Memory Propagation-based Target-aware Segmentation Tracker with Adaptive Mask-attention Decision Network

Huanlong Zhang[1*], Weiqiang Fu[1], Bin Zhou[1], Keyan Zhou[1], Xiangbo Yang[1], and Shanfeng Liu[2]
[1] College of Electric and Information Engineering, Zhengzhou University of Light Industry
Zhengzhou, 450000, China
[e-mail: zhl_lit@163.com]
[2] State Grid Henan Electric Power Research Institute
Zhengzhou, 450000, China
[e-mail: 179555872@qq.com]
[*]Corresponding author: Huanlong Zhang

Received November 12, 2023; revised July 7, 2024; revised July 28, 2024;
accepted August 8, 2024; published September 30, 2024

Abstract

Siamese-based segmentation and tracking algorithms improve accuracy and stability for video object segmentation and tracking tasks simultaneously. Although effective, variability in target appearance and background clutter can still affect segmentation accuracy and further influence the performance of tracking. In this paper, we present a memory propagation-based target-aware and mask-attention decision network for robust object segmentation and tracking. Firstly, a mask propagation-based attention module (MPAM) is constructed to explore the inherent correlation among image frames, which can mine mask information of the historical frames. By retrieving a memory bank (MB) that stores features and binary masks of historical frames, target attention maps are generated to highlight the target region on backbone features, thus suppressing the adverse effects of background clutter. Secondly, an attention refinement pathway (ARP) is designed to further refine the segmentation profile in the process of mask generation. A lightweight attention mechanism is introduced to calculate the weight of low-level features, paying more attention to low-level features sensitive to edge detail so as to obtain segmentation results. Finally, a mask fusion mechanism (MFM) is proposed to enhance the accuracy of the mask. By utilizing a mask quality assessment decision network, the corresponding quality scores of the “initial mask” and the “previous mask” can be obtained adaptively, thus achieving the assignment of weights and the fusion of masks. Therefore, the final mask enjoys higher accuracy and stability. Experimental results on multiple benchmarks demonstrate that our algorithm performs outstanding performance in a variety of challenging tracking tasks.

Keywords: Object Segmentation and Tracking, Siamese Network, Mask Propagation, Mask evaluation, Attention Mechanism.

http://doi.org/10.3837/tiis.2024.09.008 **ISSN : 1976-7277**

1. Introduction

Video object tracking (VOT), a fundamental but challenging task in computer vision, typically aims to locate and track an object of interest across frames in a video sequence, using a bounding box in the first frame. In a broad extent of applications, like human-computer interaction [1], automatic vehicles [2], traffic management [3], and video surveillance [4], object tracking is inevitable. Given the unknown and frequent changes in objects and their surroundings, developing a tracker that can effectively handle changes in target appearance, eliminate background clutter, and maintain real-time tracking is a highly challenging task in computer vision.

Various tracking methods based on deep convolutional neural networks have been proposed in recent years. One mainstream methodology, the template-matching method, addresses VOT as a similarity-matching problem between an initial target template image and the search images. Siamese trackers [5, 6, 7, 8, 9] are among the most commonly used template-matching methods. Since these approaches typically fail to update the template, they are poorly capable of resisting changes in target appearance resulting from factors like occlusions, non-rigid deformations, etc. In addition, these trackers adopt box representation, and much background information is introduced by the predefined spatial limit of the box representation.

To address the issues mentioned above, although some trackers [10, 11] have implemented advanced template updating mechanisms to enhance their robustness, these methods can be computationally expensive and may hinder real-time tracking. Additionally, the customized updating strategies [12, 13, 14] used by these trackers introduce hyperparameters that require tricky tuning, which can be a challenging task. Meanwhile, in order to suppress background interference, some approaches [15, 16] are equipped with temporal context before enhancing the target feature and inhibiting background information. Some Siamese trackers [17, 18, 19, 20] utilize the characteristic of cosine windows, which emphasize the center area and reduce the impact of the target on the image edge. The above method can help reduce boundary effects but is unable to make full use of the target information and further improve tracking performance.

