frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 29 November 2023
DOI 10.3389/fphy.2023.1269638

OPEN ACCESS

EDITED BY
Gang (Gary) Ren,
The Molecular Foundry, United States

REVIEWED BY
Fasheng Wang,
Dalian Nationalities University, China
Guohui Wang,
Xi'an Technological University, China

*CORRESPONDENCE
Jin Duan,
✉ duanjin@vip.sina.com

RECEIVED 30 July 2023
ACCEPTED 02 October 2023
PUBLISHED 29 November 2023

CITATION
Huang D, Yu S, Duan J, Wang Y, Yao A, Wang Y and Xi J (2023), Spatio-temporal interactive fusion based visual object tracking method.
Front. Phys. 11:1269638.
doi: 10.3389/fphy.2023.1269638

Spatio-temporal interactive fusion based visual object tracking method

Dandan Huang, Siyu Yu, Jin Duan*, Yingzhi Wang, Anni Yao, Yiwen Wang and Junhan Xi

College of Electronic Information Engineering, Changchun University of Science and Technology, Changchun, China

Visual object tracking tasks often struggle with utilizing inter-frame correlation information and handling challenges like local occlusion, deformations, and background interference. To address these issues, this paper proposes a spatio-temporal interactive fusion (STIF) based visual object tracking method. The goal is to fully utilize spatio-temporal background information, enhance feature representation for object recognition, improve tracking accuracy, adapt to object changes, and reduce model drift. The proposed method incorporates feature-enhanced networks in both temporal and spatial dimensions. It leverages spatio-temporal background information to extract salient features that contribute to improved object recognition and tracking accuracy. Additionally, the model's adaptability to object changes is enhanced, and model drift is minimized. A spatio-temporal interactive fusion network is employed to learn a similarity metric between the memory frame and the query frame by utilizing feature enhancement. This fusion network effectively filters out stronger feature representations through the interactive fusion of information. The proposed tracking method is evaluated on four challenging public datasets. The results demonstrate that the method achieves state-of-the-art (SOTA) performance and significantly improves tracking accuracy in complex scenarios affected by local occlusion, deformations, and background interference. Finally, the method achieves a remarkable success rate of 78.8% on TrackingNet, a large-scale tracking dataset.

KEYWORDS

object tracking, spatio-temporal context, feature enhancement, feature fusion, attention mechanism

1 Introduction

Visual object tracking technology is one of the important research directions in the field of computer vision, which is widely used in intelligent surveillance, unmanned driving, human–computer interaction, *etc.* The tracking method using correlation filtering is shown in [1], but with the emergence of SiamFC [2], the Siamese network-based tracking framework has become the mainstream of the single-object tracking algorithm framework, and a series of tracking algorithms have been generated based on it. However, such tracking algorithms have the following shortcomings:

1) Most of the current Siamese network-based trackers generally use the initial frame as a template tracking strategy, which makes it difficult for the algorithm to adapt to situations such as severe object deformation and occlusion, resulting in poor robustness and accuracy. In addition, these trackers use only the appearance information of the current frame, failing

to take full advantage of the abundant temporal contextual information in the historical frame sequences and ignoring the temporal and spatial correlation of the objects.

To improve this situation, some trackers introduce a template update mechanism or use multiple templates, such as UpdateNet [3], while others consider the spatio-temporal correlation and introduce an attention mechanism approach for improvement, such as SiamAttn [4] and SiamADT [5]. The aforementioned approaches can enhance the robustness of the tracker to some extent, but such strategies are mainly of limited use and inevitably increase the computational effort.

2) Most of the existing popular template frame feature and query frame feature trackers still use correlation operations for fusion, such as SiamFC, which causes the lack of semantic information and global information to some extent. The absence of all this information causes such tracking algorithms to encounter difficulty adapting to changes in the appearance of the object in the face of challenges, such as local occlusions and deformations, thus reducing the tracking performance of the algorithms and considerably limiting their use.

In response to the aforementioned analysis, this paper stores multiple historical frame information as memory frames in the tracking process and enhances the features in both temporal and spatial dimensions, with the aim of breaking the conventional pattern that most algorithms invariably use the initial frame *a priori* information as the tracking template and completely exploiting the hidden spatio-temporal contextual information in the historical frame sequence, while enabling the tracker to better adapt to the changing appearance of the object; in order to improve tracking accuracy to achieve a strict comparison between candidate objects and template, this paper designs a spatio-temporal interactive fusion (STIF) network for establishing the relationship between memory frame and query frame features and also obtains a more robust global feature representation by feature interaction between the two based on feature enhancement. By mining and exploiting the aforementioned information, the proposed algorithm improves the accuracy and robustness of the tracking model. The main contributions of this paper are summarized as follows.

- An end-to-end spatio-temporal interactive fusion object tracking framework is proposed. The whole network is not only simple in structure but also has a strong adaptive capability to the changing appearance of the object in different scenarios.
- A feature enhancement network is established through which the feature sequences are processed to capture the dynamic features of the object in both the temporal and spatial dimensions. The network can achieve a meticulous capture of the object information in all directions, making the extracted object features more significant, thus leading to higher accuracy of the tracking algorithm.
- A spatio-temporal interactive fusion network is proposed, based on the feature enhancement network, to achieve fine alignment and matching of the features of the two branches through the mutual transfer and influence of information between the memory and query branches and enhance the feature expression capability, which can effectively solve the problem of model drift caused by occlusion and deformation. The adversarial interference capability of the model will be improved during the tracking process to achieve more robust tracking.

2 Related work

In recent years, the rapid development of object tracking technology has led to the emergence of many object tracking algorithms. Among them, the emergence of Siamese network-based object tracking methods has transformed the search problem into similarity matching and solved the previous problem of large computational effort for visual tracking tasks; however, in the reasoning process of such algorithms, the given template information is generally intercepted from the first frame of the video sequence and matched with the current search region for feature information to achieve tracking. It has excellent performance and real-time tracking speed in many traditional tracking scenarios. Facing the aforementioned problems, DSiam [7] proposed a tracking algorithm based on the dynamic Siamese network, which uses a dynamic model to analyze the motion of the object and improves the accuracy and stability of tracking by dynamic feature extraction and template update. In addition, UpdateNet linearly updated the object template using a moving average method with a fixed learning rate and proposed an adaptive template updating strategy that combines the initial frame template, the accumulated template, and the current frame template to achieve the prediction of the best template for the next frame. STARK [8] sets up a dynamic template and determines whether the dynamic template is updated by setting a threshold on the prediction confidence. STMTrack [9] proposes a tracking framework based on spatio-temporal memory networks, which can make full use of the historical information associated with the object, thus avoiding template updates and achieving a template-free framework.

In addition, when the tracking object contains multiple spatial dimensions as well as a temporal dimension, the spatio-temporal information needs to be used to focus on the dynamic characteristics of the object at different locations and different moments for more accurate tracking. In order to make full use of spatio-temporal contextual information, object tracking methods based on attention mechanisms are also widely used. [10] introduced an encoder–decoder attention module to filter different features by compressing the feature map and establishing relationships between channels in the Siamese network. Efficient visual tracking with a stacked channel–spatial attention (SCSAtt) mechanism [11] improves the accuracy of the model by introducing an attention mechanism to explore the object features and designing a linearly stacked attention mechanism. SA-Siam [12] extracts different features of the object through a dual-branching structure of semantics and appearance and uses a channel attention mechanism for feature selection of the object but ignores template updates.

Different from these methods, our network considers spatio-temporal inter-frame correlation and makes full use of multiple historical frames of the tracking process, enhances features, and improves focus on candidate object areas by capturing dynamic features of the object in time and space. Based on this, the spatio-temporal interaction fusion network communicates the feature information of the memory branch with the query branch to achieve secondary screening of features and further enhances the feature expression capability to obtain global features, which can significantly improve the tracking performance of the model in challenging scenarios.

