IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 July 2024, accepted 26 August 2024, date of publication 28 August 2024, date of current version 13 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3451229

RESEARCH ARTICLE

The Segmentation Tracker With Mask-Guided Background Suppression Strategy

ERLIN TIAN[1], YUNPENG LEI[2], JUNFENG SUN[3], KEYAN ZHOU[2], BIN ZHOU[2], AND HANFEI LI[1]

[1]School of Electronic Information, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[2]School of Electric and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[3]China Tobacco Guangxi Industrial Company Ltd., Nanning 530000, China

Corresponding author: Junfeng Sun (Junfengsun1981@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62272423, Grant 62072416, Grant 62006213, and Grant 62102373; in part by the Program for Science and Technology Innovation Talents in Universities of Henan Province under Grant 21HASTIT028; in part by the Natural Science Foundation of Henan under Grant 202300410495; and in part by the Key Scientific Research Projects of Colleges and Universities in Henan Province under Grant 21A120010.

ABSTRACT Segmentation-based tracking is currently a promising tracking paradigm with pixel-wise information. However, the lack of structural constraints makes it difficult to maintain excellent performance in the presence of background interference. Therefore, we propose a Segmentation tracker with mask-guided background suppression strategy. Firstly, a mask-aware module is designed to generate more accurate target masks. With the guidance of regression loss, features were selected that are sensitive only to the target region among shallow features that contain more spatial information. Structural information is introduced and background clutter in the backbone feature is suppressed, which enhances the reliability of the target segmentation. Secondly, a mask-guided template suppression module is constructed to improve feature representation. The generated mask with clear target contours can be used to filter the background noise, which increases the distinction between foreground and background of which. Therefore, the module highlights the target area and improves the interference resistance of the template. Finally, an adaptive spatiotemporal context constraint strategy is proposed to aid the target location. The strategy learns a region probability matrix by the object mask of the previous frame, which is used to constrain the contextual information in the search region of the current frame. Benefiting from this strategy, our method effectively suppresses similar distractors in the search region and achieves robust tracking. Broad experiments on five challenge benchmarks including VOT2016, VOT2018, VOT2019, OTB100, and TC128 indicate that the proposed tracker performs stably under complex tracking backgrounds.

INDEX TERMS Object tracking, Siamese network, object segmentation, background interference.

I. INTRODUCTION

Single object tracking plays a crucial role in computer vision, and it has various practical applications in various fields such as medical diagnosis, video surveillance and human-computer interaction. Only the ground truth of the target to be tracked in the first frame is given, VOT aims to locate the target and report its location and size with a bounding box in the subsequent frames. Presently, VOT has still been considered a highly difficult task because of a number of factors for example background clutter and deformation, especially for Siamese-based trackers.

The shape of the target is often irregular, however, tracking is considered a matching process in Siamese-based trackers, which take fixed-size rectangular features as target templates. As a result, interference information is also included in the target template, which may reduce the reliability of the target representation. In addition, background interference in the search area tends to cause tracking drift.

To suppress the background distractor, many efforts have been made on Siamese-based trackers recently. Spatio-temporal constraints, attentional mechanisms, and

The associate editor coordinating the review of this manuscript and approving it for publication was Young Jin Chun.

exploration of distinguishing features are common approaches. Some Siamese trackers employ cosine windows to constrain the response map, which achieves the goal of focusing more on the centers and ignoring the surrounding areas. For example, the cosine window is used in SiamRPN [1] and the scores of proposals are ranked for the best proposal. And there are some trackers [2], [3], [4], [5] that penalize large displacements through temporal context prior. This distribution is too simple as a temporal context prior and therefore has limited capability for constraining background interference. And some incorporated attention mechanisms into the tracker to make the network focus more on the target of interest and thus ignore irrelevant information. There are various attention mechanisms applied in the Residual attentional siamese network(RASNet) [6] to enhance its discrimination, which includes channel, residual, and general attention. SA-Siam [7] integrates a channel attention module into the semantic branch to calculate the weight of the channel based on the channel activities that are surrounding the target location. Distraction-aware module is designed in the Distractor-aware siamese network [8] in order to help the tracker pay more attention to semantic interference. Some trackers combat background interference by exploring distinguishing features. C-RPN [9] adopts the fusion of multi-layer features and multiple steps of regressions to progressively refine the representation of the target. For SiamDW [10], the depth and width of the backbone framework are enhanced to achieve more accurate and robust tracking results.

Despite the effectiveness of these methods, they are limited by the form of target representation, which makes it difficult to further improve the performance of the tracker. Because existing trackers commonly represent targets by an axis-aligned rectangular box, which is unreliable and inevitably introduces background information in many cases. For example, when the target is a person with open hands, although the rectangular box surrounds the target, it contains a lot of background information. This situation is illustrated in Figure 1. The output of the semi-supervised segmentation task consists of a binary segmentation mask indicating whether a pixel belongs to the target or not. Consider that it is advantageous to solve the background interference problem because the target and the background can be clearly distinguished.

Recently, segmentation has often been introduced into trackers to improve their performance thanks to its pixel-level information. SiamMask [11] introduces a segmentation branch following the SiameseRPN, which allows the regression of bounding boxes and the segmentation of objects to be learned jointly in training. Then much work is done based on SiamMask on reverse optimization paradigm [12] and rotated box estimation [13]. Later, D3S [14] combines a segmentation branch with online DCF [15] to incorporate the process of target classification and inference of pixel-level segmentation. However, segmentation plays only an auxiliary role in these trackers, and few studies have used segmentation information to solve the background interference problem in trackers. In addition, tracking accuracy will be compromised due to the direct reuse of features in the tracking network. This is because, without structural constraints, interference can be introduced in the learning process of segmentation due to background clutter outside the bounding box.

To solve the problems mentioned above, we propose a Segmentation tracker with mask-guided background suppression strategy that aims to mitigate the background interference problem in Siamese-based trackers by exploiting the segmentation information at the pixel level. First, a mask-aware module is proposed to guarantee the accuracy of segmentation. Regression Loss is introduced to guide the selection of features at the shallow layers, which are only sensitive to the target region. The shallow features extracted by the neural network have more spatial information. This module reduces the interference in the backbone network and introduces structural information to ensure the reliability and accuracy of segmentation mask generation. With the pixel-level information, the segmentation mask can clearly separate the target from the background and can thus be used to solve the problem of background interference. Then, a background suppression model is constructed to reduce background interference in the target template by using segmentation information. Finally, a mask-guided spatiotemporal contextual constraint model is constructed to aid target localization. A region probability matrix is calculated via the previous frame's mask, which represents the spatial distribution between the object of interest and its contextual information. This matrix is then applied to the search region features to constrain the contextual information.

Our method based on the Siamese tracking framework shows better performance in suppressing the background due to the exploitation of segmentation information. Extensive experiments on popular tracking benchmarks including VOT2016, VOT2018, VOT2019, OTB100, and TC128 have validated the effectiveness of our tracker.