To take full advantage of the information of the target, we introduce the binary mask of the target and apply the segmentation algorithm to the tracking. The video object segmentation (VOS) task is intended to estimate the segmentation of specific object instances in the video sequence. The VOS focuses on analyzing large objects that have been observed within 100 frames, which produces pixel-level segmentation results $i.e.$ binary masks. In essence, VOS is capable of offering a more accurate assessment of the state of an object. The solution to the VOS task can be applied to the VOT task due to its detailed representation.

Recent research has focused on bridging the gap between VOT and VOS through the development of approaches that integrate segmentation algorithms into tracking. Semantic-aware tracker [21] offers a federated approach to integrate semantics into tracking. A pre-existing image segmentation network, such as Box2Seg [22], serves to transform the semantic information within the target bounding box into a pixel-level semantic segmentation result. The bounding box is generated by a tracking model [23], which can become less accurate as errors accumulate, leading to incorrect segmentation. Additionally, running the approach in real-time is challenging as the image segmentation network requires significant computation. To reduce computational costs, SiamMask [5] and D3S [24] provide a combined segmentation and tracking network, containing a segmentation branch to binary a mask by refining the features resulting from the backbone network. Nevertheless, the methodologies remain contingent upon the initial representation of the object and fail to account for the annotated mask. Hence, they may prove inadequate in addressing tracking challenges such as appearance variations and occlusions across the video sequence. This motivates us to explore a mask propagation-based tracking algorithm that fully mines the rich historical mask information available of the object. Moreover, background interference beyond the bounding boxes possibly leads to inaccurate segmentation results without spatial constraints, causing the network not to distinguish the target from background clutter with similar semantics.

In this work, we propose a mask propagation-based attention method to learn to read relevant information from historical frames and constrain the background interference in backbone features. In particular, the mask propagation-based attention module (MPAM) produces target attention maps by searching for a memory bank (MB), which stores features (*addresses*) and binary masks (*values*) of the first frame and the previous frame. We can reduce the negative effects of background noise by utilizing the attention maps to enhance the features of t-th frame**.** In addition, we design an attention refinement pathway (ARP) to further refine the segmentation profile and weaken the effect of background interference. It merges different resolution features with attention refinement modules to generate the corresponding mask. To further enhance the accuracy of the mask, we propose a mask fusion mechanism (MFM) to evaluate the importance of different masks. It can obtain the corresponding mask quality score by a quality evaluation net, which is used as the fusion ratio for generating a reliable mask.

Summarily, four main contributions are made by this work.

• We construct a mask propagation-based attention module (MPAM) for exploring the inherent correlation among image frames, which can mine relevant mask information from historical frames. Target attention maps are produced to highlight the target region on backbone features, thus relieving the adverse effects of background clutter.

• An attention refinement pathway (ARP) is designed to further refine the segmentation profile in the process of mask generation. It pays more attention to the low-level features sensitive to edge detail and enhances target characterization.

• To obtain a more accurate mask, the mask fusion mechanism (MFM) is proposed. By using a quality assessment network, corresponding quality scores of the “initial mask” and the “previous mask” can be obtained to distinguish their significance, thus achieving the assignment of weights and the fusion of masks. Benefiting from this approach, the final mask enjoys higher accuracy and stability.

• Comprehensive analysis and experiments on OTB100, VOT2016, TC128, and UAV123 tracking benchmarks show that our method performs outstanding performance in a variety of challenging tracking tasks.