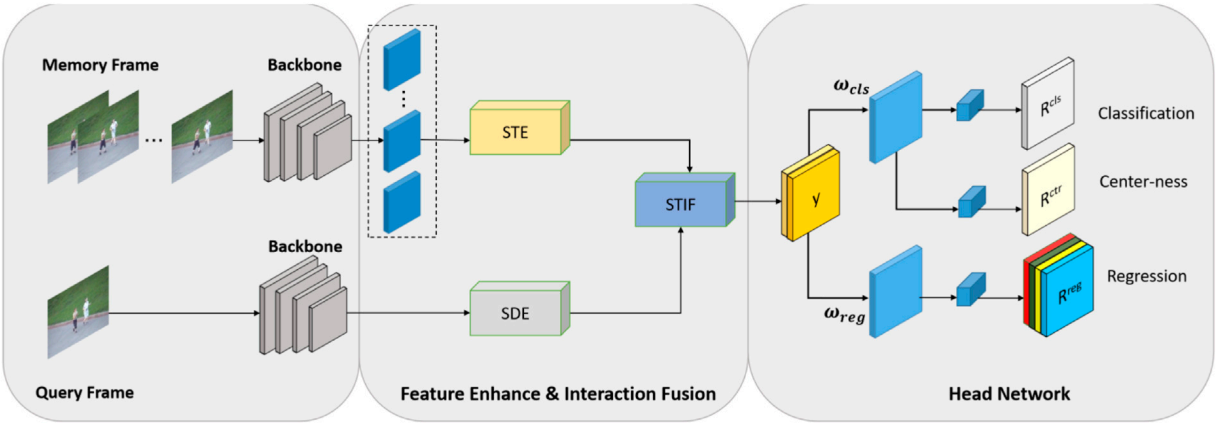


FIGURE 1
Schematic diagram of the processing flow of the algorithm in this paper. The backbone network is used to extract the memory frame features and query frame features as inputs to the overall structure. Subsequently, the feature enhancement operation is carried out, where the memory frame performs the spatio-temporal dimension enhancement (STE and SDE) operation to fully combine the spatio-temporal context information to enrich the template information, and the query frame performs the spatial dimension enhancement (SDE) operation to improve the focus of the object region. Then, spatio-temporal interactive fusion (STIF) of features is carried out to achieve the interaction of the two branches of features, and the final feature map is obtained and sent to the head network for tracking result estimation. Reproduced from the OTB100 dataset, [20], with permission from IEEE.

3 Methods

In this section, we first introduce the general framework of the proposed spatio-temporal interactive fusion-based object tracking method. In addition, we describe each part in detail in Section 3.1, Section 3.2, and Section 3.3.

3.1 Feature enhancement network

The spatio-temporal association between frames can enrich the template information, while the use of the spatio-temporal context can accurately capture the temporal and spatial changes of the object, ensuring that the features have sufficient characterization power. We use spatio-temporal feature enhancement to enrich the template information and query the features of the frames. The feature enhancement module is mainly divided into two parts: memory frame enhancement and query frame enhancement. The memory frame enhancement includes both temporal and spatial dimensions to enhance the temporality and expressiveness of the features, while for the query frame, as it only provides the image information of the current moment, only spatial enhancement is used, focusing more on the expression and extraction of spatial information. Figure 1 shows the overall framework of our network.

3.1.1 Memory frame enhancement

The memory frame features are mainly used in a spatio-temporal enhancement (STE) network, as shown in Figure 2, which mainly includes time dimension enhancement (TDE) and spatial dimension enhancement (SDE). The features are adaptively optimized by assigning different weights to each time and location based on the response to the object. End-to-end training can be achieved without adding additional parameters using spatio-temporal feature enhancement networks. This process is shown in the following equation:

$$\begin{cases} \boldsymbol{F}_m^T = \boldsymbol{N}_T(\boldsymbol{F}_m) \otimes \boldsymbol{F}_m, \\ \boldsymbol{F}_m^S = \boldsymbol{N}_S(\boldsymbol{F}_m^T) \otimes \boldsymbol{F}_m^T, \end{cases} \tag{1}$$

where F_m is the memory branch input feature, N_T denotes the feature processed by STE, N_S denotes the feature processed by SDE, F_m^T and F_m^S are the features enhanced by STE and SDE, respectively, and ⊗ denotes element-wise multiplication.

Due to the temporal correlation between video sequence frames, multiple consecutive frames are used as reference frames to capture subtle appearance variations in the object during motion, thus enhancing the discriminative features in the tracking instances. Therefore, in STE, we mainly use the T memory frames adjacent to the frame to be tracked as reference frames, and the overall process is as follows: first, we use the feature extraction network to obtain the memory frame feature map F_m; then, we feed the feature map F_m into the STE network; finally, we obtain the enhanced features F_m^T after the F_m time-attentive network according to the following strategy:

$$\boldsymbol{F}_m^T = \sum\nolimits_{i=t-T}^{t-1} \boldsymbol{\omega}_i \boldsymbol{F}_{m_i}, \tag{2}$$

where ω_i is the fusion weight of the corresponding reference frame, and the reference frame weight is positively correlated with its contribution to the tracking template. F_{m_i} is the feature of the ith reference frame. Since the initial frame contains more comprehensive information about the appearance of the object, it has a certain reference value when the object is occluded or motion blur occurs. The similarity between the reference frame features and the initial frame features is computed to obtain the fusion weight coefficients:

$$\boldsymbol{\omega}_i = \boldsymbol{SoftMax}\left(\frac{\boldsymbol{F}_{m_i}\boldsymbol{F}_{m_0}^{Trans}}{\sqrt{\boldsymbol{C}}}\right), \tag{3}$$

where F_{m_0} is the video initial frame image feature, C is the number of channels, and $F_{m_0}^{Trans}$ is the transpose of the initial frame. Finally, these weights are applied to the feature weights, and the adaptive assignment

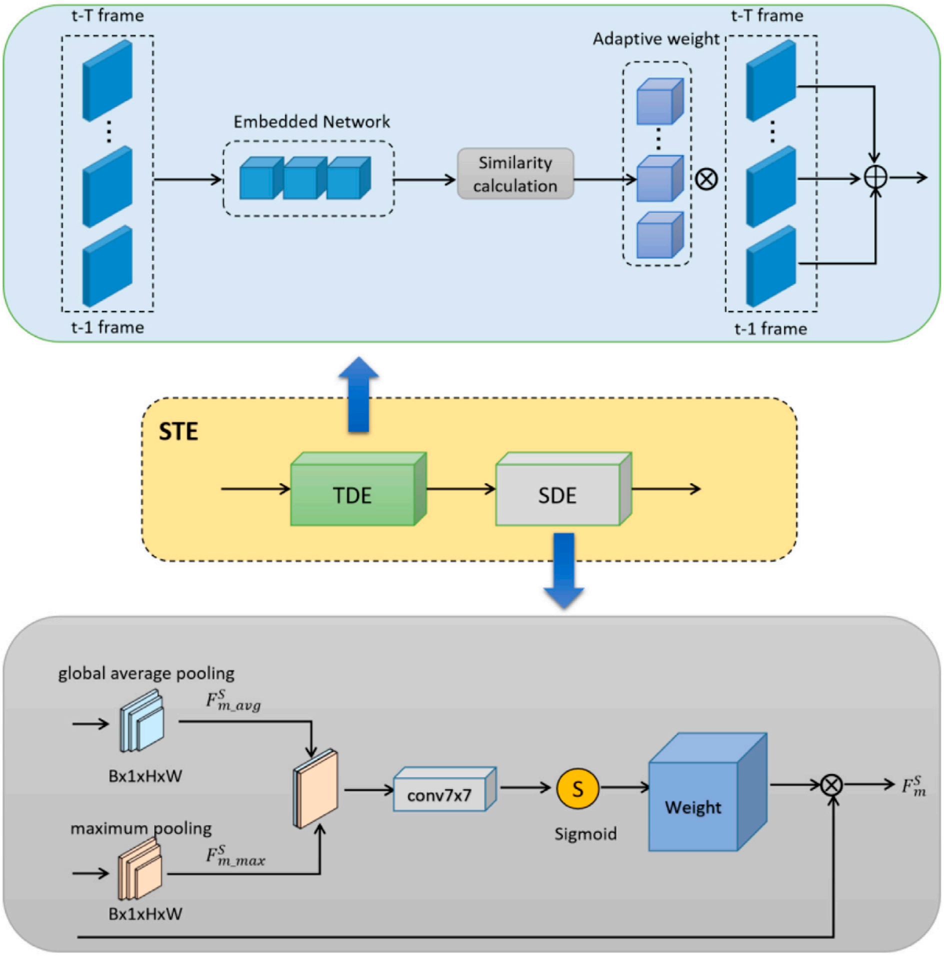


FIGURE 2
Spatio-temporal feature enhancement network structure.

of fusion weights can be implemented according to Eqs 2, 3 to obtain the temporally enhanced memory frame feature output.

