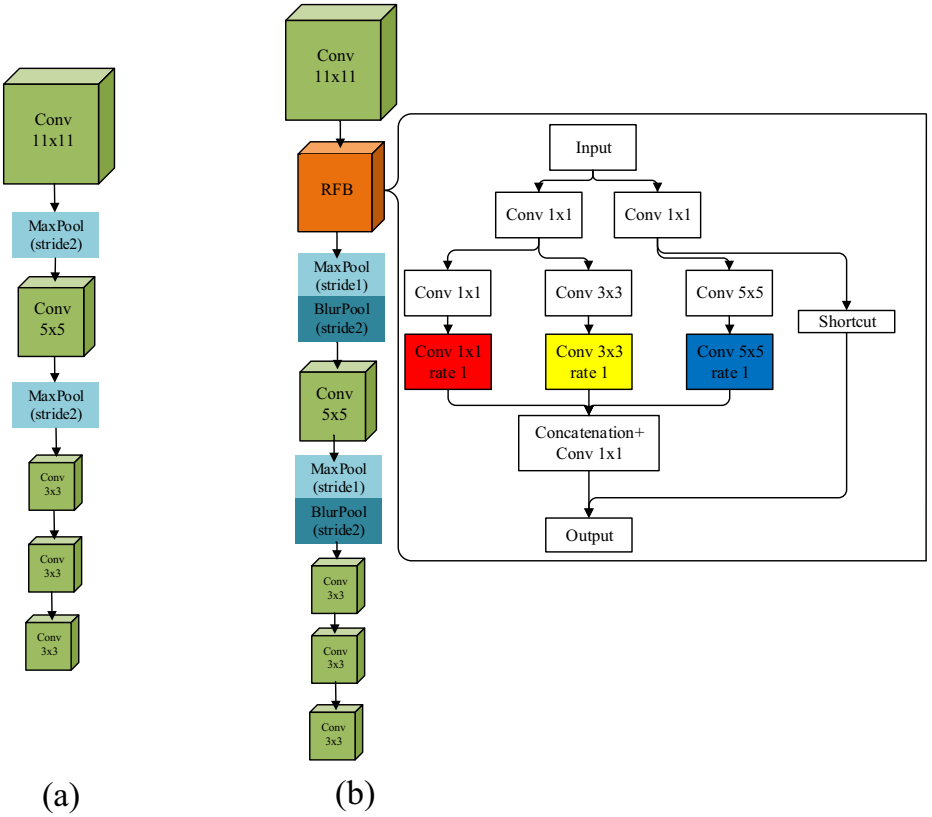

Fig. 3 Network architecture of the designed bionic visual network

rich semantic information. However, due to excessive downsampling operations, a lot of detailed information about the target is lost, making the model more robust but with reduced localization accuracy.

To produce more discriminative and robust features, in our implementation, we keep the convolution operation of AlexNet and add the receptive field block between the first and second convolution layers of AlexNet.

The first step of the receptive field block employs 1∗1 convolutional layer in each branch to reduce the number of channels in the feature map, then passes through n∗n convolutional layers. The second step adds three different sizes of dilated convolutions, 1∗1, 3∗3 and 5∗5. The feature maps are padded before the dilated convolution operation to ensure that the three branches output the same feature size. Eventually, the feature maps of all the branches are concatenated, merging into a spatial pooling or convolution array as in Fig. 3(b). The receptive field block imitates the perceptual field structure of the human visual systems. The dilated convolutions of different sizes capture multi-scale information about the target and maintain some critical details. This allows the network to extract more discriminative and robust features.

3.2.2 The blurpool

According to recent studies [1, 49], the convolutional neural network is not translation invariant, which leads to a large change in the features of the same target in different states extracted by the convolutional neural network. We believe that the lost translational invariance of convolutional neural networks is one of the reasons, which results in the performance degradation of some siamese-based trackers after the introduction of template updates. Therefore, to alleviate the lost of translation invariance of the convolutional neural network, we introduce blurpool into the basic network. The blurpool integrates low-pass filtering to anti-alias. In our implementation, the blurpool (stride=2) is combined with the maxpool (stride=1) to replace all the maxpool (stride=2) in AlexNet as in Fig. 3(b). The introduction of blurpool maintains the translation invariance of the convolutional neural network and improves the generalization ability of the network. This provides rationality in the subsequent design of the template update strategy.

3.2.3 Ground-truth and Loss

During the whole process of end-to-end training, the binary cross entropy loss is used for classification loss. Let R denote the network output, which is obtained by convolution operation and sigmoid operation of the target template feature with the search region feature. V denotes the sample label, the center point of this label is set to 1 within the radius of the positive label, and the rest is set to 0 (in this work, radius = 16). W denotes the label weight, which is determined by the number of positive and negative labels in V. The loss function formula is as follows:

$$L_{cls} = -mean\left((V \times log\left(R\right) + (1 - V) \times log\left(1 - R\right)) \times W\right) \tag{1}$$

As shown in Fig. 4, we substitute the tracking network of the basic tracker with the bionic visual network for tracking tests and visualize the response maps. These response maps show a higher response at the center of the target.

Fig. 4 (a)Initial template. (b)Search region. (c)The response map obtained by the basic tracker using AlexNet. (d)The response map obtained by the basic tracker using Bionic Visual Network (BVNet)

3.3 The template update strategy with mixed time-event triggering mechanism

During target tracking, the appearance information of the template is the key to the tracking results. The first frame of a tracking video sequence contains a lot of highly reliable target appearance information. Therefore, most of the siamese-based trackers only use the template given in the first frame for feature models. However, only the first frame is used as the fixed template without updating the target template. When the target appearance changes greatly, it is easy to lose the target. The recent frame is very related to the current state of the target, which contains more useful information about the appearance changes of the target. However, if the recent frame is used directly as a template without quality assessment, the tracking shift will occur in all subsequent frames when the recent frame is tracking drift. Therefore, this paper proposes a template update strategy with the mixed time-event triggering mechanism, which combines the quality-evaluated recent frame template with the first frame template to generate the final response map, to improve the tracking robustness. And to balance the performance and speed of the algorithm, the mixed time-event triggering

mechanism is used to guide the template update. Next, we explain the template update strategy in three parts, the quality assessment model, the adaptive template fusion mechanism, and the mixed time-event triggering mechanism.