2. Related Work

2.1 Siamese Network-based Tracking

Recently, the Siamese network-based tracking algorithm offers an excellent balance between efficiency and performance [25]. There are two inputs in a Siamese tracker, a target template at the first frame, and a search region within subsequent frames. The tracker aims to learn a similarity mapping of the two inputs by localizing the target template in the search regions [9]. SiamFC [26] applies a pre-trained backbone network to maximize object-background discrimination by cross-correlation operations between the target templates and the search regions. Following the idea of SiamFC, several multi-stage Siamese extensions have been developed to enhance tracking performance. The Siamese Region Proposal Network

(SiamRPN) [17] utilizes region proposal techniques to improve the efficiency and accuracy of object tracking. Based on SiamRPN, the work of [27] employs a distractor-aware module, and data augmentation to suppress the influence of distractors on the tracking process and use a local-to-global strategy to redetect the target for long-term tracking in DaSiamRPN. As an improved version of the original SiamRPN algorithm for visual object tracking, the main improvement of SiamRPN++ [7] is the utilization of a deeper and more complex network architecture, which includes multiple residual blocks and a spatial-aware attention mechanism to better capture the spatial context of the target object. SiamMask [5] is the first algorithm to introduce segmentation to the tracker. It adds a mask branch to SiamRPN to predict the object mask in real time and generates a box from the segmentation result. Considering that only a single-channel response map lacking correlation features is generated in [26], SiamCAR [9] generates a multi-channel response map using a depth-wise correlation layer, which consists of multiple single-channel response maps folded along the channel dimension. Another architecture very similar to SiamCAR, SiamBAN [6] takes a unified FCN classification and regression bounding box to accurately estimate the scale and aspect ratio of the target. These Siamese networks are poorly adapted to changes in target appearance caused by occlusions or non-rigid deformations since they usually do not update the template. Moreover, the box representation adopted by these trackers introduces a lot of background information due to the predefined spatial limit, which makes visual tracking very difficult because of the presence of distractors and frequent changes in appearance. Our algorithm solves these problems by introducing the mask information of historical frames into the Siamese network.

2.2 Segmentation-based Tracking

In Semi-supervised video object segmentation (SVOS), the initial frame provides ground truth annotations, which are used to identify the objects to be automatically segmented from the subsequent frames. The representative SVOS approaches, namely matching-based methods can be divided into three categories: pixel-level matching [28, 29, 30] and region-of-interest matching [31, 32, 33], and mask propagation-based techniques [34, 35, 36], which explore the inherent correlation among image frames by propagating the segmented masks of the previous frames to the subsequent frame.

Video object segmentation and tracking can be integrated to effectively address their tasks at the same time and improve the accuracy and stability of the tracking process. Recently, several researchers have combined tracking and semi-supervised video object segmentation using offline and online CNN-based methods [5, 37] and demonstrated excellent performance in their results. In the work of [5], a Siamese network is proposed to predict bounding boxes, object scores, and binary masks simultaneously. Based on matching features extracted from the first frame, D3S [24] conducts segmentation, which improves segmentation accuracy and speeds up processing. Nevertheless, these methods still use only the first frame as a template to create a model, resulting in inadequate performance when faced with difficulties such as changes in appearance. In this study, we leverage a propagation method and acquire the ability to extract pertinent information from past frames.

2.3 Attention Mechanism

As deep learning advances, the applications of attention mechanisms are becoming more and more diverse, with crucial implications for human perception [38]. Some works use the attention mechanism for classification [39], detection [40], tracking [41], and segmentation [42]. By integrating spatial and channel attention sequentially, CBAM [43] determines where and what to focus on according to cross-channel and spatial correlations. A novel non-local network is presented in [44] which incorporates self-attention into computer vision, resulting in significant success in object detection and video understanding. In object tracking, SiamAtt [41] applies self-attention mechanisms to improve the target estimation based on a weight fusion of the classification and attention branch scores. However, the method has difficulty dealing with frequent changes in the appearance of the target. LANet [42] proposes two modules for enhancing the representation of features based on the exploitation of local attention, bridging the high-level and low-level features gap. Xiao et al. [45] proposed an online updated Siamese-block attention network for augmenting target representations, addressing the problem of occlusion and change in tracking. Several attention mechanisms can also be applied to segmentation, such as OCNet [46], using a self-attention-affected object context pooling mechanism, Expectation-Maximization Attention (EMANet) [47], Split-Attention Networks (ResNet) [48], and Height-driven Attention Networks [49].