The main contributions of this work are summarized in three-fold:

- We propose the mask-aware module to ensure the accuracy of the generated segmentation masks. Regression loss is introduced to guide the selection of shallow features that are sensitive only to the target region to generate the mask. This introduces structural information constraints in the process of generating the mask, thus reducing the interference of background clutter in the backbone features.
- We propose a mask-guided template suppression module and adaptive spatiotemporal context constraint strategy that aims to suppress background interference by exploiting the pixel-level segmentation information. For the template branch, MTSM employs the mask of the initial frame to filter out background interference in the target template, which improves its discriminative capacity. For the search region branch, ASTCCS calculates a region probability matrix based on the mask of the

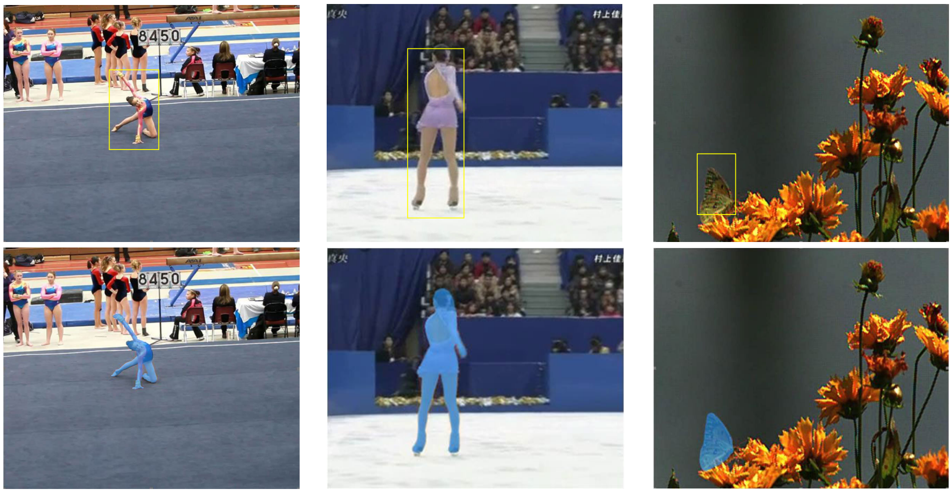


FIGURE 1. An illustration of the annotation of a target with a rectangular box is shown in the first line. It can be seen that the irregularity of the target shape thus introduces an amount of background information in the bounding box. And the segmentation results are shown in the second line. It can be seen that the pixel-level segmentation result has a clear target contour and can clearly distinguish the target from the background, which shows the natural advantage of using the segmentation result to suppress background interference.

previous frame to constrain the contextual information in the current frame to aid the target location.

- The proposed module is seamlessly integrated into the siamese-based segmentation tracking framework, and the results of extensive experiments on popular datasets show that our approach achieves excellent performance.

II. RELATED WORKS

The main content of this section is to introduce the technologies related to this work. Our approach introduces segmentation information into the Siamese-based tracking framework for removing its background interference. We first present the Siamese-based tracking, then the segmentation based tracking and the tracking with background information included at the end.

A. SIAMESE-BASED TRACKERS

The Siamese-based trackers have aroused a lot of attention with balanced speed and precision achieved. Siamese networks were originally proposed for signature verification tasks and were then applied for visual tracking in SINT [16] and SiamFC. SINT turned out the efficiency of the learned matching function. It matches the initial patch of the first frame with the candidates of the new frame by means of a learned matching function and returns the most similar patch. SiamFC trained a Siamese-based fully-convolutional network to perform cross-correlation between a template image within a larger search region. Subsequent Siamese-based trackers followed this thinking and treat tracking as a similarity-matching method. The features of the target template and the search area branches are extracted with branches of shared weights, and then their similarity is calculated by correlation operations. There are a number of strategies being proposed to enhance SiamFC, for instance bounding box regression, exploration of deeper networks [10], [17], [18], loss function [19] and updating of model [20], [21]. CFNet [15] integrates the correlation filters into SiamFC as a CNN layer. DSiam [3] equips the SiamFC with a fast transformation learning model to online adaptability of both target appearance and background suppression. Region proposal network (RPN) is embedded in the Siamese-based network in Siamrpn [1], which employs predefined anchors instead of the multi-scale searching method as used in previous trackers to solve the scale estimation problem. Siamrpn++ [17] integrates a spatial aware sensing sampling strategy to the ResNet [22]-driven Siamese tracker, which mitigates the restriction of network depth for tracker performance. Subsequent works [18], [23] propose an anchor-free algorithm to reduce the calculation of parameter adjustment and improve the adaptive capability of the network to achieve the effect of manual intervention reduction.

Siamese-based trackers have made great progress in terms of accuracy and speed, however, background interference is still a significant challenge for them. Therefore, in this paper, we address this problem by introducing pixel-level segmentation information with clear target edges. Segmentation information is introduced to filter out background interference in the template and to constrain the contextual information in the search region.

B. VIDEO OBJECT SEGMENTATION

Semi-supervised VOS [3], [24], [25], [26] task aims to predict a detailed mask representation of the interest object given in the first frame. Therefore with the pixel-wise segmentation information, many studies have developed various segmentation trackers by optical flow [27], filter [28], [29], boosting decision tree [30], and Absorbing Markov Chain [31].

Depending on the network structure, segmentation-based trackers can be categorized roughly into two groups, cascaded and parallel. The cascade paradigm [32], OceanPlus [12] SiamMargin [12] adapts a well-trained bounding-box-to-segmentation (Box2Seg) network [33], [34] to output binary target masks within the bounding boxes, which are presented by a separate branch. For SiamRCNN [32], a Box2Seg network from PReMVOS [34] was employed to predict the mask, which is a fully convolutional DeepLabV3+ [33] network with an Xception-65 [35] backbone and has been trained on Mapillary [36] and COCO [37]. The parallel paradigm [11], [14] typically contains a segmentation branch composed of up-sampling modules for refinement of the features extracted from the backbone network of the tracker, similar to the classical U-Net construction [38]. For the Siamese-based tracker, SiamMask [11] introduces a segmentation branch following the SiameseRPN to implement combined training of regressing the bounding box and segment target. For correlation filter-based tracking, D3S [14] combines a segmentation branch with online DCF [15] to get the combination of target classification and pixel-wise segmentation reasoning. Our approach utilizes a parallel architecture and proposes a mask-aware module to guarantee the quality of mask generation. Afterward, the mask is used in the tracking process to reduce the impact of background interference on the tracker.

C. TRACKERS WITH BACKGROUND INFORMATION

Although some progress has been gained in visual tracking recently, for example, Siamese and segmentation-based solutions, background information is also a challenging task for visual tracking. The capacity of trackers can be remarkably enhanced if it is equipped with a background suppression component. In correlation filter-based trackers, temporal context prior is employed in diverse forms to suppress background distractors. Some approaches [21], [39], [40], [41] use cosine windows or Gaussian windows as temporal context priors to structure constraint models. Zhang et al. [42] formulate spatiotemporal relationships between the target and its surrounding regions based on a Bayesian framework. Based on the Siamese-based tracker, Galoogahi et al. [43] learn a novel background suppression module by mining the information of the background features in the first frame. Tan et al. [44] detects and suppress background interference through the calculation of activation maps of them in the previous frame. Wei [45] et al. improve the feature representation by filtering out background interference in the template, which mitigates background interference. In contrast to the above tracker, segmentation information is considered a priori information in our tracker to suppress the background noise both in the template branch and search region branch. For the template branch, the initial frame is employed to filter out background interference in the target template and we also compute a region probability matrix to constrain the contextual information in the search region.