3.3.1 The quality assessment model

We adopt Mdnet [32] as our base quality assessment model. The model consists of three convolutional layers and two fully connected layers. It crops the input image to 107∗107∗3 and outputs a single evaluation value S. The mdnet is trained online by long-short term updates and collects a large number of positive and negative samples. These strategies result in Mdnet to run at a very slow speed, only 5 frames per second. To maintain the real-time nature of the algorithm, we replace the long-short update strategy with a fixed frame update interval, (In this work, the value is 10). During the tracking process, the predicted result is fed into the quality assessment model, and when the evaluated value is greater than the set threshold γ, (in this work, $\gamma = 0.6$), it is further decided by the mixed time-event triggering mechanism whether to update the dynamic template.

3.3.2 The mixed time-event triggering mechanism

The template update based target tracking algorithm still has many problems, due to the fact that each time the acquirement of target features needs to be computed by the convolutional network. If we only rely on the fixed threshold to determine whether the algorithm template is updated, which often consumes a lot of computational power and results in a decrease in tracking speed. However, to improve the tracking speed, an update strategy using fixed interval frames is chosen, which makes the tracker unable to obtain some feature changes of the target in time. It is a key issue to balance the performance and speed of the tracking algorithm. Therefore, we design the mixed time-event triggering mechanism, which achieves a finer balance between the performance and speed of the algorithm during tracking.

Before introducing the mixed time-event triggering mechanism, we first describe the definition of time trigger mechanism and event trigger mechanism respectively. In the operation of dynamic systems, event triggering is updated at the moment of the next evaluation event, and time triggering is updated using a fixed length of time. The use of event triggering can provide timely and accurate response to the trust relationship of entities. Time triggering updates the system by a fixed length of time. If event triggering is initiated frequently, it will lead to a large amount of wasted computational power. While time triggering can fully adjust the time step for updating, this way has a certain delay, but the load of the system is less. To obtain the advantages of both triggering mechanisms, the mixed time-event triggering mechanism splits the template update problem and uses different update models for different tracking situations, which are in turn processed according to switching rules between these update models. Next, we describe in detail the mixed time-event triggering mechanism. For a tracking system without template updates, the construction of an input-output system is represented as:

$$B_i = f_1\left(B_1, im\right) \tag{2}$$

where B_i denotes the target template, which is not updated during the tracking process. B_1 denotes the predicted result. im denotes the search region.

For our tracking system, the input-output system is constructed as follows:

$$B_i, S_i = f_2\left(B_1, B_{t_k}, S_j, im\right) \tag{3}$$

where B_{t_k} denotes the dynamic template. S_j denotes the evaluation value of the dynamic template. S_i denotes the evaluation value of the predicted result.

Before describing the template update switching rule, the predicted results sequence is set as follows: $\{B_1, B_2, B_3, ...\}$.

Set the sequence of dynamic templates to be used for each frame as follows: $\left\{B_{t_1}, B_{t_2}, B_{t_3}, ...\right\}$.

The template update rule is to replace B_{t_i} with $B_{t_{i+1}}$, where t_{i+1} is defined as follows:

$$t_{i+1} = \text{Inf}\left\{k > t_i, S_{t_k} - S_{t_i} \geqslant \alpha \text{ or } t_k - t_i \geqslant \beta\right\} \tag{4}$$

where α denotes the event trigger factor. β denotes the time trigger factor. (In this work, $\alpha = 0.1, \beta = 5$.)

During the tracking process, the tracking system determines the dynamic template by the mixed time-event triggering mechanism after the tracking result is obtained for each frame. The mixed time-event triggering mechanism, by mixing dynamic tracking system with discrete event regulation, is able to increase the tracking speed with less increase in system load.

3.3.3 The adaptive template fusion mechanism

In the research of template update strategy, some tracking methods adopt the simplest averaging strategy to update the target appearance model, which gives the same weight to each target template. However, this update strategy treats each template equally, making the tracker unable to obtain the feature changes of the target timely. Another part of researchers introduces a linear update strategy to update the target appearance model, which is a function that decays exponentially with time to assign weights to different templates. Due to its acceptable performance, it has long been a widely adopted strategy for online updating. However, these template update strategies tend to make the tracker lose the first frame of template information, which is undoubtedly the most reliable. Therefore, we propose an adaptive template fusion mechanism. The formula is expressed as follows:

$$R_i^{new} = \left(1 - w\right) \times S_X \times R_i^X + w \times R_i^I \tag{5}$$

Where w is a predefined hyperparameter. R_i^I is the current frame response map generated from the first frame template. R_i^X is the current frame response map generated from the X-frame template. R_i^{new} is the final response map. S_X is the evaluation value of the X-frame template.

4 Experiment

This section presents the results of our algorithm on multiple benchmark datasets, with comparisons to the state-of-the-art tracking algorithms. Ablation studies are also provided to analyze the effects of the components in the proposed networks.

4.1 Implementation details

Training. The experiment is conducted on a PC with Intel (R) Core (TM) i7-9700K CPU 3.60GHZ and NVIDIA Quadro RTX 4000. Our novel bionic visual network and template update strategy are implemented based on the Pytorch framework for Python with OpenCV

4.5. The training image pairs for our algorithm are collected from the ImageNet VID dataset [28]. The size of an exemplar image is $127 * 127$ pixels, while the size of a search image is $255*255$ pixels. The network parameters are initialized with the normal distribution. We use stochastic gradient descent (SGD) to train the network from scratch, with momentum set to 0, weight decay set to 0, and initial learning rate set to 0.01. The learning rate is dynamically adjusted by StepLR, with the decay rate set to 0.1 and the step size set to 25. The model is trained for 30 epochs with a mini-batch size of 32. The loss curve during the training process is shown in Fig. 5.