Most attention modules currently generate weight maps using a convolution layer or self-attention mechanism that solely relies on the feature itself, which does not take into account the specific target being attended to. Conversely, we incorporate both visual features and pixel-wise mask information and diffuse temporal information of the target, resulting in a weight map that is more specific and focused on the object of interest.

3. Method Description

This section provides a comprehensive overview of our proposed segmentation tracker. The section is subdivided into four parts, beginning with a general introduction to the motivation and architecture of the network. Next, we describe the implementation of our mask propagation-based attention module (MPAM), followed by a detailed explanation of the concatenate attention refinement pathway (ARP). Finally, we discuss the mask fusion mechanism (MFM), which is used to generate a reliable mask and obtain a bounding box. The complete tracking process is depicted in **Fig. 1**.

3.1 Framework Overview

The framework we propose is specifically designed to tackle complex segmentation and tracking challenges. It comprises three main components: a mask propagation-based attention module (MPAM), an attention refinement pathway (ARP), and a mask fusion mechanism (MFM). To begin with, our framework utilizes the Siamese network SiamMask [5] for object tracking, which is a well-established technique in the field. As is typical in this type of network, the template z with the size of 127×127 is a cropped image centered in the target of the first frame and the search region x with the size of 255×255 is another larger cropped image centered in the predicted target location in the previous frame. Then, a dense response map is calculated through a depth-wise crossing-correlation operation of two extracted features. In addition to establishing two branch networks to compute the classification score and location regression for a set of k anchor boxes encoded in a region, a mask branch is proposed to form a two-task network for both segmentation and tracking in SiamMask. It is a three-branch network to generate the classification score, location regression, and binary mask. The anchor box with the highest score will be selected to form a new box through its corresponding mask as the tracking result.

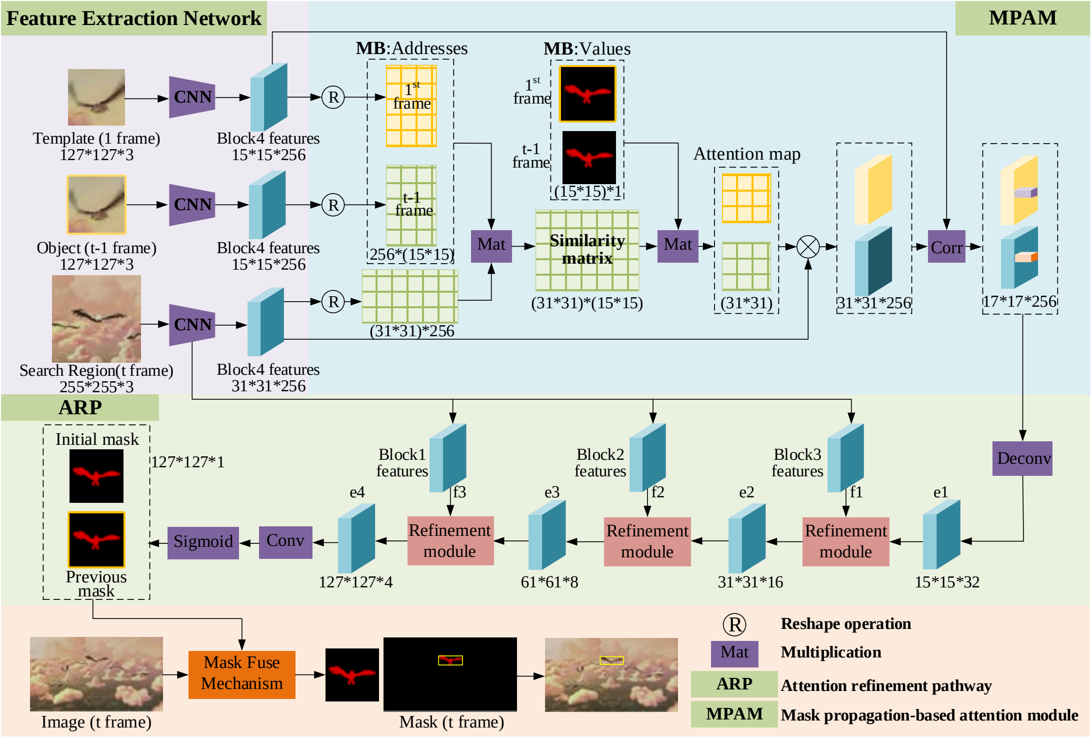


Fig. 1. Architecture of the proposed network.