III. METHOD

In the following, the overall framework of our proposed method is first presented in Section III-A, and the mask-aware module is explained in Section III-B. The benefits of our mask-guided template suppression module design are demonstrated in Section III-C. Finally, we describe the adaptive spatiotemporal context constraint strategy in detail in Section III-D.

A. OVERALL ARCHITECTURE

The proposed framework aims to address the problem of background interference in Siamese-based networks by segmentation information at the pixel level. The overall pipeline is shown in Figure 2.

The proposed method first follows the recent Siamese-based tracker SiamMask [11] to implement object tracking. Centered on the target annotated in the initial frame, the template image is cropped to a size of 127*127 as in the classical Siamese structure. In the following tracking process, the search region image is obtained by cropping at the center of the target position in the last frame, which has a size of 255*255. Then, we follow the original SiamMask inference procedure to predict the target location in the search image. The target template image and the search region image are extracted features through a fully-convolutional network with the same weights and obtain template feature z and search region feature X. The two features are performed with a depth-wise cross-correlation to obtain a multi-channel response map, and the process can be described as:

$$f(z, X) = \varphi(z) * \varphi(X) \tag{1}$$

where $\varphi()$ is the function to obtain deep features from the pre-trained neural network resnet50. $*$ denotes depth-wise cross-correlation.

In addition to similarity scores and bounding box candidates, the response map also contains the necessary information that allows the generation of binary masks. For target mask generation, we first extract the information of the highest scoring position in the response map. The strategy of [46] is then followed to merge low-resolution features and high-resolution features with multiple refinement modules composed of up-sampling layers and skip connections.

Direct reuse of features in the tracking network can compromise segmentation accuracy. To improve the reliability and accuracy of generating segmentation masks, the mask-aware module is proposed, in which structural constraints are introduced in the generation process of masks. Based on SiamMask, tracking-oriented backbone features are fed into the mask-aware module, and shallow features that are sensitive to the target region only are selected to generate the mask under the guidance of regression loss. The target mask provides pixel-level classification information of the target and background which can be used to suppress background interference. To enhance the discrimination of the target template, the initial frame target mask is fed into the mask-guided template background suppression module

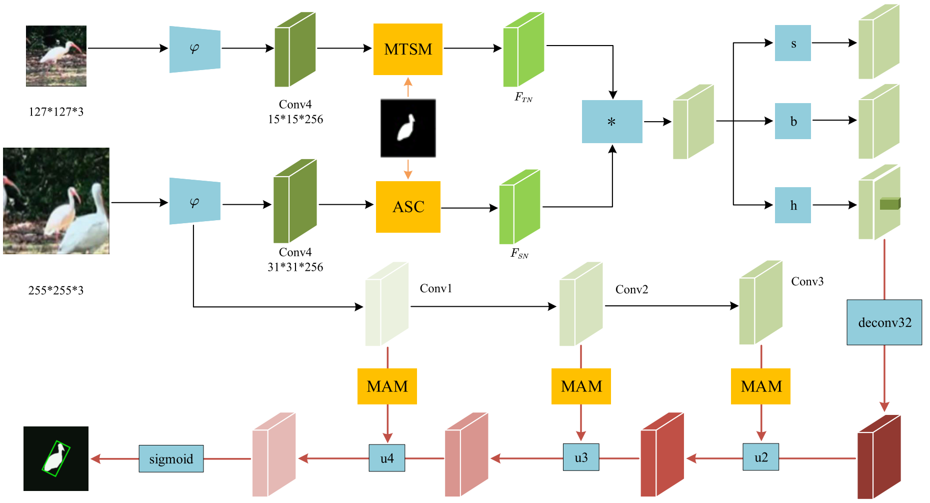


FIGURE 2. The overall framework of our method. The aim of our method is to reduce background interference in Siamese-based trackers by exploiting pixel-level segmentation information. Firstly, the mask-aware module (MAM) is proposed to lead the network to pay more attention to the features of the target area and thus obtain more accurate object masks. Then, the mask-aware module (MAM) and the adaptive spatiotemporal context constraint strategy (ASC) are proposed to introduce the target mask into the template branch and the search area branch to suppress background interference respectively.

to filter out the noise of the background in the template. To help target location, the target mask of the previous frame is used to construct a region probability matrix to constrain the contextual information in the search region of the current frame. the improved Siamese-based tracking method can be described as:

$$f_n(z, X) = F_{TN} * [R \odot \varphi(X)] \tag{2}$$

where F_{TN} denotes the target template after background suppression. R denotes the region probability matrix, and $\odot$ denotes the Hadamard product.

B. MASK-AWARE MODEL (MAM)

The segmentation branch is often designed in a segmentation-based tracker as a structure like the classical U-Net, which includes up-sampling layers for the refinement and integration of the low- and high-resolution features extracted from the backbone network of the tracker. In this paper, the mask with a clear target edge is used to suppress background interference in the proposed method, so the accuracy of the mask representation is important. However, the general features extracted consist of several feature patterns representing different target priors, of which the object of interest represents only a part. Other redundant features may compromise segmentation accuracy and introduce false positives in segmentation. Therefore, we constructed a mask-aware features model for the segmentation branch to select features that are active for the target region, as shown in Figure 3.

Specifically, we compute the importance of features along the channel depending on a gradient-based approach and then retain the most important features as mask-aware features. Given the extracted general features X, mask-aware features that are active for the target can be generated according to the channel importance Δ as:

$$\chi' = y(\chi; \Delta) \tag{3}$$

where y denotes the mapping function that selects channel features with top importance. Equation (3) is inferred from a combination of the chain rule and Equation (2) where x is the output prediction.

We regress all the samples $X_{i,j}$ in a template to a Gaussian label map $Y(i,j) = e^{-\frac{i^2+j^2}{2\sigma^2}}$, where (i,j) is the offset against the target and σ is the kernel width. The problem can be formulated as the regression loss,

$$L_{reg} = \|Y(i,j) - W * X_{i,j}\|^2 - \lambda \|W\|^2 \tag{4}$$

where W denotes the regressor weight and $*$ indicates the convolution operation. The importance of each channel feature can be calculated based on its contribution to the fitted label map, i.e., the derivation of L_{reg} with respect to the input feature X_{in}.

$$\begin{aligned} \frac{\partial L_{reg}}{\partial X_{in}} &= \sum_{i,j} \frac{\partial L_{reg}}{\partial X_o(i,j)} \times \frac{\partial X_o(i,j)}{\partial X_{in}(i,j)} \\ &= \sum_{i,j} 2\left(Y(i,j) - X_O(i,j)\right) \times W \end{aligned} \tag{5}$$

The mask-aware module has the following advantages over directly reusing features extracted by the tracking backbone. We select a portion of features that are sensitive only to the target region for segmentation refinement by introducing structural information. This not only reduces irrelevant information in the depth features but also improves the representation of the target, which guarantees the quality of the segmentation results.