4.2 Experiments on the tracking benchmark

4.2.1 Experiment on the GOT-10k dataset

GOT10K is a large-scale single-object tracking benchmark consisting of 10k video sequences, covering most of 563 object classes and 87 motion patterns, and is the most abundant motion trajectory dataset. This dataset contains more video sequences with deformation and partial occlusion than OTB. Results of GOT-10k testset need to be uploaded to the official website for analysis. The provided evaluation indicators include average overlap (AO) and success rate (SR). The AO represents the average overlaps between ground-truth boxes and estimated bounding boxes. The SR0.5 is the rate of successfully tracked frames that overlap more than 0.5, while SR0.75 is the rate of successfully tracked frames that overlap more than 0.75.

We evaluate our algorithm on GOT-10k testset and compare it with trackers such as SiamFC, SiamRPN, C-COT [9], KCF [19], ECO [5], Staple [2], MEEM [23], Mdnet, DTDU [29], MFESiam [11], DCANet [30] and ESiamFC [20]. As shown in Table 1, our tracker is ranked 1st in all indicators except HZ indicator. Among these algorithms, classical Siamese algorithms (SiamFC and SiamRPN) usually fail to track in the challenges of deformation, occlusion or fast motion, mainly due to the difficulty in adapting to the feature changes of the target. Our tracker handles these situations well due to the online update strategy. For the classic correlation filtering based trackers (KCF, Staple and ECO), the frequency domain computation method greatly improves the speed of the tracker. But compared with the deep features used in our algorithm, its generalization ability is insufficient, and it is prone to tracking failure. Tracking algorithms based on classification ideas (MEEM and MDnet) have high requirements for training samples, and it is difficult to maintain tracking performance in complex tracking environments. The latest siamese-based algorithms (MFESiam,

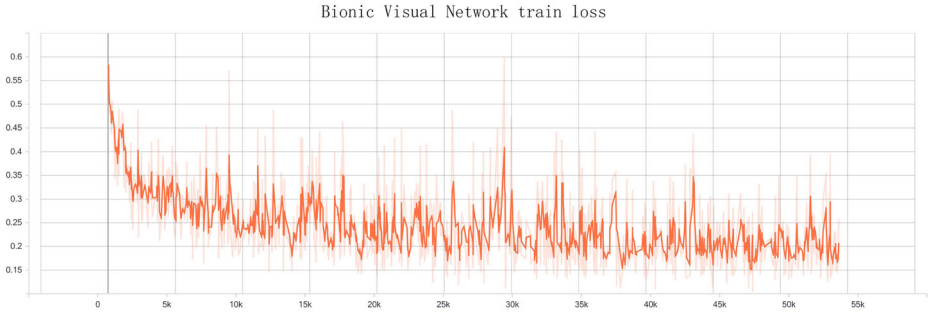

Fig. 5 Loss curve of the bionic visual network

Table 1 Details about the state-of-the-art trackers in GOT-10K

Tracker	AO	SE0.50	SR0.75	HZ
SiamRPN	0.367	0.425	0.103	<u>47.10</u>
Siamfc	0.348	0.353	0.098	44.15
C-COT	0.325	0.328	0.107	0.68
ECO	0.316	0.309	0.111	2.62
Mdnet	0.299	0.303	0.099	1.52
MEEM	0.253	0.235	0.068	20.59
Staple	0.246	0.239	0.089	28.87
KCF	0.203	0.177	0.065	**96.66**
DTDU	0.375	0.416	0.133	44.0
MFESiam	0.389	0.400	0.143	–
DCANet	<u>0.403</u>	<u>0.466</u>	<u>0.150</u>	28.06
ESiamFC	0.381	0.435	0.132	26.36
Our	**0.425**	**0.467**	**0.307**	30.67

The best and second best values are highlighted in bold and underlined

DTUD, DCANet and ESiamFC), the innovations of these algorithms include multi-feature fusion, online update, special attention structure, and new construction of a backbone network. They all have good performance in tracking performance, but our algorithm considers the problem of tracker performance degradation caused by the loss of translation invariance of the neural network when designing the online update strategy. And considering the practical application of the target tracking algorithm, we introduce the mixed time-event triggering mechanism into the online update strategy, which can effectively balance the performance and speed of the algorithm. Therefore, our tracker has good performance and speed.

4.2.2 Experiment on the UAV123 dataset

UAV123 contains 123 fully annotated HD video sequences over 110K frames. Acquired by low-altitude UAVs, this dataset provides an aerial point of view video sequences that are inherently different from traditional temporal tracking benchmarks such as OTB and VOT. This dataset contains video sequences such as viewpoint changes, drastic object appearance changes, complete occlusions and partial occlusions. Objects in the dataset have large scale differences in the early and late stages. Due to the movement of the camera and objects can be seen fast movement, occlusion and other difficulties, which makes tracking with this data set challenging.

As shown in Table 2, we evaluate the proposed method with several representative methods including SiamRPN, DaSiamRPN [55], ECO, SiamFC, DeepSRDCF [7], Staple, MEEM, A3CT [10], FF-Siam-CA [14], SiamFF-AV [15], SiamRPN++ [27]. In the algorithm comparison of this dataset, DeepSRDCF is a deep features based algorithm, which shows good performance in the early simple tracking environment. Other algorithms (SiamRPN++, DaSiamRPN, FF-Siam-CA, SiamFF-AV and A3CT.), most of these algorithms try to improve the quality of target modeling, the methods used include multi-layer feature fusion, deepen the network, depth reinforcement learning methods, etc. However, most tracking networks contain a large number of down-sampling operations, which results

Table 2 Details about the state-of-the-art trackers in UAV123

Tracker	ARE	AOR
SiamRPN	0.710	**0.577**
DaSiamRPN	0.724	<u>0.569</u>
ECO	0.688	0.525
Siamfc	0.648	0.485
DeepSRDCF	0.627	0.463
Staple	0.614	0.450
MEEM	0.570	0.412
A3CT	0.622	0.471
FF-Siam-CA	0.708	0.505
SiamFF-AV	0.700	0.497
SiamRPN++	<u>0.745</u>	0.551
Our	**0.746**	0.564

The best and second best values are highlighted in bold and underlined

in the inability of the tracker to improve target localization. Since our algorithm solves this problem, it performs well in localization accuracy.