Based on SiamMask, in the first stage, the block4 feature (as shown in **Fig. 1**) extracted by the backbone network is fed to the mask propagation-based attention module (MPAM) for generating attention maps with target awareness (see Sect.3.2). Next, the block4 feature is heightened by the attention maps and used to perform depth correlation operation with template features to a reliable dense response. Then the second stage, a $1 \times 1 \times 256$ response bar is employed as the input in the attention refinement pathway (ARP). The ARP (see Sect.3.3) uses tree attention refinement modules consisting of upsampling layers and skip connections to produce a more accurate segmentation mask. In MPAM, since employing the method of mask propagation, we can get two masks after the above two stages, which are input into the mask fusion mechanism (MFM) intending to get a robust mask (see Sect.3.4). Considering the second strategy ($i.e.$ rotated minimum bounding rectangle) in SiamMask, a robust binary mask can be used to generate a bounding box as the tracking result.

3.2 Mask Propagation-based Attention Module (MPAM)

Only a bounding box of the target is needed to be input in the first frame in SiamMask and the segmentation can be done for the following frames. It does not consider using the predicted mask information in the previous frames but only uses the bounding box in the first frame as the template. Therefore, we consider a mask propagation method to better adapt to target changes during tracking. Meanwhile, the attention maps are generated to constrain the background clutter in the backbone network features, guiding the tracker to concentrate on the most informative part of the current frame. We first detail how to build the history frame mask set ($i.e.$ Memory Bank), and then demonstrate how the Memory Bank (MB) can be used to predict attention maps and how to conduct the history information to the current frame.

3.2.1 Build Memory Bank (MB)

The MB can be viewed as an ordered set that consists of *address*-*value* pairs, where each *address* corresponds to a unique *value*. In other words, each element in the set is a tuple (*address*, *value*), where the *address* is reference features $A \in R^{H \times W \times C}$ stored in a memory bank and the *value* is corresponding pixel-wise foreground probabilities (binary masks) $V \in R^{H \times W \times 1}$, where H is the height, W is the width, and C is the feature dimension. To ensure that our module can run efficiently with limited memory resources, we avoid storing intermediate features and results. The first frame and the previous frame have a relatively large impact on the current frame. The target mask of the first frame can extract the most reliable appearance and position information, and the mask of the previous frame has the motion state of the target. We simply construct the MB with the first frame and the previous frame ($i.e.$ $t-1$ frame). Therefore, the mask generated in the current frame usually has higher accuracy and can better represent the location and shape of the target.

Specifically, based on the ground truth bounding box presented in the first frame, a feature map of size $H_1 \times W_1 \times C$ is obtained by passing the template $127 \times 127 \times 3$ cropped in the first frame through a backbone network and a 1×1 convolution layer. In our module, $H_1 = W_1 = 15$, $C = 256$. Here, subscript 1 denotes the “first frame”. Given that the first frame of a video does not have an annotated mask, we first input the template and the search region cropped on the first frame, which is input into the original SiamMask algorithm to get the predicted mask as the pseudo annotated mask. The corresponding mask has the same spatial size $H_1 \times W_1 \times 1$. At this point, the size of the MB equals $H_1 \times W_1$. Again, by executing our proposed framework, we can get the object feature and binary mask for each frame with the same size as the first frame, $H_{t-1} = W_{t-1} = 15$, $C = 256$. The second *address*-*value* pair of MB is dynamically updated with the $t-1$ frame object feature and predicted mask.