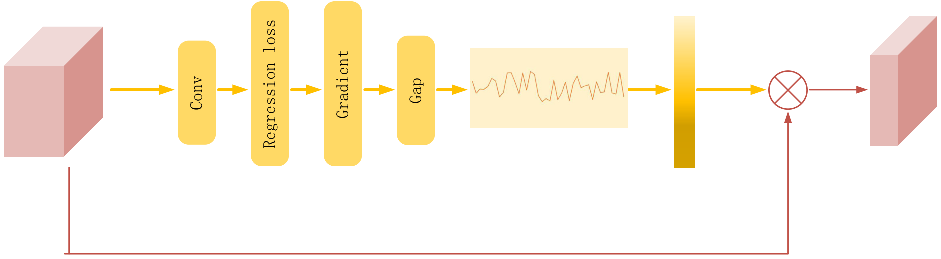


FIGURE 3. An illustration of the proposed mask-aware module. With the guidance of regression loss, the module selects features that are sensitive only to the target region for refinement of the mask within the features extracted by the tracking framework.

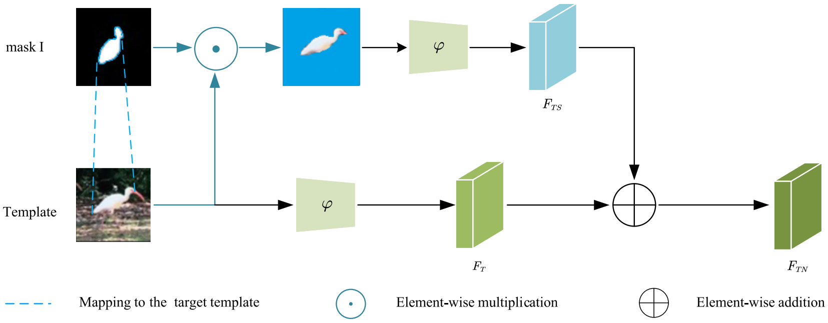


FIGURE 4. The framework of the mask-guided template suppression module.

C. MASK-GUIDED TEMPLATE SUPPRESSION MODULE (MTSM)

Most Siamese-based trackers regard tracking as a matching process, with the template patch as fixed kernels to match the search patch. However, the template contains some background information in addition to the target. While this contextual information can aid target location, it causes tracking to drift or even fail in more cases. Considering binary segmentation mask can clearly indicate whether a pixel belongs to the target, it was introduced to reduce the interference of background clutter. The mask-guided template suppression module is proposed to achieve this goal, which is presented in Figure 4. We consider the target mask of the initial frame to be credible because the ground truth value of the initial frame is given. Firstly, the target template feature Ft extracted by resnet50 and the target mask of the initial frame are fed into this module. Each pixel of the target is also mapped to the F_T, forming a region of interest M which can be identified by the mask I. Mask I is the output of the segmentation branch, which clearly indicates whether a pixel belongs to the target or not. Thus, with mask I, we construct a binary mask $M_I \in [0, 1]^{w\times h}$ corresponding to the elements on the template patch. Its elements for each pixel are calculated with Equation (6). The information of spatial location (i, j) is considered target-related information when the value of (i, j) on mask I is 1, and the corresponding value of $M\ (i, j)$ is 1; otherwise, the corresponding value of $M\ (i, j)$ is 0. Hence, the region of interest is located by all elements with a value of 1.

$$M\ (i, j) = \begin{cases} 1, & if\ \ (i, j) \in R \\ 0, & otherwise \end{cases} \tag{6}$$

So that we can obtain a background-suppressed feature by

$$F_{TS} = \varphi\ (T\ (i, j) \odot M_I\ (i, j)) \tag{7}$$

where T is the target template patch and φ () is the function to obtain deep features from the pre-trained neural network resnet50. φ () denotes Hadamard product. The new template patch F_{TN} is derived from the fusion operation between F_T and F_{TS} as follows:

$$F_{TN} = \beta * F_T + (1 - \beta) * F_{TS} \tag{8}$$

where β is a hyperparameter to control the influence of the background information. Then, the new template patch F_{TN} is reused inside the search region in the subsequent tracking process.

Through the above operations, background information is suppressed, while the information related to the target is enhanced, so that the updated template patch Ft can pay more attention to the target-related information. The background suppression template fs which contains only partial information related to the target, was not chosen as the new template patch. For the background information in the template cannot be completely discarded as the closer ones can be regarded as contextual information to locate the target. In addition, the module is only applied once

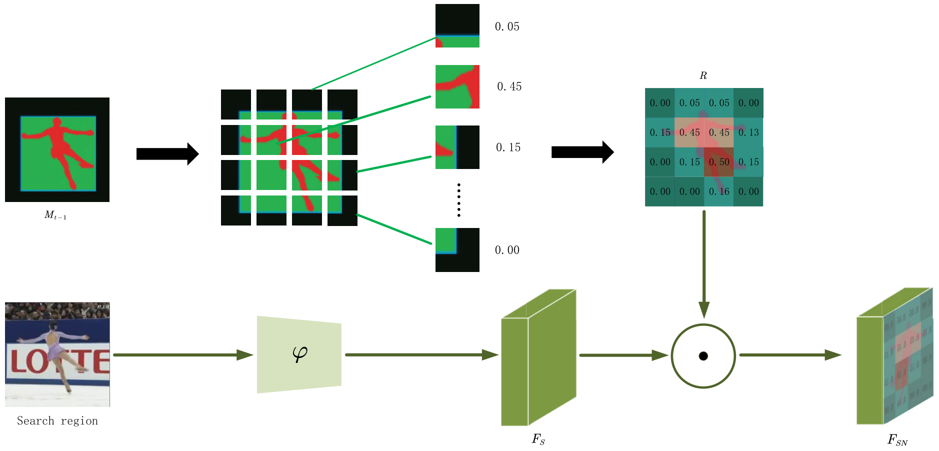


FIGURE 5. The overall pipeline of adaptive spatiotemporal context constraint strategy.

at initialization, so it improves performance without the additional burden of computation. (The effectiveness of the module was demonstrated in section IV-C by ablation experiments.)

D. ADAPTIVE SPATIOTEMPORAL CONTEXT CONSTRAINT CTRATEGY

During tracking, we only want to select a candidate with the highest similarity to the target template. However, other candidates from the background compromise the tracking accuracy. Therefore, an adaptive spatiotemporal context constraint strategy is proposed to alleviate this problem, as shown in Figure 5. The aim of this strategy is to suppress the background noise in the current frame, by means of learning the spatiotemporal distribution between the target and the interference in the previous frame.

In visual tracking, the search region is an area three times larger than the template centered on the target position of the previous frame. The information in the search area can therefore be considered the local context of the target. Since the time interval between video frames is usually small, it is reasonable to assume that successive frames are smooth between them. So that the spatial distribution of the target and surrounding interference remains essentially unchanged. Furthermore, the spatial relationship between an object and its surrounding context provides specific scene configuration information which aids to differentiate the target from the context. The closer to the target location the more important the information is and a larger weight should be given to it.