4.2.3 Experiment on the OTB100 dataset

OTB100 contains one hundred videos with eleven attributes, focusing on testing and analyzing the tracker's ability to handle different scenes, such as illumination variation, deformation, occlusion, fast motion, etc. It is widely used to track literature. The proposed method is compared with different algorithms, the correlation filters based trackers, SRDCF [8], Staple, fDSST, the Siamese network based trackers, SiamDW [50], SiamRPN and SiamFC. Figure 6 shows the DP and AUC score of OTB100 datasets. The results show that our method greatly outperforms the correlation filter-based tracking method. Compared

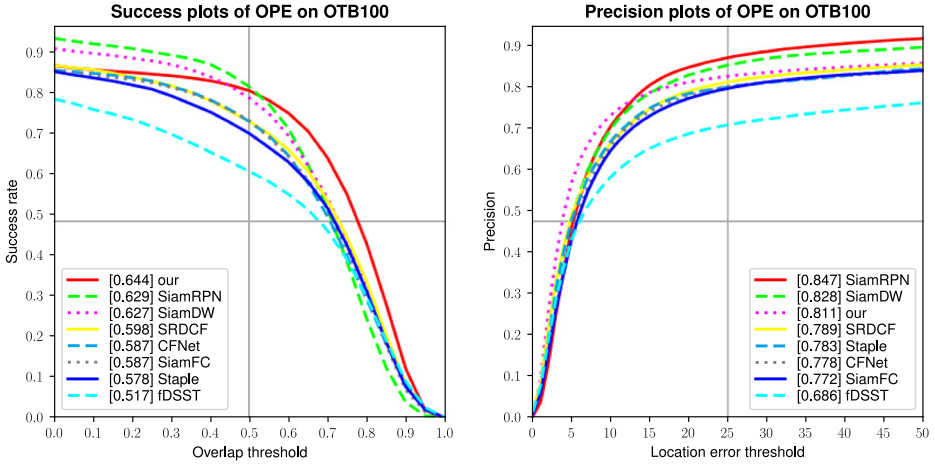

Fig. 6 Success and precision plots on the OTB-100 dataset

with the tracking algorithm based on the Siamese network, our algorithm shows a good performance. This indicates that the proposed method can effectively improve the target representation to achieve better robustness.

4.2.4 Experiment on the VOT2016 dataset

VOT2016 contains 60 challenging public videos. This challenge has been held every year since 2013, and many trackers are tested on it. The test criteria of VOT2016 is different from other datasets, the tracker is reset using ground-truth after 5 frames of tracking failure. Therefore, a robustness index is included in the evaluation criteria, which is calculated based on the number of tracking failures. For the assessment on VOT2016, we report the performance of some of the best non-siam-based trackers for reference, including Mdnet, ECO-HC [5], Staple, and EBT [53]. And compared with other siam-based trackers, such as SiamRPN, SiamAN [23], SiamRN [23]. The EAO curve evaluated on VOT2016 is presented in Table 3 and 7 other state-of-the-art trackers are compared. Table 3 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 Accuracy score of 0.610 and a Robustness score of 0.545. The first tracker, SiamRPN, is much better than our tracker in terms of EAO, while much lower in terms of Accuracy and Robustness, suggesting that the bionic vision network and template update mechanism introduced improves tracking performance.

4.2.5 Experiment on the VOT2018 dataset

VOT2018 contains 60 video sequences, VOT2018 and VOT2016 have 10 different video sequences, and the tracking dataset is also more challenging. We compare the proposed tracker with 7 state-of-the-art tracking algorithms on VOT2018 dataset. These trackers are: SiamFC, Staple, DAT [34], SiamLM [23], MEEM, KCF and Dsiam [16]. We evaluate the proposed method on VOT2018, and report the results in Table 4. As shown in Table 4, our method achieves the best accuracy score of 0.577 and the EAO score of 0.209. Notably, our method sets a new state-of-the-art by improving 0.022 absolute value, i.e., 2.2% relative improvement, compared to SiamFC, indicating that the bionic vision network and template update mechanism can significantly decrease the tracking failure.

Table 3 Details about the state-of-the-art trackers in VOT2016

Tracker	EAO	Accuracy	Robustness
SiamRPN	**0.344**	<u>0.56</u>	1.08
SiamRN	0.277	0.55	1.37
SiamAN	0.235	0.53	1.65
Mdnet	0.257	0.54	1.2
ECO-HC	<u>0.322</u>	0.53	1.08
Staple	0.295	0.54	1.35
EBT	0.291	0.47	<u>0.9</u>
Our	0.258	**0.610**	**0.545**

The best and second best values are highlighted in bold and underlined

Table 4 Details about the state-of-the-art trackers in VOT2018

Tracker	EAO	Accuracy	Robustness
SiamFC	0.187	0.503	0.585
Staple	0.169	<u>0.530</u>	0.688
DAT	0.144	0.435	0.721
SiamLM	**0.230**	0.500	**0.297**
MEEM	0.193	0.463	<u>0.534</u>
KCF	0.134	0.477	0.773
Dsiam	0.196	0.512	0.646
Our	<u>0.209</u>	**0.577**	0.693

The best and second best values are highlighted in bold and underlined

4.3 Qualitative evaluation

Figure 7 shows some results of the top-performing tracker: SiamDW, SiamRPN, SiamFC, GradNet and fDsst on 7 challenging sequences (from top to down: Bird2, Bolt2, Box, DragonBaby, Ironman, MotorRolling, Skater2, respectively). These video sequences contain deformation and occlusion challenges.

In the Bird2 and Bolt2 sequences, the scene contains the motion of many objects and the frequent deformation of the target. We can observe that when the target is deformed, these trackers drift and fDsst even fails to track. However, our tracker can maintain good tracking performance. This is due to the fact that the bionic visual network alleviates the loss of translation invariance and enlarges the receptive field. The target positioning capability of the tracker is guaranteed.