Unlike temporal dimension enhancement, spatial dimension enhancement focuses on the location information of the object in the feature image and improves the sensitivity of the network to spatial feature information by highlighting or weakening the feature information at different spatial locations. The specific procedure is shown in the following equation:

$$\xi_1 = \sigma\left(f^{7\times7} C_{cat}\left(F^S_{m_{avg}}, F^S_{m\_max}\right)\right), \tag{4}$$

where ξ_1 denotes the feature output after the aforementioned processing; $C_{cat}(\cdot,\cdot)$ denotes the connected maximum pooling and average pooling; $f^{7\times7}$ is the convolution layer with the convolution kernel of 7×7; σ denotes the sigmoid function normalization; and $F^S_{m\_avg}$ and $F^T_{m\_avg}$ are the features of F^T_m after global average pooling and maximum pooling, respectively. The final spatially enhanced memory frame features are shown as

$$F^S_m = \xi_1 F^T_m. \tag{5}$$

3.1.2 Query frame enhancement

For the query frame, as it only provides information about the image at the current moment, only spatial enhancement is used, focusing more on the representation and extraction of spatial information. Spatial enhancement is mainly used to improve the spatial sensitivity and accuracy of the features by optimizing them, making them more accurate in reflecting spatial information such as the morphology and edges of the object.

Like memory frame enhancement, the result of processing after spatial pooling is shown in Eq. 6:

$$\xi_2 = \sigma\left(f^{7\times7} C_{cat}\left(F^S_{m_{avg}}, F^S_{m\_max}\right)\right), \tag{6}$$

where ξ_2 denotes the feature output in the query branch after the aforementioned processing. The final spatially enhanced query frame features are shown as

$$F^S_q = \xi_1 F_q. \tag{7}$$

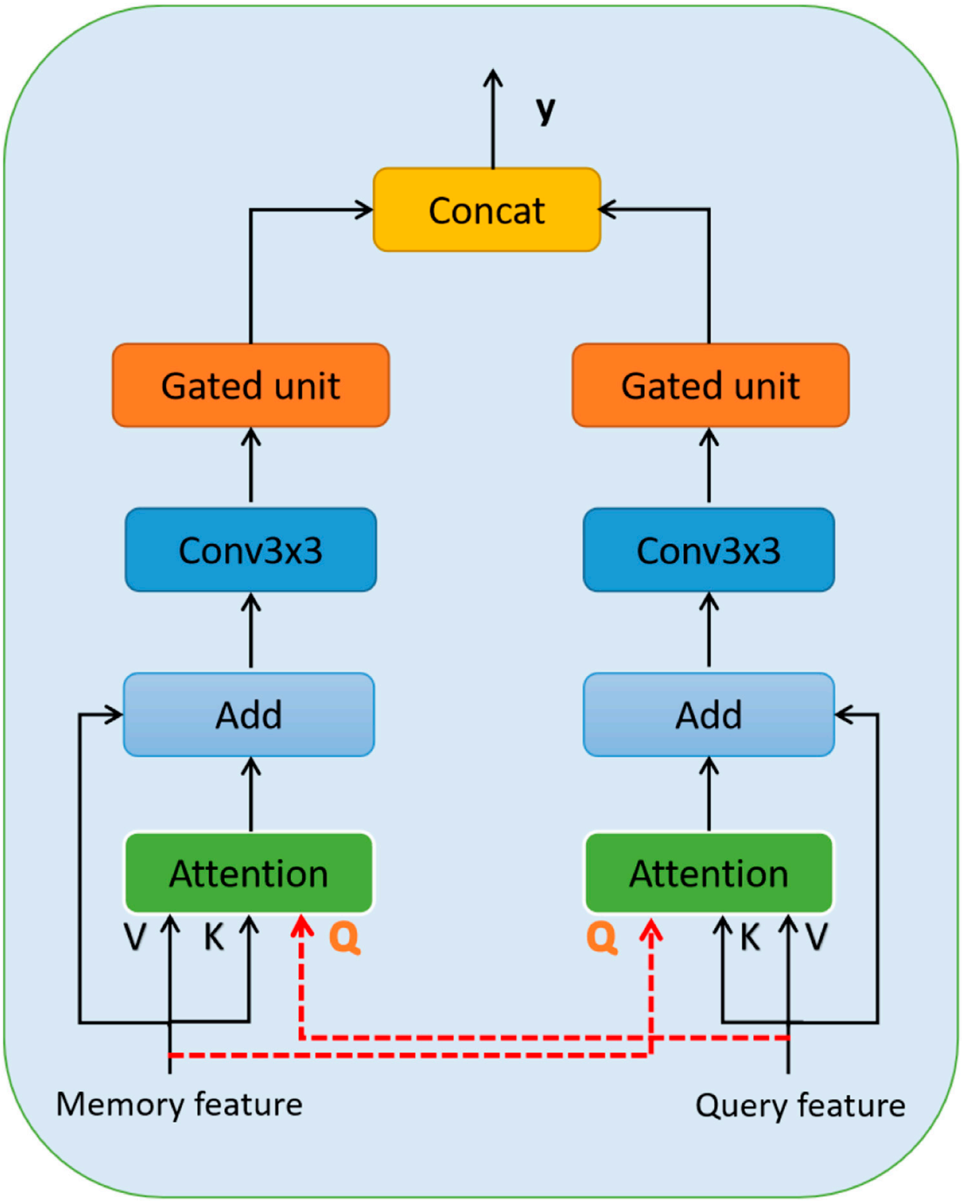


FIGURE 3
spatio-temporal interactive fusion network structure.

3.2 Spatio-temporal interactive fusion network

The main purpose of the spatio-temporal interactive fusion (STIF) network is to compare and complement features from different frames to achieve better feature representation and tracking results. As shown in Figure 3, the memory branch and the query branch after feature enhancement are used as network inputs. By cross-comparing the features of the memory and query frames and emphasizing the importance and priority between different features, the dynamic features and correlations of the object in time and space can be effectively captured, and the features are fused into more robust and accurate features to achieve effective information interaction and fusion between the two branches. The attention model used in the network is defined as shown in Eq. 8, with three inputs, Q, K, and V, as query, key, and value, respectively, and d_k is the dimensionality of the key.

$$\boldsymbol{Attention}(Q, K, V) = \boldsymbol{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{8}$$

Specifically, based on the feature enhancement network processing, we use the memory frame feature F_m^S and the query frame feature F_q^S obtained by the enhancement process as inputs to the spatio-temporal interactive fusion network. Taking the memory branch as an example, first, consider F_m^S as V and K of the memory branch and F_q^S as Q of the memory branch; the similarity measure between the memory frame features and the query frame features can be obtained by

TABLE 1 Comparison of the tracking performances of different networks in the same experimental environment and datasets.

Dataset	Baseline tracker	Fm + STE	Fq + SDE	+STIF	AO	$SR_{0.5}$	$SR_{0.75}$
GOT-10k	√				0.642	0.737	0.545
		√			0.644	0.739	0.550
			√		0.643	0.740	0.550
				√	0.644	0.742	0.549
		√	√		0.647	0.750	0.575
		√	√	√	0.652	0.781	0.590

TABLE 2 Test results of the GOT-10k dataset, where AO denotes the average overlap rate, SR denotes the success rate, and trackers are sorted by AO values from top to bottom.