3.2.2 Attention Maps

To generate the attention map of a feature, the MB is queried with the backbone network output feature of the current frame. Concretely, in the t -th frame, the search region $255 \times 255 \times 3$ cropped in the t -th frame is passed into the backbone network next to a 1×1 convolution layer, getting feature map $F_t \in R^{H_t \times W_t \times C}$. Now, $H_t = W_t = 31$, $C = 256$. Here, subscript t denotes the “t-th frame”. Based on the transductive model [36], an effective pixel-based similarity measure is required to be learned online in a video sequence. In the feature map F_t , an attention value is attached to each spatial pixel $f_i \in R^{1 \times 1 \times 256}$, which is decided by its similarity to the *addresses* in the MB. This similarity is calculated using the dot product operation between the embedding and each *address*.

$$s_{ij} = f_i \cdot a_j \tag{1}$$

Where $\cdot$ represents dot product, s_{ij} indicates the similarity between f_i and a_j , $f_i \in F_t$ and $a_j \in A$ are the spatial embedding of the feature map F_t and an address in MB. As a result, all embeddings of F_t are similar to one of the addresses in MB, so

$$S = softmax(R(F_t) * R(A_1)) \tag{2}$$

Where $*$ indicates tensor multiplication and R represents the reshape operation. The shape of S, $R(F_t)$ and $R(A_1)$ is $H_tW_t \times H_1W_1$, $H_tW_t \times C$, $C \times H_1W_1$, respectively. The resulting matrix S is used to generate the attention map M, as shown in (3).

$$M = S * V_1 \tag{3}$$

The M quantifies the likelihood that a specific spatial location in F_t will be the target of attention. Afterwards, an element-wise product is performed between F_t and M to enhance the feature, which highlights the most relevant features for the target.

$$F_t' = F_t * M \tag{4}$$

Since the foreground pixels have higher attention values, they will be underlined, whereas the background pixels will be suppressed.

3.2.3 Mask Propagation

As shown in the above operation, the first frame is subjected to a feature-based similarity metric with the current frame (t-th frame), and the ($t-1$)-th frame is also performed with the current frame to generate another weighted attention map containing historical information, which is weighted and fused with the features of the current frame, respectively. As described in the process above, we use a mask propagation approach to pass pixel labels from sampled history frames to the current frame based on feature similarity metric in an embedding space. The purpose of enhancing the target features and learning the target appearance changes are achieved, while also weakening the background information around the target.

3.3 Attention Refinement Pathway (ARP)

Similar to SiamMask [5], we follow the idea of [50], generating masks by flattening representations of objects. This representation corresponds to a $1 \times 1 \times 256$ response bar, one of a $17 \times 17 \times 256$ response map, which results from depth-wise cross-correlation.

Since the network h_φ for segmentation task is based on two 1×1 convolutional layers with 256 and 63^2 channels, respectively. This enables each $1 \times 1 \times 256$ response bar can contain the information of a candidate region. Then we employ the strategy described in [66], which combines multiple refinement modules consisting of multiple upsampling layers and skip connections, merging different resolution features to generate the corresponding mask. To further refine the segmentation profile in the process of mask generation, an ECA (Efficient Channel Attention) module [67] is added to refinement modules to encode enhanced low-resolution features (**Fig. 2**), which pays more attention to the low-level features sensitive to edge detail by calculating attention weights. Thus more precise object masks are produced by improved refinement modules with attention.

Fig. 1 shows the detailed structure of the attention refinement pathway, which explicitly shows the stack of improved refinement modules for generating the final mask. With the same feature extraction network, template z and search region x are processed to get $f_\theta(\mathrm{z})$ and $f_\theta(\mathrm{x})$, and they are depth-wise cross-correlated to obtain the features $g_\theta(\mathrm{z,x})$, where we refer to the n -th response map with $g_\theta^n(\mathrm{z,x}) \in R^{1\times1\times d}$. The feature maps extracted from the third, second, and first layers in the Siamese network in x are represented as f_1, f_2, and f_3, respectively. The feature $g_\theta^n(\mathrm{z,\ x})$ is deconvoluted and gradually fused with f_1, f_2, and f_3 , and upsampled to get mask representations with different resolutions until a final mask of $127 \times 127 \times 1$ is obtained.