In the Box and Ironman sequences, the tracked objects are often occluded due to the complexity of the scene. SiamDW, SiamRPN, SiamFC, these siamese-based trackers rely on static template matching and their performance degrades severely in this scenario. Our algorithm enables the tracker to acquire the feature changes of the target in time due to the template update strategy with mixed time-event triggering mechanism.

In the DragonBaby, Skater2 and MotorRolling sequences, although GradNet can update templates online, due to the lack of a template evaluation mechanism, the tracking performance is degraded due to the used of low-quality templates during the tracking process. And our tracker includes a quality assessment model to ensure the effective operation of the template update strategy.

In summary, these tracking sequences contain occlusion and deformation challenges, which prove that the tracker proposed in this paper can handle these challenges well.

4.4 Ablation study

In this section, we perform an ablation analysis of the bionic visual network and the template update strategy with the mixed time-event triggering mechanism. To more intuitively illustrate the effectiveness of our proposed network and strategy, we experiment with an evaluation to analyze the algorithms in the UAV123 dataset. The average pixel error (ARE), average overlap rate (AOR), floating-point operations per second (FLOPs) and Params are shown in Table 5. Siambase represents the basic tracker SiamFC. BV represents the bionic vision network. Update represents the template update strategy with the mixed time-event triggering mechanism. As can be seen from Table 5, the introduction of the bionic visual network results in a 5.5% increase in the ARE of the algorithm performance compared to

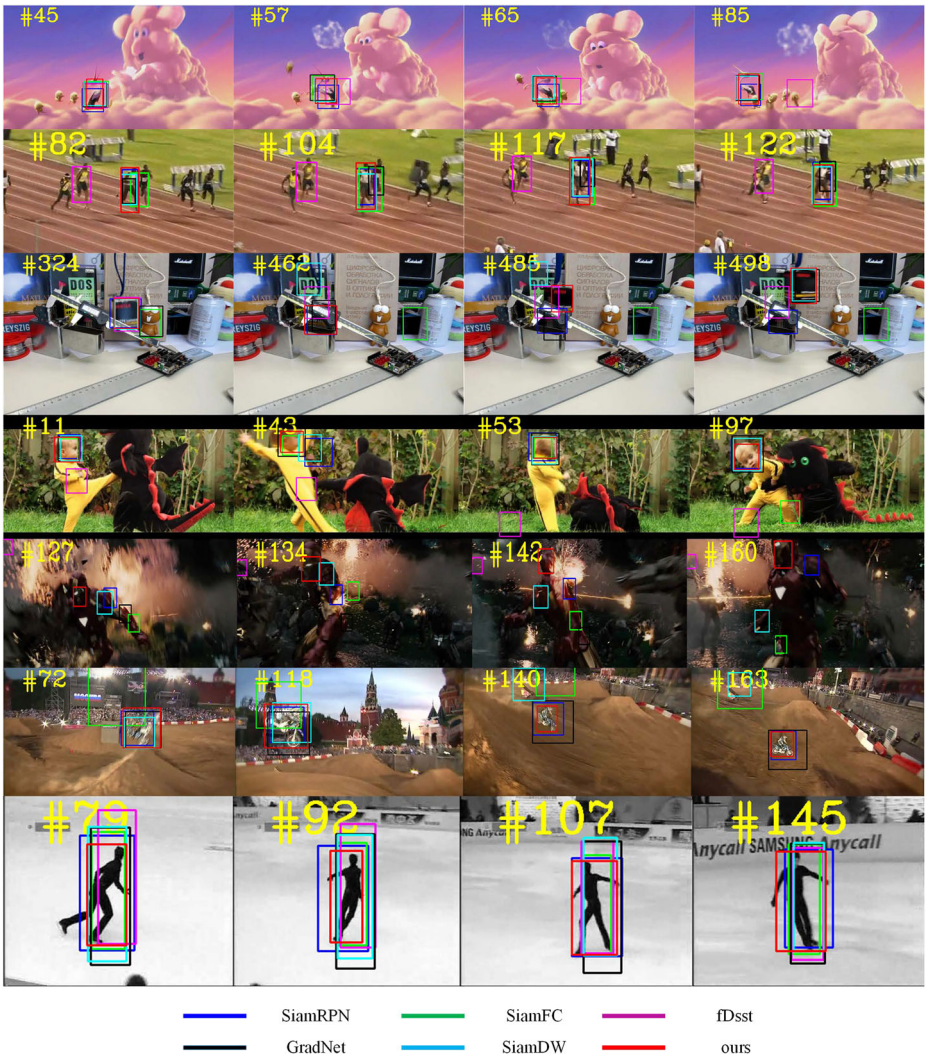

Fig. 7 Qualitative experiments

Siambase, while the AOR decreases by 0.6% compared to Siambase. The algorithm performance is further improved by introducing the template update strategy with the mixed time-event triggering mechanism, with a 4.3% increase in ARE and a 7.9% increase in AOR compared to Siambase. Further, we analyze the algorithm's precision plots and success plots on each challenging tracking attribute of OTB100. Compared with the basic tracker, our algorithm shows a large improvement in each challenge scenario. Details are shown in Fig. 8.