Tracker	AO	$\boldsymbol{SR_{0.5}}$	$\boldsymbol{SR_{0.75}}$
SiamFC [2]	0.348	0.353	0.098
SiamRPN++ [26]	0.517	0.616	0.325
DiMP [27]	0.611	0.717	0.492
Ocean [28]	0.611	0.721	0.473
SiamGAT [6]	0.627	0.743	0.575
STMTrack [9]	0.642	0.737	0.575
UAST [24]	0.648	0.751	0.578
Ours	**0.652**	**0.781**	**0.590**

The meaning of the bold values in table primarily represents the training results of the methods proposed in this article.

computing the similarity between K and Q, and the corresponding similarity weights are then weighted with the memory frame features to enable feature information interaction between the memory branch and the query branch. In addition, before feature fusion, we use a gating unit to adjust the features, mainly to strengthen the nonlinear nature and generalization ability of the fusion network to avoid gradient explosion and increase the number of parameters during model training. Finally, the memory frame features after the interaction are combined with the query frame features in the second dimension to obtain the final fused feature map y. The query branch operation is the same as mentioned previously. The specific process is formulated as follows:

$$\boldsymbol{y_1 = C_{conv}\left(Attn\left(F_m^S, F_q^S\right)\right)}, \tag{9}$$

$$\boldsymbol{y_2 = C_{conv}\left(Attn\left(F_q^S, F_m^S\right)\right)}, \tag{10}$$

$$\boldsymbol{y = concat\left(y_1, y_2\right)}, \tag{11}$$

where y_1 denotes the memory frame features after interaction, y_2 denotes the query frame features after interaction, $concat(\cdot,\cdot)$ denotes the fusion join operation; C_{conv} denotes a convolutional layer with a convolutional kernel of 3 × 3, and Attn denotes the attention operation.

3.3 Classification regression head network

Inspired by the fact that the first-level anchorless detector [13] has better detection and fewer computational parameters than the first-level detector [14] based on anchor frame point boxes for object detection, we employ an anchorless head network that contains a classification branch to classify the information in the image and an anchorless regression branch for direct estimation of the object's bounding box.

A lightweight classification convolutional network ω_{cls} is first used to encode the fused feature mapping, which combines information from memory frames and query frames and is more suitable for the classification task. The output dimension of ω_{cls} is then reduced to 1 using a linear convolutional layer with a 1 × 1 kernel, and finally, the classification response map $R^{cls} \in R^{1\times H\times W}$ is obtained.

Furthermore, since positive samples close to the object boundary tend to predict low-quality object bounding boxes, a branch is bifurcated after ω_{cls} for the centrality response map $R^{ctr} \in R^{1\times H\times W}$, as shown on the right side of Figure 1. In the inference process, R^{ctr} is multiplied with R^{cls} to suppress the classification confidence score of pixels at a distance from the object center. In the regression branch, we use the feature mapping y for another lightweight regression convolutional network ω_{reg} and then reduce the dimensionality of the output features to 4 to generate the regression response mapping $R^{reg} \in R^{4\times H\times W}$. Finally, the classification loss L_{reg} and regression loss L_{cls} are used as functions, as described in [7, 15], respectively, and the final overall loss function can be expressed as

$$\boldsymbol{L = L_{cls} + \lambda_1 L_{reg} + \lambda_2 L_{ctr}}, \tag{12}$$

where λ_1 and λ_2 are both weighting factors.

4 Experimental results and analysis

4.1 Experimental details

4.1.1 Datasets

In the training phase, we mainly used the training sets of TrackingNet [15], LaSOT [16], GOT-10k [17], ILSVRC VID [18], ILSVRC DET [18], and COCO [19] as the training datasets except for the GOT-10k [18] benchmark; in the testing phase, we mainly used TrackingNet [15], LaSOT [16], GOT-10k [17], OTB100 [20], and WATB [21]datasets for testing and comparing with other object tracking methods. In addition, four sequences with occlusion and appearance deformation properties were extracted from the LaSOT [16] dataset for testing.

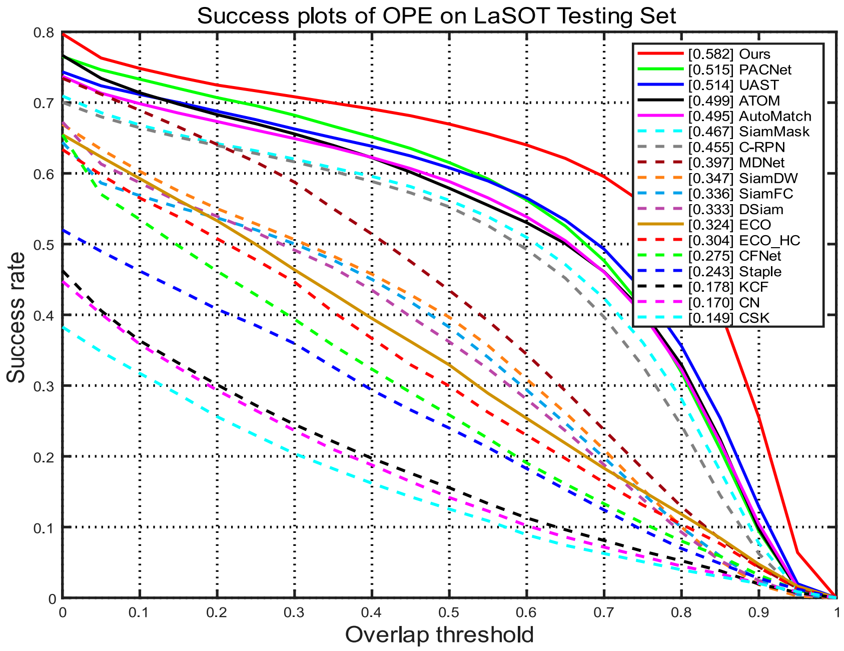


FIGURE 4
Success rate of our method on the LaSOT dataset.

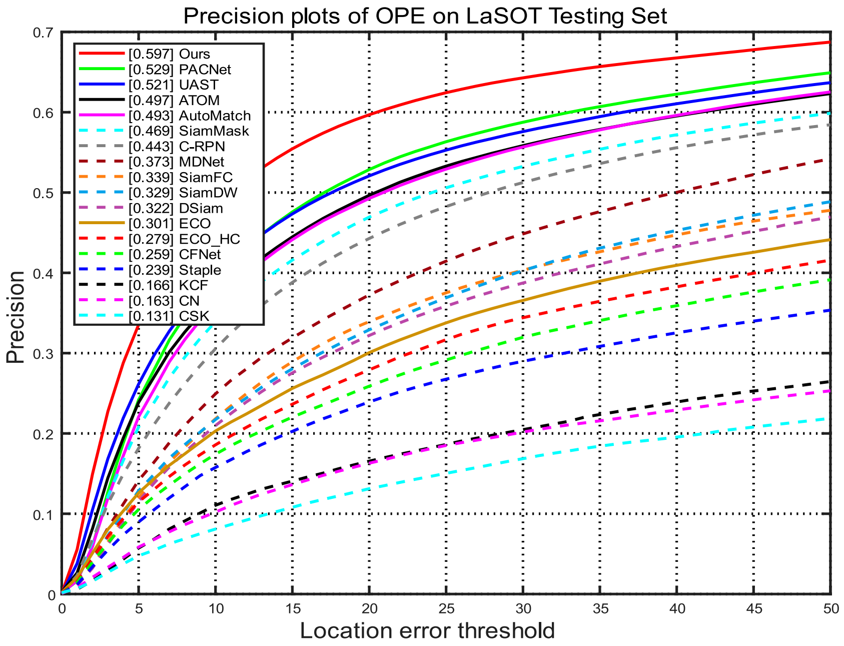


FIGURE 5
Precision of our method on the LaSOT dataset.

4.1.2 Experimental setting

4.1.2.1 Model setup

We used the PyTorch [22] framework for our experiments, employing GoogLeNet [23] as the backbone of our feature extraction network $\varphi_m \varphi_q$ and both the classification convolutional network ω_{cls} and the regression convolutional network ω_{reg} consisting of seven convolutional layers. Each convolutional layer in ω_{cls} and ω_{reg} is followed by a ReLU activation function.