Fig. 2 illustrates the structure of the refinement module, which includes the detailed fusion with shallow features and the upsampling process. In this paper, with the ECA module, low-level features with rich edge details of f_2 can be highlighted. The feature map f_2 is used to output a new feature f_2' via the ECA module, two convolutional layers, and two non-linear layers sequentially. The mask representation e_2 is utilized to obtain a new mask representation e_2' with the same size as f_2' via two convolutional layers and two non-linear layers. Then we fuse f_2' and e_2' by element-wise addition and producing a new mask representation e_3 by an upsampling layer.

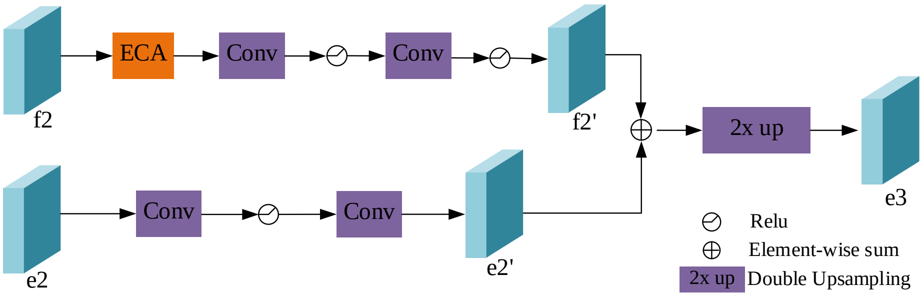


Fig. 2. The figure shows the structure as an example of a refinement module.

The ECA achieves a remarkable performance improvement with the addition of only a few parameters, as it applies a local cross-channel interaction technique without dimensionality reduction [51]. The ECA module structure is shown in Fig. 3. A fast 1 D convolution is available for achieving ECA efficiently. It is worth noting that the size of the convolution kernel determines the extent of cross-channel interaction, which refers to neighboring channels that contribute to predicting the attention of a particular channel. The resulting channel-wise attention map is then obtained by applying a sigmoid activation function to the output of the convolutional operation.

There exists a mapping relationship between channel dimension C and kernel size k . Then, given C , k can be defined adaptively by

$$k = \Psi(C) = \left|\frac{\log_2 C}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{5}$$

Where $|t|_{odd}$ indicates the nearest odd number of t. b and γ are the parameters of the mapping function set to 1 and 2 respectively. Evidently, through the mapping ψ, the ECA module enables high-dimensional channels to interact over longer ranges and low-dimensional channels to interact over shorter ranges. This is achieved by applying a non-linear transformation to the channel-wise attention map, which enhances the discriminative power of the features.

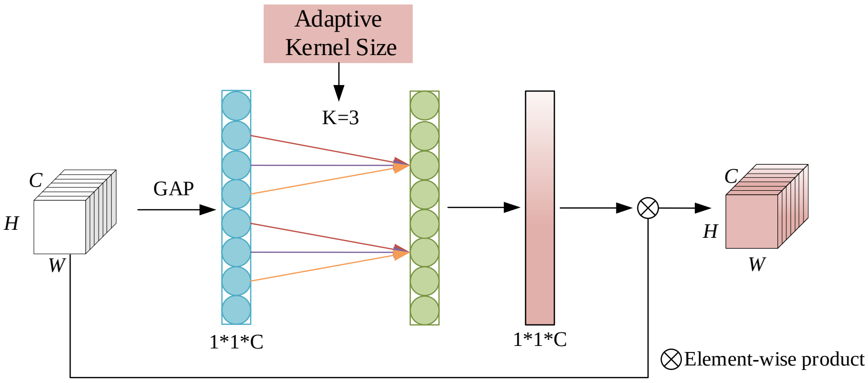


Fig. 3. Structure diagram of ECA module.

3.4 Mask Fusion Mechanism (MFM)

We propose a mask fusion mechanism (MFM) to obtain a more accurate mask by fusing the two masks, called the “initial mask” and the “previous mask”, generated from the attention refinement pathway (ARP). It can obtain the corresponding mask quality score adaptively by a neural network-based decision network [52] called Quality Evaluation Net in our paper, which is used as the fusion ratio for generating a reliable mask. Due to the fusion of the "initial mask" with the reliable information of the first frame and the "previous mask" of the previous frame with the motion state of the target, the final mask enjoys higher accuracy and stability.