We compare the FLOPs and Params of the above three trackers. As shown in Table 5. The input tensor of the backbone network is $1 * 3 * 255 * 255$. According to data analysis, the introduction of BV not only reduces the FLOPs and Params of the original feature extraction network, but also improves the performance of tracker target localization. This proves that

Table 5 The algorithm complexity analysis of all trackers are compared on ARE, AOR, HZ, FLOPs (MB), Params (MB) on the UAV123 dataset

Tracker	ARE	AOR	FLOPs (MB)	Params (MB)
Siambase	0.648	0.485	1989.31	2.34
Siambase+BV	0.703	0.479	1698.43	2.31
Siambase+BV+Update	0.746	0.564	1822.72	6.74

the operations of alleviating the loss of translation invariance and expanding the receptive field are effective in improving network performance. Among them, we analyze that the reason for the reduction of FLOPs and Params is to set the output channel to 16 in the first layer of the convolution operation of the network, and enlarge it to 96 by the receptive

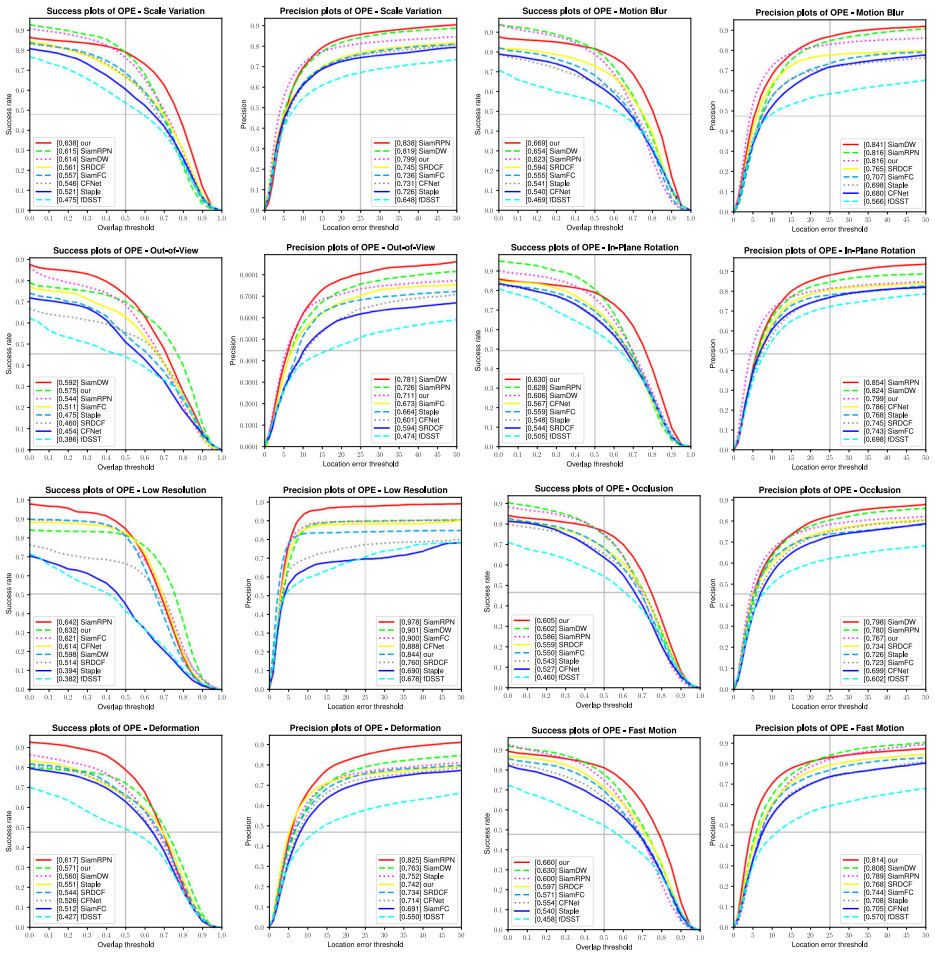

Fig. 8 Comparison of precision plots and success plots on challenging attributes for tracking, including scale variation, motion blur, out of view, in-plane rotation, low resolution, occlusion, fast motion, and deformation

Table 6 Influence of the parameters α, β tuning on tracking performance

	$\beta = 1$		$\beta = 5$		$\beta = 10$		$\beta = 15$	
	AO	HZ	AO	HZ	AO	HZ	AO	HZ
$\alpha = 0$	0.437	15.19	0.429	26.67	0.426	24.7	0.409	23.46
$\alpha = 0.1$	0.437	15.19	0.425	30.67	0.419	28.55	0.409	25.00
$\alpha = 0.2$	0.437	15.19	0.423	32.15	0.420	31.04	0.412	29.89
$\alpha = 0.3$	0.437	15.19	0.420	35.84	0.421	35.38	0.418	35.38

field block. The dilated convolution operation in the receptive field block greatly reduces the rise in network complexity. After the introduction of the template update strategy, the FLOPs of the tracking algorithm increased slightly, and the Params increased significantly. Among them, the rise of Params is mainly due to the three-layer fully connected layer of the quality assessment model. But from the performance point of view, the template update strategy helps the tracker to get a good improvement in both evaluation metrics (ARE and AOR).

In addition, we conduct experiments on the parameters α, β of the mixed time-event triggering mechanism and the parameter w of the adaptive template fusion mechanism. These experiments are conducted out on the GOT10k dataset, based on two metrics, AO and HZ. The AO represents the average overlaps between ground-truth boxes and estimated bounding boxes. The HZ represents the tracking speed of the tracker. The parameter influence is shown in Table 6. According to the data in Table 6, with the increase of β related to the time triggering mechanism, the AO gradually decreases. This is because the event-triggering mechanism is not always activated. The longer the time interval, the less frequently the update template mechanism is triggered, resulting in degraded tracker performance. As the α associated with the event-triggered mechanism increases, the total number of times the event-triggered mechanism is activated decreases, reducing the number of template updates. So the speed metric of the tracker has improved. With the increase of w, the proportion of the response map generated by the first frame template in the final response map increases, and the AO increases and then decreases. This is due to the fact that the static template cannot reflect the target feature changes, which leads to the performance degradation of the tracker. The parameter w has little effect on HZ, and in a limited number of trials, the speed of the tracker is keeping at around 30 frames (Fig. 9).