4.1.2.2 Optimization and training strategies

We mainly use SDG to optimize the loss function during the training process. A total of 20 epochs are set for the experiment,

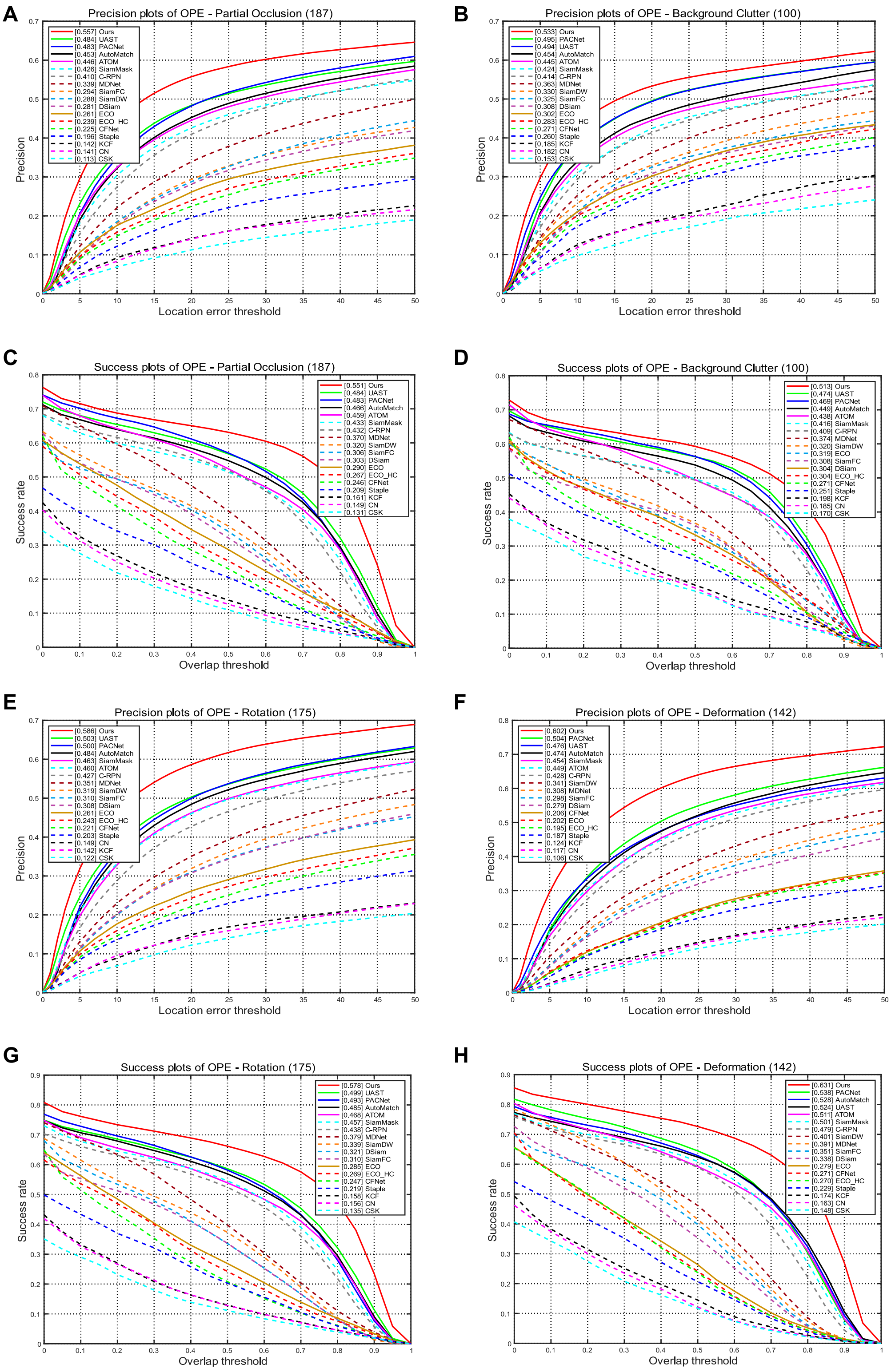


FIGURE 6
Some of the challenges tested on the LaSOT dataset include partial occlusion, similar appearance, deformation, and rotation. **(A, B, E, F)** Results of accuracy evaluation; **(C, D, G, H)** results of success rate evaluation.

and for the GOT-10k benchmark test, the number of samples per epoch is set to 150,000, the mini-batch size is set to 20, the momentum decay rate is 0.9, the weight decay rate is 1×10^{-4}, the initial learning rate is set to 1×10^{-2} to 1×10^{-8} in the first epoch when the backbone network is trained, and in the other epochs, the learning rate increases from 8×10^{-2} to 1×10^{-6}. In the online

TABLE 3 TrackingNet dataset test results, with trackers ranked from top to bottom according to "success" values, where "Norm. Prec." is an abbreviation for normalized precision.

Tracker	Success	Precision	Norm. Prec.
ATOM [30]	0.703	0.648	0.771
SiamRPN++	0.733	0.649	0.800
SiamAttn	0.752	-	0.81
SiamFC++ [31]	0.754	0.705	0.800
AutoMatch [32]	0.760	0.726	-
TrSiam [33]	0.781	0.721	0.829
TrDiMP	0.784	0.731	0.833
Ours	**0.788**	**0.754**	**0.836**

The meaning of the bold values in table primarily represents the training results of the methods proposed in this article.

tracking process, the input memory frame template image size is 289 × 289 × 3, the search image is 289 × 289 × 3, where the memory frame is stored in T = 3, and the image pairs are entered into the respective network branches to finally obtain a score map of size 25. Furthermore, in the hyperparameter settings, λ1 and λ2 are the two hyperparameters that control the weights between the three losses, and we set them to the default value 1.

4.2 Ablation experiments

To verify the effectiveness of the proposed spatio-temporal interactive fusion-based object tracking method using the spatio-temporal feature enhancement module in the memory branch, only spatial dimension enhancement (SDE) in the query branch, and the spatio-temporal interaction model (STIF), we performed a series of ablation experiments on the GOT-10k [17] dataset, and it should be noted that the baseline tracker we compared were mainly STMTrack tracker. The specific experiments are as follows: 1) adding STE, SDE, and STIF separately to the overall network framework to compare with the baseline network; 2) adding STE and SDE simultaneously to the network framework (i.e., using the feature enhancement network as a whole) to compare with STE and SDE alone; 3) adding STE, SDE, and STIF simultaneously to the network framework to compare with the training performance when each module is added separately.

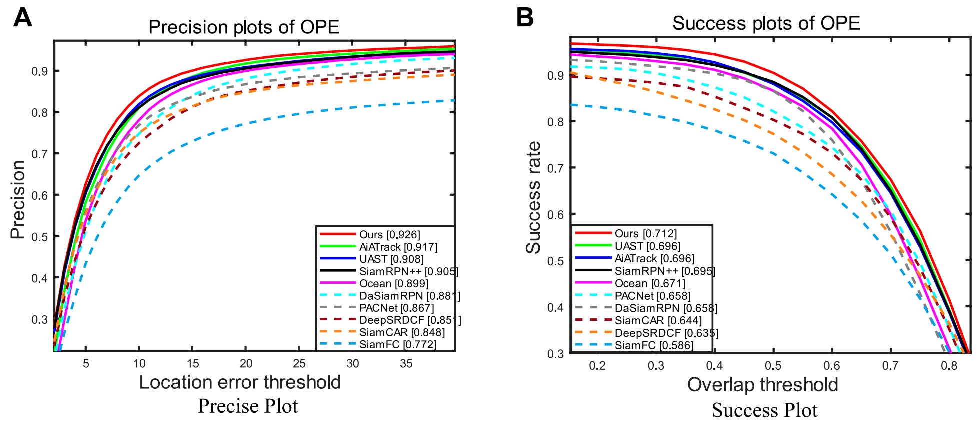
The results of the ablation experiments are shown in Table 1, where AO represents the average overlap rate and SR represents the success rate. From Table 1, for AO, adding each network individually improves the value. While the overall usage of the feature enhancement network improves by approximately 2% compared to adding STE and SDE networks alone, adding STE, SDE, and STIF simultaneously to the network framework achieves the best performance, improving the value of AO by 10% compared to the baseline network. In addition, the results in Table 1 show that the STE, SDE, and STIF modules achieve more significant improvements in the tracking success rate metrics SR0.5 and SR0.75. The results of the ablation experiments sufficiently demonstrate the effectiveness of the feature enhancement module and the interaction fusion module designed in this paper.