First, a region probability matrix is calculated with the search region feature and the mask. Specifically, we have the mask M_{t-1} after tracking at the $(t-1)\,th$ frame, in which the corresponding part of the search region can be cropped from M_{t-1} and mapped to the same size of the search patch. Then a grid of size $H \times W$ is computed, which is of the same spatial resolution as the search region feature F_S. We demonstrate how to calculate the regional probability matrix using a grid of 4×4 size In Figure 5. Denotes the grid box at the position (i, j) as $\left(x_{ij1}, y_{ij1}, x_{ij2}, y_{ij2}\right)$ and each grid value can be calculated as follows:

$$G_k^{(i,j)} = \frac{n}{\left(x_2^{ij} - x_1^{ij}\right)\left(y_2^{ij} - y_1^{ij}\right)} \tag{9}$$

The numerator n is the number of pixels in the grid area that belong to the target and the G_k indicates the overlap ratio of this grid with the target. All grid values together form a regional probability matrix R which represents the spatial distribution of the target in the search region. Then, the search patch F_S is multiplied bitwise by the regional probability matrix R to generate an interference-suppressed patch F_{SN}, which is formulated as Equation (9).

The strategy takes the target mask of the previous frame as a priori information to constrain the contextual information of the search area in the current frame. With the proposed strategy, contexts in the search patch are re-weighted, with information close to the target being given more weight, and far from the target being relatively suppressed. This effectively mitigates the effects of interference on tracking and improves tracking performance. The ablation experiment in this module is in subsection IV-C.

IV. EXPERIMENTS

In this section, the setup of this work is first presented in Section IV-A. Then, experiments are conducted in Section IV-B on five popular trace datasets to evaluate the performance of our method. Next, attribute-based experiments and ablation experiments are conducted in Sections IV-C and IV-D, respectively.

A. IMPLEMENTATION DETAILS

All experiments are implemented on a PC with an Intel i7-10700CPU 2.90 GHz) processor and a single NVIDIA GeForce GTX 1650 with 16 GB RAM. Like SiamMask, the template patch input size is 127 pixels and the search area input size is set to 255 pixels, the modified ResNet-50 in [33] is used as the base Siamese subnetwork. In the initial training, the convergence loss threshold is set to 0.002 and the maximum iteration number is 100. We select the top 250 important channel features from the Conv1, Conv2, and Conv3 layers respectively for the refinement of target masks. Our algorithm is evaluated on five benchmarks, including VOT2016 [47], VOT2018, VOT2019, OTB2015, and TC128.

B. EXPERIMENTS ON THE TRACKING BENCHMARK

1) VOT2016 BENCHMARK

In order to evaluate the performance of our method, we conduct experiments on the popular VOT2016 [47], which comprises 60 short sequences with various challenges. The VOT benchmark evaluates a tracker from three aspects: expected average overlap, accuracy, and robustness. The accuracy was indicated by calculating the average overlap ratio between the ground truth values and predicted results and the robustness is calculated by a number of tracking failures. The tracking is considered a failure when the overlap ratio between the predicted and ground truth values is zero. And the tracker is reinitialized for object tracking after 5 frames of the failure. The Expected Average Overlap (EAO) is used for overall performance ranking according to these two measures. We compared some state-of-the-art trackers (SiamFC [2], TCNN [48], CCOT [49], CSRDCF [50], SiamRPN [1], TADT [51], ECO [52], ASRCF [53], ATOM [54], SPM [55], and SiamMask [11]. Their scores are shown detailed in Table 1. ATOM and ASRCF obtained better robustness, but their EAO scores and overlap scores are lower than ours. In comparison to the SiamMask [11], which is a segmentation-based tracker as well, our EAO scores and Accuracy scores were 1.3% and 1.5% higher, respectively, and our robustness score is 1.8% lower. Overall, our method has the highest accuracy and EAO and better robustness score, indicating that the method effectively reduces the background interference problem.

TABLE 1. Details about the state-of-the-art trackers in VOT2016 [47]. red, blue and green, represent 1st, 2nd and 3rd respectively.

Tracker	EAO	Accuracy	Robustness
SiamFC	0.235	0.532	0.461
TCNN	0.325	0.550	0.268
CCOT	0.331	0.540	0.238
CSRDCF	0.338	0.510	0.238
SiamRPN	0.344	0.560	0.302
TADT	0.360	0.560	0.299
ECO	0.375	0.550	0.569
ASRCF	0.391	0.560	0.187
ATOM	0.430	0.610	0.180
SPM	0.434	0.620	0.210
SiamMask	0.436	0.621	0.214
OURS	0.449	0.636	0.196

2) OTB100 BENCHMARK

To further validate the performance of our tracker, we conducted experiments on VOT2018. VOT2018 contains 60 video sequences and, like VOT2016 [47], takes expected average overlap, accuracy, and robustness to measure the performance of the tracker. The videos contained in the VOT2018 dataset [56] are somewhat longer than those in VOT2016 and the challenges are more complex. In Table 2, we compare our tracker with others including SiamFC [1], ECO [52], ASRCF [53], SPM [55], SiamMask [11], DaSiamRPN [8], SiamRPN [1], LADCF [57], and ATOM [54]. DaSiamRPN [8] suppresses distractors by learning interference-aware features and obtains 0.383 EAO scores and 0.590 accuracy scores. Our method achieves EAO scores equivalent to its accuracy scores 1.7% higher, which indicates the advantage of employing segmentation information to suppress background interference. However, our method has not yielded the best performance. ATOM obtains better robustness and EAO scores, only our accuracy scores were higher than ours. This is probably because our tracker cannot cope with target disappearance and reappearance well.

TABLE 2. Details about the state-of-the-art trackers in VOT2018 [56]. red, blue and green, represent 1st, 2nd and 3rd respectively.

Tracker	EAO	Accuracy	Robustness
SiamFC	0.188	0.506	0.506
ECO	0.280	0.270	0.480
ASRCF	0.328	0.490	0.234
SPM	0.338	0.580	0.300
SiamMask	0.375	0.592	0.272
DaSiamRPN	0.383	0.590	0.276
SiamRPN	0.384	0.588	0.276
LADCF	0.389	0.503	0.159
ATOM	0.401	0.590	0.204
OURS	0.387	0.607	0.272

3) VOT2019 BENCHMARK

Then we evaluate our method on VOT2019 [58], in which sequences are replaced by 20% compared to the VOT2018. Table 3 shows the EAO, robustness, and accuracy of our trackers, SA-Siam [7], SiamCRF_RT [58], SPM [55], SiamMask [11], SiamRPN++ [17], ATOM [54], ARTCS [58], SiamDW_ST [10], DCFST [58] respectively. Although the overall performance metric EAO scores of our tracker are lower than ATOM in the vot2018 comparison results, in vot2019, the EAO and accuracy scores are comparable to or even slightly higher than it. And in the case where EAO scores are slightly lower than SiamDW_ST [10] and DCFST [58], our accuracy scores are higher than both. In general, our method ranks first on the accuracy metric on the VOT2019 dataset and also ranks high in EAO metrics. This shows that our tracker obtains better tracking accuracy, which validates the effectiveness of our target-guided mask refinement module. This module improves the reliability and accuracy of the object mask by introducing structural constraints in the mask generation process, thus improving the tracking accuracy of the tracker.