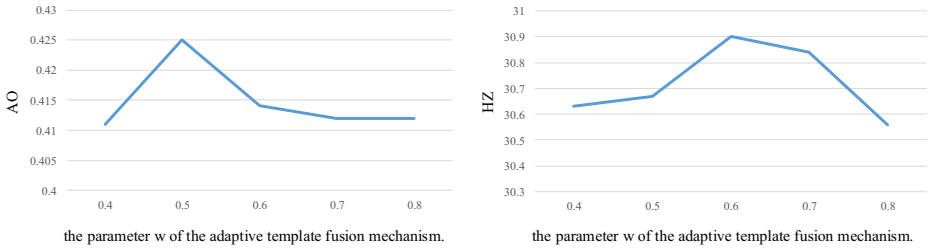

Fig. 9 Influence of the parameters w tuning on tracking performance

5 Conclusions

In this paper, we first introduce the definition of target tracking and the development of tracking algorithms. Then, we analyze the reasons why the siamese-based trackers lead to severe performance degradation in target deformation and occlusion scenarios: most siamese-based trackers extract target features only in the first frame, and generate response maps in subsequent frames only by the static features of the target. After that, we find that in many online update based tracking algorithms, it is desirable to obtain high-quality templates to prevent the degradation of tracker performance. Therefore, these algorithms introduce a large number of attention modules into the network structure to fuse the features of each layer or design a special update mechanism, however, the impact of the translation invariance of the network on the online update-based tracking algorithm is ignored. The quality of features extracted by convolutional neural networks for the same target at different locations fluctuates widely.

This paper is based on the existing research on network characteristics, an online bionic visual siamese tracking framework based on the mixed time-event triggering mechanism is designed. The bionic vision network that combines the dilated convolutions to expand the perceptual field and blurpool to mitigate the loss of translation invariance to improve the stability and quality of extracted features. The template update strategy with a mixed time-event triggering mechanism is proposed to improve the accuracy of the tracker in target deformation scenarios. In the experimental section, five mainstream single-target tracking datasets (GOT-10k, UAV123, OTB100, VOT2016, VOT2018) are used to evaluate the performance of the tracker proposed in this paper, and these tracking datasets contain multiple challenging scenarios such as deformation and occlusion. We also analyze in detail the reasons why the tracker achieves the advantage in specific scenarios through qualitative experiments. Extensive quantitative and qualitative experiments show that alleviating the lost of translation invariance of the network can improve the performance of the tracker in occlusion and deformation scenarios. Therefore, the online bionic visual Siamese tracker based on mixed time-event triggering mechanism has better competitive performance and tracking speed in the tracking scene including deformation and occlusion.

Although the proposed tracking approach achieves better performance than several state-of-the-art trackers, it cannot handle the problem of similar object interference well. In the future, we will investigate the redetection component to recover the tracked target in case of tracking failure caused by similar object interference.

Acknowledgements This work is supported by the National Natural Science Foundation of China under Grant (61873246, 62072416, 62006213, 62102373), Program for Science & Technology Innovation Talents in Universities of Henan Province (21HASTIT028), Natural Science Foundation of Henan (202300410495), Key Scientific Research Projects of Colleges and Universities in Henan Province (21A120010).

Data Availability The raw/processed data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

Declarations

Conflict of Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Azulay A, Weiss Y (2018) Why do deep convolutional networks generalize so poorly to small image transformations?. arXiv:1805.12177
2. Bertinetto L, Valmadre J, Golodetz S, Miksik O, Torr HS (2016) Staple: Complementary learners for real-time tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1401–1409
3. Chen Z, Zhong B, Li G, Zhang S, Ji R (2020) Siamese box adaptive network for visual tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6668–6677
4. Collins RT, Liu Y, Leordeanu M (2005) Online selection of discriminative tracking features. IEEE Trans Pattern Anal Machine Intell 27(10):1631–1643
5. Danelljan M, Bhat G, Shahbaz Khan F, Felsberg M (2017) Eco: Efficient convolution operators for tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6638–6646
6. Danelljan M, Häger G, Khan FS, Felsberg M (2016) Discriminative scale space tracking. IEEE Trans Pattern Anal Machine Intell 39(8):1561–1575
7. Danelljan M, Hager G, Shahbaz Khan F, Felsberg M (2015) Convolutional features for correlation filter based visual tracking. In: Proceedings of the IEEE international conference on computer vision workshops, pp 58–66
8. Danelljan M, Hager G, Shahbaz Khan F, Felsberg M (2015) Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the IEEE international conference on computer vision, pp 4310–4318
9. Danelljan M, Robinson A, Shahbaz Khan F, Felsberg M (2016) Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: European conference on computer vision. Springer, pp 472–488
10. Dunnhofer M, Martinel N, Luca Foresti G, Micheloni C (2019) Visual tracking by means of deep reinforcement learning and an expert demonstrator. In: Proceedings of The IEEE/CVF international conference on computer vision workshops, pp 0–0
11. FEI D, SONG H, ZHANG K (2020) Multi-level feature enhancement for real-time visual tracking. J Comput Appl 40(11):3300
12. Fu L, Ding Y, Du Y, Zhang B, Wang L, Wang D (2020) Siammn: Siamese modulation network for visual object tracking. Multimed Tools Appl 79(43):32623–32641
13. Gündoğdu E, Alatan AA (2016) The visual object tracking vot2016 challenge results
14. Guo D, Wang J, Zhao W, Cui Y, Wang Z, Chen S (2021) End-to-end feature fusion siamese network for adaptive visual tracking. IET Image Proc 15(1):91–100
15. Guo D, Zhao W, Cui Y, Wang Z, Chen S, Zhang J (2018) Siamese network based features fusion for adaptive visual tracking. In: Pacific Rim international conference on artificial intelligence. Springer, pp 759–771
16. Guo Q, Feng W, Zhou C, Huang R, Wan L, Wang S (2017) Learning dynamic siamese network for visual object tracking. In: Proceedings of the IEEE international conference on computer vision, pp 1763–1771
17. He A, Luo C, Tian X, Zeng W (2018) A twofold siamese network for real-time object tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4834–4843
18. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
19. Henriques JF, Caseiro R, Martins P, Batista J (2014) High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Machine Intell 37(3):583–596
20. Huang H, Liu G, Zhang Y, Xiong R, Zhang S (2022) Ensemble siamese networks for object tracking. Neural Comput Appl 34(10):8173–8191
21. Huang L, Zhao X, Huang K (2019) Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Trans Pattern Anal Mach Intell 43(5):1562–1577
22. Jepson AD, Fleet DJ, El-Maraghi TF (2003) Robust online appearance models for visual tracking. IEEE Trans Pattern Anal Machine Intell 25(10):1296–1311
23. Kristan M, Leonardis A, Matas J, Felsberg M, Pflugfelder R, Čehovin Zajc L, Vojir T, Bhat G, Lukezic A, Eldesokey A et al (2018) The sixth visual object tracking vot2018 challenge results. In: Proceedings of the European conference on computer vision (ECCV) workshops, pp 0–0
24. Kristan M, Leonardis A, Matas J, Felsberg M, Pflugfelder R, Čehovin Zajc L, Vojir T, Hager G, Lukezic A, Eldesokey A et al (2017) The visual object tracking vot2017 challenge results. In: Proceedings of the IEEE international conference on computer vision workshops, pp 1949–1972
25. Kristan M, Matas J, Leonardis A, Felsberg M, Cehovin L, Fernández G, Vojir T (2016) Hager, and et al. the visual object tracking vot2016 challenge results. In: ECCV workshop, vol 2, p 8
26. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25