4.3 Comparison experiments

In order to evaluate the performance of our method, this section compares the performance of the proposed tracking algorithm with the current mainstream algorithms on the datasets of GOT-10k [17], LaSOT [16], TrackingNet [15], OTB100 [20],and WATB [21].

4.3.1 Evaluation on the GOT-10k dataset

To evaluate the generalization ability of our method, we choose the GOT-10k dataset, which is a large-scale generic benchmark dataset with more than 10,000 videos of real-life scenarios and a test set with 180 video sequences. A key feature of GOT-10k is that there is no overlap between the classes of tracked objects in the training and test sets, which can be used to evaluate the generalization ability of the tracker. To ensure a fair comparison, this paper follows the GOT-10k testing protocol, and only the GOT-10k training set is used to train the tracker.



Precise Plot

Success Plot

FIGURE 7
Test results of our method on the OTB100 dataset. **(A)** Precise plot and **(B)** success plot.

TABLE 4 Video frame challenge attributes for the selected OTB100 dataset.

Sequence name	Frame number	Challenge attribute
Soccer	35,165,215	IV, SV, OCC, MB, FM, IPR, OPR, BC
Toy	171,216,271	IV, IPR, OPR
Jogging	18,56,84	OCC, DEF, OPR
Skating1	103,182,253	OCC, DEF, OPR

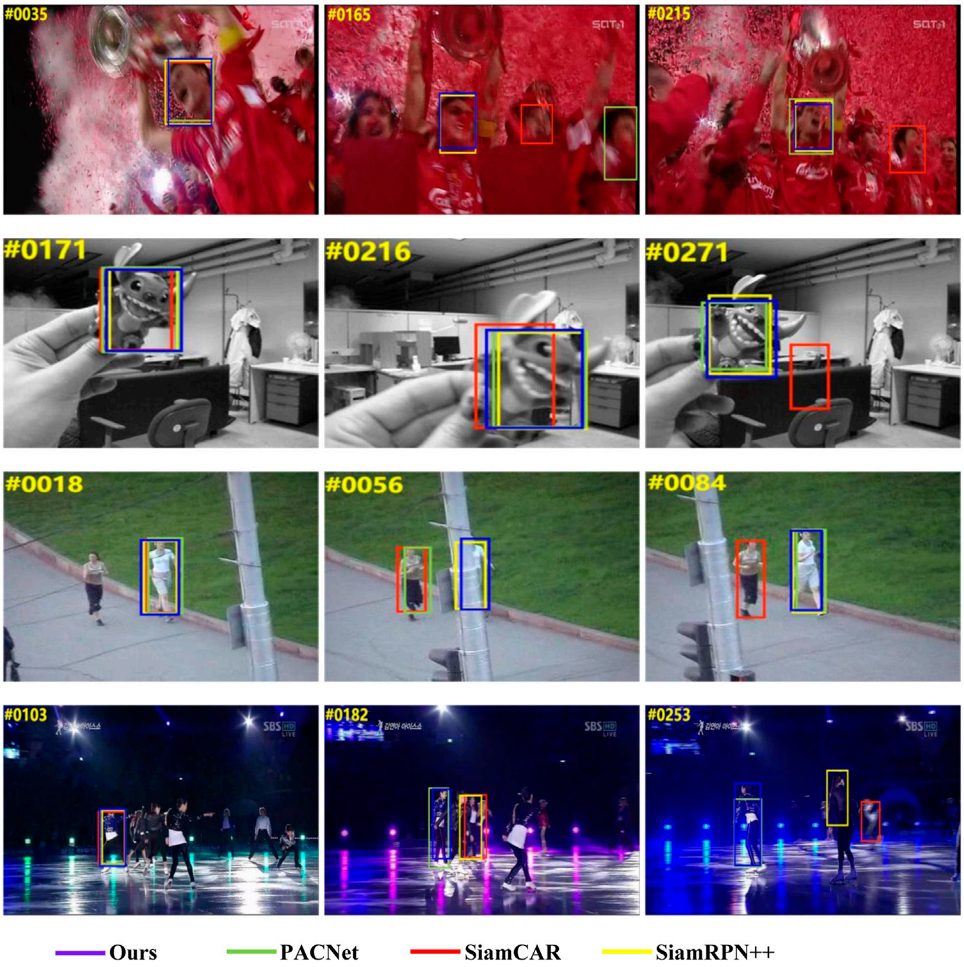


FIGURE 8
Comparison of our proposed tracker with some advanced trackers on four challenging OTB100 video sequences; the results show that our approach can effectively address these challenges. Reproduced from the OTB100 dataset, [20], with permission from IEEE.

As shown in Table 2, by comparing with some existing advanced trackers, including UAST [24], STMTrack [9], and SiamGAT [25], the algorithm proposed in this paper has the best results, where AO is 65.2%, $SR_{0.5}$ is 78.1%, and $SR_{0.75}$ is 59.0%. The method in this paper mainly makes full use of the spatio-temporal context information and can improve the generalization ability of the model, thus outperforming these advanced trackers in terms of performance.

4.3.2 Evaluation on the LaSOT dataset

To evaluate the adaptability of our method in the face of different scenario variations, we tested it using the LaSOT dataset. LaSOT is also a large-scale single-object tracking dataset with high-quality annotations. Its test set consists of 280 long videos, with an average of 2,500 frames each, including 70 tracking object classes, each containing 20 tracking sequences, containing many video sequences with different attributes, covering a wide range of scenarios at multiple scales, speeds, backgrounds, and poses. LaSOT analyzes the performance of each algorithm mainly using accuracy maps based on position error metrics and success maps based on overlap metrics.

The success plot is shown in Figure 4, and the accuracy plot is shown in Figure 5. Compared with a variety of comparable trackers,

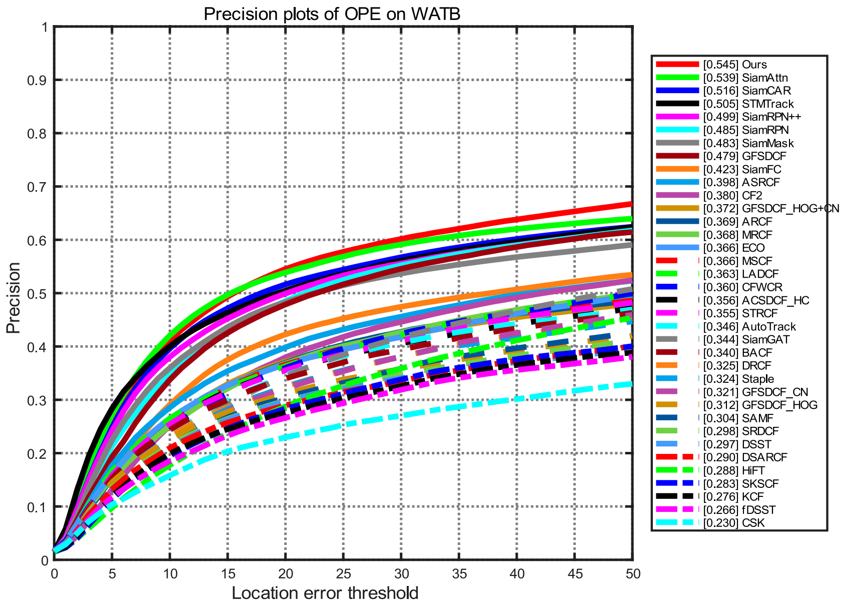


FIGURE 9
Precision plot of our method on the WATB dataset.