4) OTB100 BENCHMARK

To verify the versatility of our tracker, we conducted experiments on OTB100 [59]. It contains 100 videos, some

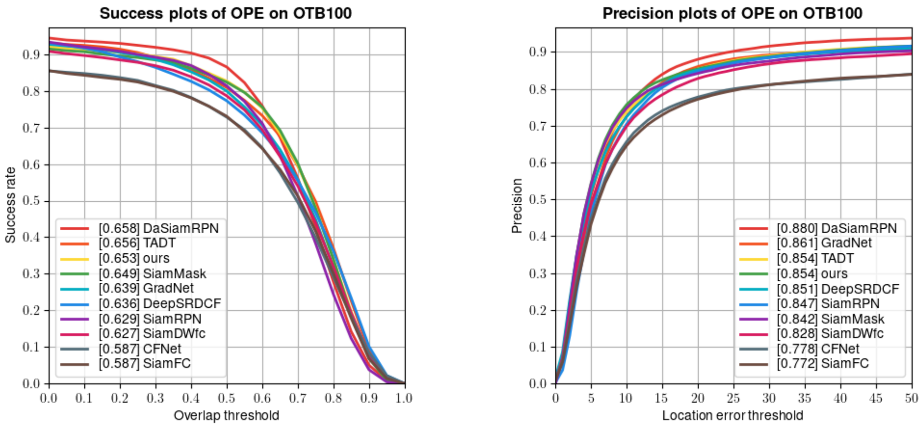


FIGURE 6. The Success plot and Precision plot on the Object Tracking Benchmark (OTB) 100 dataset.

TABLE 3. Details about the state-of-the-art trackers in VOT2019 [58]. red, blue and green, represent 1st, 2nd and 3rd respectively.

Tracker	EAO	Accuracy	Robustness
SA-Siam	0.252	0.547	0.492
SiamCRF_RT	0.252	0.549	0.346
SPM	0.262	0.577	0.507
SiamRPN++	0.275	0.590	0.492
ATOM	0.280	0.601	0.411
ARTCS	0.287	0.602	0.482
SiamDW_ST	0.287	0.600	0.467
DCFST	0.299	0.585	0.371
OURS	0.281	0.607	0.497

TABLE 4. Details about the state-of-the-art trackers in VOT2019 [58]. red, blue and green, represent 1st, 2nd and 3rd respectively.

Tracker	Success	Precision
KCF	0.386	0.557
BACF	0.495	0.660
CF2	0.495	0.692
SiamFC	0.503	0.688
CREST	0.533	0.708
SCS-Siam	0.538	0.742
SiamMask	0.540	0.725
SCSAtt	0.549	0.744
ECO	0.552	0.740
TADT	0.562	0.758
DeepSRDCF	0.543	0.730
OURS	0.548	0.736

of which are in grayscale. Different from VOT, OTB100 takes success and precision as the major evaluation metrics. The precision represents the percentage of error between the center point of the ground truth and the prediction bounding box. The success indicates the percentage of successfully tracked frames to the total number of video frames. The success of the frame tracking is determined when the percentage of overlap between the ground truth and the prediction bounding box is greater than a certain threshold. We compare our method with numerous trackers including Siammask [11], GradNet [60], DeepSRDCF [61], SiamRPN [1], SiamDWfc [10], SRDCF [62], CFNet [63], SiamFC [2], Staple [64]. The performance of the above trackers by one-pass evaluation (OPE) including both the accuracy plots and the success plots are shown in Figure 6.

Our method achieves a score of 0.653 and 0.854 on success and precision respectively. In comparison to the Siamese-based method SiamFC, our method obtained a 6.6% improvement in the success metric. This demonstrates the advantage of our approach of using pixel-level segmentation information to suppress background noise. TADT learns target-aware features that improve the tracker's ability to identify targets that undergo significant changes in appearance, and obtains 0.656 success scores. Our method achieves scores equivalent to its, which also proved that the proposed mask-guided background suppression module improves the discriminative ability of the target template. However, in contrast to the good performance on the VOT datasets, the overall performance of the proposed method on OTB100 is still somewhat distant from the state-of-the-art methods. The reason for this phenomenon is the targets in the OTB100 dataset are annotated with rectangles, while the tracking results generated by our method are presented in the form of quadrilaterals. In addition, the OTB dataset is not set up with a mechanism to reinitialize after tracking failure, which also proves that our approach cannot cope well with the challenge of target loss.

5) TC128 BENCHMARK

To demonstrate the universality of the proposed approach, additional experiments were conducted on the TC128 dataset [47]. We evaluated three distinct categories of tracking algorithms: correlation filter-based trackers such as KCF [41] and BACF [43], CNNs and CF-based trackers including CF2 [65], DeepSRDCF [61], and ECO [52], as well as Siamese network structure-based trackers including SiamFC [2], SiamMask [11], SCS-Siam [66], TADT [51], and SCSAtt [67]. CREST [68] was also included in the evaluation. The comparison results are presented in Table 4.