27. Li B, Wu W, Wang Q, Zhang F, Xing J, Yan J (2019) Siamrpn++: Evolution of siamese visual tracking with very deep networks. CVPR 4282–4291
28. Li B, Yan J, Wu W, Zhu Z, Hu X (2018) High performance visual tracking with siamese region proposal network. In: 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 8971–8980
29. Liu J, Wang Y, Huang X, Su Y (2022) Tracking by dynamic template: Dual update mechanism. J Vis Commun Image Represent 84:103456
30. Ma X, Guo J, Tang S, Qiao Z, Chen Q, Yang Q, Fu S (2020) Dcanet: Learning connected attentions for convolutional neural networks. arXiv:2007.05099
31. Mueller M, Smith N, Ghanem B (2016) A benchmark and simulator for uav tracking. In: European conference on computer vision. Springer, pp 445–461
32. Nam H, Han B (2016) Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4293–4302
33. Noor S, Waqas M, Saleem MI, Minhas HN (2021) Automatic object tracking and segmentation using unsupervised siammask. IEEE Access 9:106550–106559
34. Pu S, Song Y, Ma C, Zhang H, Yang M-H (2018) Deep attentive tracking via reciprocative learning. Advances in neural information processing systems 31
35. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28
36. Ross DA, Lim J, Lin R-S, Yang M-H (2008) Incremental learning for robust visual tracking. Int J Comput Vision 77(1):125–141
37. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition
38. Tao R, Gavves E, Smeulders WM (2016) Siamese instance search for tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1420–1429
39. Valmadre J, Bertinetto L, Henriques J, Vedaldi A, Torr HS (2017) End-to-end representation learning for correlation filter based tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2805–2813
40. Wang Q, Teng Z, Xing J, Gao J, Hu W, Maybank S (2018) Learning attentions: residual attentional siamese network for high performance online visual tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4854–4863
41. Wei L, Xi Z, Hu Z, Sun H (2022) Siamsyb: simple yet better methods to enhance siamese tracking. Multimedia Tools Appl 1–20
42. Wu Y, Lim J, Yang M-H (2013) Online object tracking: A benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2411–2418
43. Xiao D, Tan K, Wei Z, Zhang G (2022) Siamese block attention network for online update object tracking. Appl Intell 1–13
44. Xu Y, Wang Z, Li Z, Yuan Y, Yu G (2020) Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. AAAI 12549–12556
45. Xu Z, Luo H, Hui B, Chang Z, Ju M (2019) Siamese tracking with adaptive template-updating strategy. Appl Sci 9(18):3725
46. Yang T, Chan AB (2018) Learning dynamic memory networks for object tracking. In: Proceedings of the European conference on computer vision (ECCV), pp 152–167
47. Yuan T, Yang W, Li Q, Wang Y (2021) An anchor-free siamese network with multi-template update for object tracking. Electronics 10(9):1067
48. Zhang L, Gonzalez-Garcia A, Weijer JVD, Danelljan M, Khan FS (2019) Learning the model update for siamese trackers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 4010–4019
49. Zhang R (2019) Making convolutional networks shift-invariant again. In: International conference on machine learning. PMLR, pp 7324–7334
50. Zhang Z, Peng H, Wang Q (2019) Deeper and wider siamese networks for real-time visual tracking. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR 2019) 4586–4595
51. Zhao F, Zhang T, Song Y, Tang M, Wang X, Wang J (2020) Siamese regression tracking with reinforced template updating. IEEE Trans Image Process 30:628–640
52. Zhou Y, Li J, Du B, Chang J, Ding Z, Qin T (2021) Learning adaptive updating siamese network for visual tracking. Multimedia Tools Appl 80(19):29849–29873
53. Zhu G, Porikli F, Li H (2016) Beyond local search: Tracking objects everywhere with instance-specific proposals. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 943–951
54. Zhu W, Zou G, Liu Q, Zeng Z (2021) An enhanced visual attention siamese network that updates template features online. Secur Commun Netw 2021
55. Zhu Z, Wang Q, Li B, Wu W, Yan J, Hu W (2018) Distractor-aware siamese networks for visual object tracking. In: Proceedings of the European conference on computer vision (ECCV), pp 101–117

Affiliations

Huanlong Zhang[1] · Zhuo Zhang[1] · Jiapeng Zhang[1] · Yanchun Zhao[2] · Miao Gao[3]

Zhuo Zhang
youngzhangzhuo@163.com

Jiapeng Zhang
zjpzzuli@163.com

Yanchun Zhao
1354823440@qq.com

Miao Gao
1248139671@qq.com

[1] College of Electrical and Information Engineering, Zhengzhou University of Light Industry, Dongfeng Road, Zhengzhou, 450002, Henan Province, People's Republic of China

[2] Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology of China, Yatai Road, Huzhou, 313001, Zhejiang Province, People's Republic of China

[3] China Tobacco Henan Industrial CO.,LTD, Henan, People's Republic of China