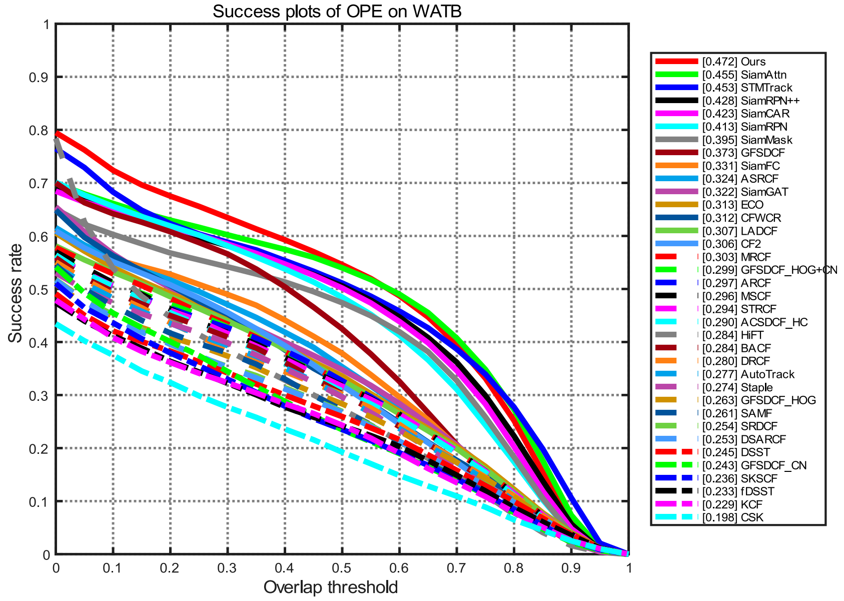


FIGURE 10
Success plot of our method on the WATB dataset.

our tracker reaches the forefront in terms of success, accuracy, and standardized precision, including 6.7% and 6.8% improvement in the success rate and 6.8% and 7.6% improvement in accuracy compared to PACNet [29] and UAST [24], respectively.

In addition, for some complex scenario challenges, such as the similar appearance as the object (BC), the object is deformable during tracking (DEF); the object rotates in the image (ROT), and the object is partially occluded in the sequence (POC). We also report results on the LaSOT [16] dataset. The results shown in Figure 6 demonstrate that our method exhibits optimal performance when facing the aforementioned challenges. This indicates that our model is capable of effectively adapting to variations in the images without easily experiencing drifting, showcasing strong robustness.

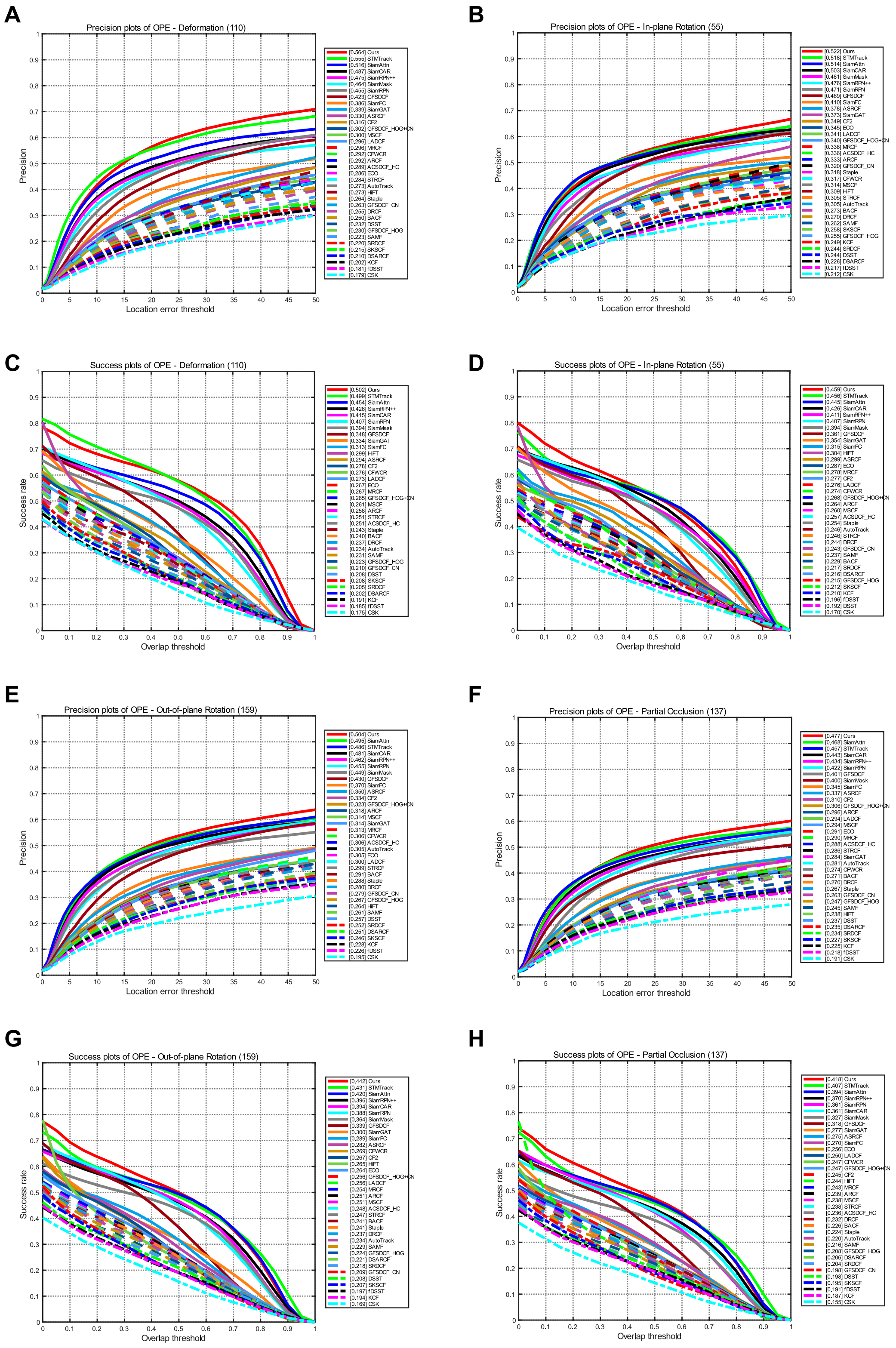


FIGURE 11
Some of the challenges tested on the WATB dataset include out-of-plane rotation, in-plane rotation, deformation, and partial occlusion. **(A, B, E, F)** Results of precision plots and **(C, D, G, H)** results of success plots.

4.3.3 Evaluation on the TrackingNet dataset

TrackingNet is a large-scale short-term tracking dataset that provides many field videos for training and testing. The test set contains 511 video sequences. Evaluation on the TrackingNet dataset enables testing of the model's tracking performance in a variety of scenarios, including the effectiveness of tracking different objects, adaptability to different locations, scales, lighting, and background changes, and evaluation of various metrics, such as the model's recognition and tracking accuracy for many different attributes.

We evaluate the tracker on the test set and obtain the results from a dedicated evaluation server. As shown in Table 3, our tracker outperforms some other advanced tracking algorithms to a large extent exhibiting better performance.

4.3.4 Evaluation on the OTB100 dataset

OTB100 is a classical benchmark for visual object tracking and contains 100 short-term videos, with an average of 590 frames per video. The results of our proposed algorithm against other algorithms on the OTB100 dataset are shown in Figure 7, and the results demonstrate that our method outperforms other algorithms in terms of success rate metrics (AUC) as well as accuracy. In addition, to be able to visualize the actual tracking effect of the object tracking algorithm, we selected four representative video frames from the 100 videos of the OTB100 dataset for visualization, which contain most of the challenges encountered in object tracking scenarios. As shown in Table 4, these four sequences selected include the following challenges: scale variation, deformation, object partial occlusion, and background interference. By using different colored object tracking frames to represent the tracking effect of the algorithm in this paper and other compared mainstream object tracking algorithms (SiamRPN++, SiamCAR, and PACNet) in the same image (Figure 8), it is possible to analyze more intuitively these algorithms in the same image (Figure 8), allowing a more visual analysis of the tracking performance of these algorithms. From the visualization results, the algorithm proposed in this paper shows long-term stability in tracking sequences under complex challenges. The results of the first line and the fourth line of the video sequences are mainly related to the challenges of cluttered backgrounds as well as scale variations, which shows that our tracker maintains a high level of robustness, while the other three methods drift during the tracking process. Tests from the second row of video sequences show that our method exhibits stable and consistent tracking performance when the object undergoes rotational changes. The information obtained from the third row of video sequences shows that our model can accurately track the object even in the presence of object occlusions and recognize the object accurately when it reappears. By visualizing the tracking process as described previously, we again validate the high tracking performance of our model.