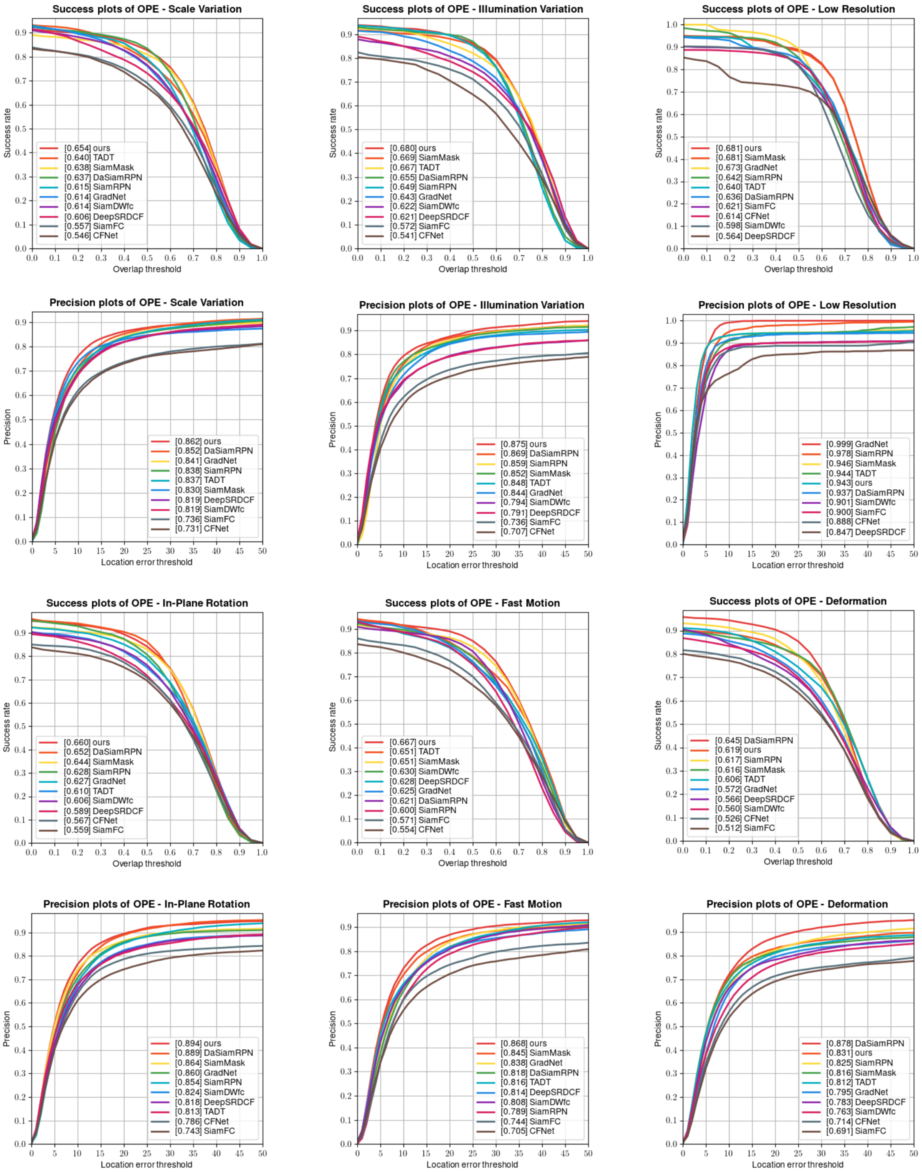


FIGURE 7. Comparison of success plots and precision plots on challenging attributes for tracking, including scale variation (SV), illumination variation (IV), low resolution, In-plane rotation (IPR), Fast Motion, and Deformation for OTB-100.

C. ATTRIBUTE-BASED COMPARISON

We compared the performance of nine trackers using success plots and precision plots on six attributes that are closely related to our method: fast motion, scale variation, illumination variation, low resolution, in-plane rotation, and deformation. The experiments were conducted on the OTB100 dataset. Figure 7 displays the performance of our proposed tracker and other trackers on six distinct attributes. Overall, the performance of our algorithm was found to be superior to that of most of the other trackers evaluated in the comparison study. What is worth emphasizing is that our method outperforms SiamMask for each attribute, which proves that our tracker makes fuller use of the segmentation information. For the attributes of scale variation and fast motion, our algorithm obtains the first scoring rate which indirectly proves that our object-guided background suppression module effectively enhances the discriminability of the target template.

D. ABLATION EXPERIMENTS

To verify the effectiveness of the proposed algorithm, we performed ablation experiments on the VOT2016 dataset. we investigate the effects of different combinations of three main components in our method, including the mask-aware module (MAM), mask-guided template suppression module (MTSM), and adaptive spatiotemporal context constraint strategy (ASC). Their contributions are validated by removing them separately from the entire framework. Detailed results are reported in Table 5. The baseline algorithm is a method without the mask-aware model and the background constraint method, which is described in detail in the subsections.

TABLE 5. Ablation studies on VOT2016.In the table, E represents the EAO metrics. The larger the value, the better the performance of the tracker. R indicates the robustness metrics. The smaller the value, the better the robustness of the tracker.

Tracker name	EAO	A	R	Lost Number
Baseline	0.436	0.621	0.214	46
Baseline + MAM	0.438	0.631	0.205	48
Baseline + MAM + MTSM	0.441	0.640	0.205	44
Baseline + MAM + MTSM + ASC	0.449	0.636	0.196	42

1) EFFECTIVENESS OF MASK-AWARE MODEL

Based on the baseline algorithm, the mask-aware model is introduced to improve the quality of the segmentation mask. Compared to the baseline algorithm, the score on the accuracy metric is improved with almost unchanged in other metrics. This demonstrates that the proposed mask-aware model can effectively improve the quality of the mask. For the model selects the features that best represent the target to generate the mask, which significantly reduces the negative impact of interference in the backbone network.

2) EFFECTIVENESS OF MASK-GUIDED TEMPLATE SUPPRESSION MODULE

In order to verify the validity of the background suppress module, we introduce this module with the other components unchanged. The mask-guided template suppression module improved the EAO score and Accuracy score and reduce the number of lost. This indicates that the utilization of this module effectively suppresses irrelevant information in the target template and highlights foreground information related to the target. For pixel-wise segmentation information can clearly distinguish whether each pixel belongs to the target and the background.

3) EFFECTIVENESS OF ADAPTIVE SPATIOTEMPORAL CONTEXT CONSTRAINT STRATEGY

Next, we further show the superiority of the adaptive spatiotemporal context constraint strategy. As can be seen from Table 5 the method with ASC performs better. Compared to the method in which the module is not included, ASC reduces the number of lost tracks while improving accuracy. This indicates that the proposed MSCM can significantly reduce the possibility of tracking drift. This is because the module learns a region probability matrix representing the spatial distribution of the target from the segmentation information of the previous frame, which is then used in the current frame. Our tracker maintains as much temporal and spatial information as possible, thus improving tracking performance.

V. CONCLUSION

Previous Siamese-based trackers use attentional mechanisms or explore distinguishable features to suppress interference, but the lack of structural information makes it difficult to highlight targets effectively. Considering that the output of semi-supervised object segmentation results in determining whether each pixel position belongs to the target, which is advantageous for separating the target from the background. We propose segmentation tracker with mask-guided background suppression strategy to suppress background interference by exploiting pixel-level segmentation information. First, we propose a mask-aware module that introduces structural information to help the network focus more on target region features and thus generate more accurate target masks. Then, we introduce the object mask with pixel-wise information into the template branch to filter the background interference, which obtains a more discriminative target template. Finally, the mask is introduced into the search region branch to help target location. The spatial distribution of the target in the search region in the previous frame is learned by the object mask to constrain the background interference in the current frame. We evaluated our tracker on the VOT2016, VOT208, VOT2019, OTB100, and TC128 datasets. The quantitative and qualitative results show that our method achieves superior tracking accuracy and robustness.

REFERENCES

[1] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, “High performance visual tracking with Siamese region proposal network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[2] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, “Fully-convolutional Siamese networks for object tracking,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Berlin, Germany: Springer, Oct. 2016, pp. 850–865.

[3] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, “End-to-end representation learning for correlation filter based tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5000–5008.

[4] W. Zhou, L. Wen, L. Zhang, D. Du, T. Luo, and Y. Wu, “SiamMan: Siamese motion-aware network for visual tracking,” 2019, *arXiv:1912.05515*.

[5] J. Zhu, T. Chen, and J. Cao, “Siamese network using adaptive background superposition initialization for real-time object tracking,” *IEEE Access*, vol. 7, pp. 119454–119464, 2019.

[6] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, “Learning attentions: Residual attentional Siamese network for high performance online visual tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4854–4863.

[7] A. He, C. Luo, X. Tian, and W. Zeng, “A twofold Siamese network for real-time object tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4834–4843.

[8] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, “Distractor-aware Siamese networks for visual object tracking,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 101–117.

[9] H. Fan and H. Ling, “Siamese cascaded region proposal networks for real-time visual tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7944–7953.

[10] Z. Zhang and H. Peng, “Deeper and wider Siamese networks for real-time visual tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4586–4595.

[11] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1328–1338.

[12] Y. Wang, J. Choi, K. Zhang, Q. Huang, Y. Chen, M.-S. Lee, and C.-C.-J. Kuo, "Video object tracking and segmentation with box annotation," *Signal Process., Image Commun.*, vol. 85, Jul. 2020, Art. no. 115858.

[13] B. Xin Chen and J. K. Tsotsos, "Fast visual object tracking with rotated bounding boxes," 2019, *arXiv:1907.03892*.

[14] A. Lukežic, J. Matas, and M. Kristan, "D3S—A discriminative single shot segmentation tracker," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7131–7140.

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

[16] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1420–1429.

[17] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4277–4286.

[18] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12549–12556.

[19] X. Dong and J. Shen, "Triplet loss in Siamese network for object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 459–474.

[20] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6181–6190.

[21] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland. Berlin, Germany: Springer, Sep. 2017, pp. 254–265.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[23] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "SiamCAR: Siamese fully convolutional classification and regression for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6268–6276.

[24] Y. Zhang, Z. Wu, H. Peng, and S. Lin, "A transductive approach for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6947–6956.

[25] Y.-T. Hu, J.-B. Huang, and A. G. Schwing, "VideoMatch: Matching based video object segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 54–70.

[26] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3491–3500.

[27] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.

[28] I. Kompatsiaris and M. G. Strintz, "Spatiotemporal segmentation and tracking of objects for visualization of videoconference image sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 8, pp. 1388–1402, Aug. 2000.

[29] V. Belagiannis, F. Schubert, N. Navab, and S. Ilic, "Segmentation based particle filtering for real-time 2D object tracking," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, Florence, Italy. Berlin, Germany: Springer, Oct. 2012, pp. 842–855.

[30] J. Son, I. Jung, K. Park, and B. Han, "Tracking-by-segmentation with online gradient boosting decision tree," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3056–3064.

[31] D. Yeo, J. Son, B. Han, and J. H. Han, "Superpixel-based tracking-by-segmentation using Markov chains," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 511–520.

[32] P. Voigtlaender, J. Luiten, P. H. S. Torr, and B. Leibe, "Siam R-CNN: Visual tracking by re-detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6577–6587.

[33] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[34] J. Luiten, P. Voigtlaender, and B. Leibe, "PReMVOS: Proposal-generation, refinement and merging for video object segmentation," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 565–580.

[35] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 1251–1258.

[36] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5000–5009.

[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland. Berlin, Germany: Springer, Sep. 2014, pp. 740–755.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, vol. 9351, Munich, Germany. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[39] T. Liu, G. Wang, and Q. Yang, "Real-time part-based visual tracking via adaptive correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4902–4912.

[40] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2014, pp. 65.1–65.11.

[41] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[42] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland. Berlin, Germany: Springer, Sep. 2014, pp. 127–141.

[43] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1144–1152.

[44] K. Tan, T.-B. Xu, and Z. Wei, "Online visual tracking via background-aware Siamese networks," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 10, pp. 2825–2842, Oct. 2022.

[45] B. Wei, H. Chen, Q. Ding, and H. Luo, "SiamOAN: Siamese object-aware network for real-time target tracking," *Neurocomputing*, vol. 471, pp. 161–174, Jan. 2022.

[46] P. O. Pinheiro, T. Y. Lin, R. Collobert, and P. Dollr, "Learning to refine object segments," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 75–91.

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

[48] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6875–6879.

[49] M. Danelljan, A. Robinson, F. Shahbaz Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 472–488.

[50] A. Lukežic, T. Vojír, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4847–4856.

[51] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-aware deep tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1369–1378.

[52] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6931–6939.

[53] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4665–4674.

[54] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4655–4664.

[55] G. Wang, C. Luo, Z. Xiong, and W. Zeng, "SPM-tracker: Series-parallel matching for real-time visual object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3638–3647.

[56] M. Kristan et al., "The sixth visual object tracking VOT2018 challenge results," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 3–53.

[57] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5596–5609, Nov. 2019.

[58] M. Kristan et al., "The seventh visual object tracking VOT2019 challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2206–2241.

[59] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.

[60] P. Li, B. Chen, W. Ouyang, D. Wang, X. Yang, and H. Lu, "GradNet: Gradient-guided network for visual object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6161–6170.

[61] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 621–629.

[62] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4310–4318.

[63] Z. Shen, Y. Dai, and Z. Rao, "CFNet: Cascade and fused cost volume for robust stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13901–13910.

[64] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1401–1409.

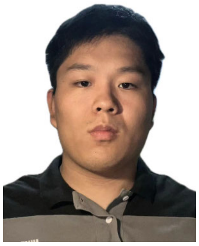
[65] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3074–3082.

[66] M. Fiaz, A. Mahmood, and S. K. Jung, "Learning soft mask based feature fusion with channel and spatial attention for robust visual object tracking," *Sensors*, vol. 20, no. 14, p. 4021, Jul. 2020.

[67] M. M. Rahman, M. Fiaz, and S. K. Jung, "Efficient visual tracking with stacked channel-spatial attention learning," *IEEE Access*, vol. 8, pp. 100857–100869, 2020.

[68] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M.-H. Yang, "CREST: Convolutional residual learning for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2574–2583.

ERLIN TIAN was born in Henan, China, in 1980. He received the M.S. degree in communication and information systems from Huazhong University of Science and Technology, Wuhan, China, in 2008. He was an Associate Professor at Zhengzhou University of Light Industry. His research interests include pattern recognition and image processing.

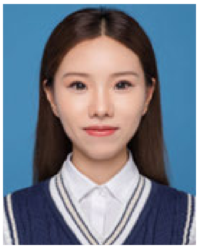


YUNPENG LEI was born in Henan, China, in 2002. He received degree in smart grid information engineering from the School of Electrical Information Engineering, Zhengzhou University of Light Industry, in 2022. His research interests include object tracking and object segmentation

JUNFENG SUN received the bachelor's degree. He is currently working as a Engineer with China Tobacco Guangxi Industrial Company Ltd. His research interests include application of power control and automation control technology.

KEYAN ZHOU was born in Henan, China, in 1999. She received the B.S. degree in smart grid information engineering and the M.S. degree in control science and engineering from Zhengzhou University of Light Industry, China, in 2021 and 2024, respectively. Her research interests include object tracking and object segmentation.

BIN ZHOU was born in Henan, China, in 1996. She received the B.S. degree in electrical engineering and automation from Zhengzhou University of Aeronautics, in 2018, and the M.S. degree in electrical engineering from Zhengzhou University of Light Industry, in 2024. Her research interests include object segmentation and object tracking.



HANFEI LI received the master's degree in rural regional development from Henan University of Science and Technology, in 2022. She is currently working with the Department of Science and Technology, Zhengzhou University of Light Industry. Her research interests include regional economic development and social public management.