4.3.5 Evaluation on the WATB dataset

The WATB dataset is a common wildlife dataset containing over 203,000 frames and 206 video sequences covering a wide range of animals on land, in the ocean, and in the sky. The average length of the videos was over 980 frames. Each video was manually labeled with 13 challenge attributes, including partial occlusion, rotation, and deformation. All frames in the dataset were labeled with axis-aligned bounding boxes. To test the performance of our model on this dataset, we tested our method using WATB, and the tracking accuracy and success plots are shown in Figure 9 and Figure 10, respectively. According to the results of the tracking success rate curve and precision rate curve, our method outperforms other algorithms and achieves optimal performance when compared to other tracking methods. In addition, we also tested the performance against some challenging attributes, including rotation, deformation, partial occlusion, and scale variation. According to the results shown in Figure 11, our method shows continuous better performance among the compared trackers.

5 Conclusion

In this paper, a novel tracking framework based on the spatio-temporal interactive fusion network is proposed. Considering the spatio-temporal correlation between frames, a feature enhancement network is used to process both memory and query branches by combining historical frame information, and a spatio-temporal interactive fusion network is proposed to achieve effective filtering and fusion of feature information of the two branches, which improves the generalization ability of the network and makes full use of contextual information. In the feature enhancement network, by introducing a spatio-temporal feature enhancement network, the memory frame features are enhanced in the temporal dimension as well as the spatial dimension, and the query frame features are enhanced only in the spatial dimension, enabling the tracker to locate the object more accurately. The method proposed in this paper can cope with most complex situations, but the problem of object loss for small objects and for situations, where there are more similar objects interfering in the background, still exists, while the method can be improved in other ways. Overall, through extensive experimental results on the GOT-10k, OTB100, TrackingNet, LaSOT, and WATB datasets, the tracking method proposed in this paper shows better performance.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

DH: conceptualization and writing–original draft. SY: investigation, software, visualization, writing–original draft, and writing–review and editing. JD: resources and writing–review and editing. YW: investigation, methodology, and writing–original draft.

AY: project administration, resources, and writing–original draft. YW: software and writing–review and editing. JX: visualization and writing–original draft.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Natural Science Foundation of China (grant number 62127813) and the Science and Technology Key Research and Development Project of Jilin Province (grant number 20230201071GX).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Wang Y, Wang F, Wang C, Sun F, He J. Learning saliency-aware correlation filters for visual tracking. *Comp J* (2022) 65(7):1846–59. doi:10.1093/comjnl/bxab026

2. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PHS. Fully convolutional siamese networks for object tracking. Proceedings of the computer vision – ECCV 2016 workshops, Amsterdam, Netherlands, October 2016.

3. Zhang L, Gonzalez-Garcia A, Weijer J, Learning the model update for siamese trackers, Proceedings of the IEEE/CVF international conference on computer vision. Seoul, Korea (South), October 2019: 4010–9.

4. Yu Y, Xiong Y, Huang W, Deformable siamese attention networks for visual object tracking, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle, WA, USA, June 2020: 6728–37.

5. Xu Y, Li T, Zhu B, Siamese tracking network with multi-attention mechanism[J] (2023). doi:10.21203/rs.3.rs-3296460/v1

6. Sun F, Zhao T, Zhu B, Jia X, Wang F. Deblurring transformer tracking with conditional cross-attention. *Multimedia Syst* (2023) 29(3):1131–44. doi:10.1007/s00530-022-01043-0

7. Guo Q, Feng W, Zhou C Learning dynamic siamese network for visual object tracking, Proceedings of the IEEE international conference on computer vision. Venice, Italy, October 2017: 1763–71.

8. Yan B, Peng H, Fu J, Wang D, Lu H, et al. Learning spatio-temporal transformer for visual tracking, Proceedings of the IEEE/CVF international conference on computer vision. Montreal, BC, Canada, October 2021: 10448–57.

9. Fu Z, Liu Q, Fu Z Stmtrack: template-free visual tracking with space-time memory networks, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville, TN, USA, June 2021: 13774–83.

10. Li D, Wen G, Kuai Y, Porikli F. End-to-end feature integration for correlation filter tracking with channel attention. *IEEE Signal Process. Lett* (2018) 25(12):1815–9. doi:10.1109/lsp.2018.2877008

11. Rahman MM, Fiaz M, Jung SK. Efficient visual tracking with stacked channel-spatial attention learning. *IEEE Access* (2020) 8:100857–69. doi:10.1109/access.2020.2997917

12. He A, Luo C, Tian X, A twofold siamese network for real-time object tracking, Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA, June 2018: 4834–43.

13. Tian Z, Shen C, Chen H, Fcos: fully convolutional one-stage object detection, Proceedings of the IEEE/CVF international conference on computer vision. Seoul, Korea (South), October 2019: 9627–36.

14. Lin TY, Goyal P, Girshick R, Focal loss for dense object detection, Proceedings of the IEEE international conference on computer vision. Venice, Italy, October 2017: 2980–8.

15. Muller M, Bibi A, Giancola S, Trackingnet: a large-scale dataset and benchmark for object tracking in the wild, Proceedings of the European conference on computer vision (ECCV). Munich, Germany, September 2018: 300–17.

16. Fan H, Lin L, Yang F, Lasot: a high-quality benchmark for large-scale single object tracking, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach, CA, USA, June 2019: 5374–83.

17. Huang L, Zhao X, Huang K. GOT-10k: a large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans pattern Anal machine intelligence* (2019) 43(5):1562–77. doi:10.1109/tpami.2019.2957464

18. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis* (2015) 115:211–52. doi:10.1007/s11263-015-0816-y

19. Lin TY, Maire M, Belongie S, Microsoft coco: common objects in context, Proceedings of the computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, September, 2014: 740–55.

20. Wu Y, Lim J, Yang MH. Object tracking benchmark. *IEEE Trans pattern Anal machine intelligence* (2015) 37(9):1834–48. doi:10.1109/tpami.2014.2388226

21. Wang F, Cao P, Li F, Wang X, He B, Sun F. WATB: wild animal tracking benchmark. *Int J Comp Vis* (2023) 131(4):899–917. doi:10.1007/s11263-022-01732-3

22. Paszke A, Gross S, Chintala S, Automatic differentiation in pytorch (2017). https://openreview.net/pdf?id=BJJsrmfCZ#:~:text=PyTorch%2C%20like%20most%20other%20deep,which%20usually%20differentiate%20a%20single.

23. do Carmo França HF, Soares A. *GoogLeNet-going deeper with convolutions* (2014). https://arxiv.org/abs/1409.4842.

24. 24Zhang D, Fu Y, Zheng Z. UAST: uncertainty-aware siamese tracking, Proceedings of the international conference on machine learning. PMLR, Baltimore, Maryland, USA, July 2022.

25. Guo D, Shao Y, Cui Y, Graph attention tracking, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville, TN, USA, June 2021: 9543–52.

26. Li B, Wu W, Wang Q, Siamrpn++: evolution of siamese visual tracking with very deep networks, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach, CA, USA, June 2019.

27. Bhat G, Danelljan M, Gool LV, Learning discriminative model prediction for tracking, Proceedings of the IEEE/CVF international conference on computer vision. Seoul, Korea (South), October 2019: 6182–91.

28. Zhang Z, Peng H, Fu J, Ocean: object-aware anchor-free tracking, Proceedings of the computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August, 2020.

29. Zhang D, Zheng Z, Jia R, Li M. Visual tracking via hierarchical deep reinforcement learning, Proceedings of the AAAI Conf Artif Intelligence. Washington DC, USA, February, 2021, doi:10.1609/aaai.v35i4.16443

30. Danelljan M, Bhat G, Khan FS, Atom: accurate tracking by overlap maximization, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach, CA, USA, June 2019: 4660–9.

31. Xu Y, Li T, Zhu B, Wang F, Sun F, Siamfc++: towards robust and accurate visual tracking with object estimation guidelines, Proceedings of the AAAI conference on artificial intelligence. New York, NY, USA, February 2020: 12549–56.

32. Zhang Z, Liu Y, Wang X, Learn to match: automatic matching network design for visual tracking, Proceedings of the IEEE/CVF international conference on computer vision, Montreal, BC, Canada, October 2021: 13339–48.

33. Wang N, Zhou W, Wang J, Transformer meets tracker: exploiting temporal context for robust visual tracking, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville, TN, USA, June 2021: 1571–80.