Using neural networks to judge the quality of a binary mask allows us to utilize the capability of deep learning models to learn complex connections between inputs and outputs. The quality of binary masks is not always straightforward and depends on factors such as the accuracy of the mask boundary, the degree of overlap with the object of interest, the smoothness of the edges, and the consistency of the semantics captured by the mask. Classical techniques are unable to capture all of these characteristics. Neural networks, on the other hand, can learn complex patterns in data and can leverage large amounts of data to improve their accuracy, we therefore adopt a learning-based approach to achieve a more nuanced and reliable evaluation of binary masks.

The structure of the Quality Evaluation Net is illustrated in **Fig. 4**. It includes a feature extractor network ResNet-18 [53] and two fully connected (FC) layers. The function of the feature extractor is to encode the binary masks to feature space that gets the relative factors for determining the qualities of the masks. The first FC layer has 512 output units combined with a dropout layer and the next FC layer is a binary classification layer with $softmax$ cross-entropy loss.

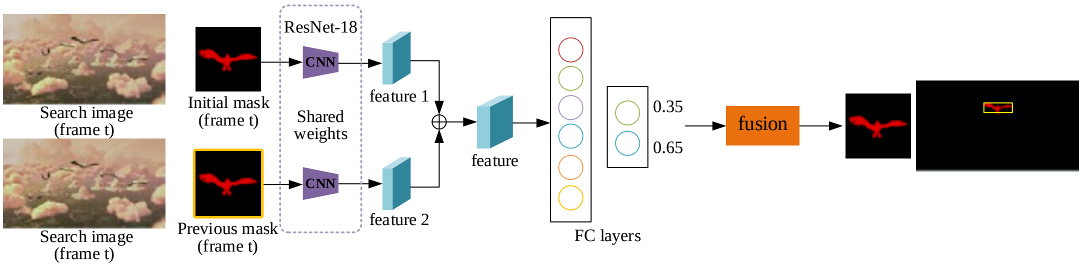


Fig. 4. Framework of the Quality Evaluation Net.

The network takes two inputs: the RGB image and the binary masks, both concatenated into a 4-channel input. The outputs of the feature extraction are flattened and combined to produce a 1024-dimensional vector. This vector is fed into two FC layers, with a dropout layer in between, to produce two final scores for the two masks using a $softmax$ output layer, where a higher score indicates better mask quality. Then we use these two scores as scale factors and fuse the corresponding two masks into one mask.

The network is trained using mean squared error loss, as shown in (6). Refer to [52] for details as we use the same formulation and parameters.

$$L = \frac{1}{N}\Sigma[(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2] \tag{6}$$

Where y_1, y_2 represent the ground truth of input and $\hat{y}_1$, $\hat{y}_2$ indicate the predicted values.

After getting two masks from the mask propagation-based attention module and attention refinement pathway, it is necessary to merge two masks into one with a certain proportion so that the final mask contains historical information of different levels of importance. To achieve this, the Quality Evaluation Net is needed that produces two scores called s_1 and s_2 , which are used as proportions. The coefficients represent the proportions of each mask that will be included in the final merged mask. The formula can be expressed as:

$$mask_{\mathrm{t}} = s_1 * mask1 + s_2 * mask2 \tag{7}$$

Where s_1 and s_2 are the two coefficients that determine the proportion of each mask in the final merged mask. The $mask1$ and $mask2$ are getting from the mask propagation-based attention module and attention refinement pathway that need to be merged. By utilizing the values of the coefficients generated automatically in the decision network rather than an artificial setting, the final mask can control the degree to which each mask contributes to the final merged mask.

4. Experiments

In this section, the proposed tracker is summarized and experimentally evaluated. Firstly, the settings and datasets of this work are introduced in Sect.4.1. Then, Sect.4.2 presents the ablation analysis of the proposed tracker to prove the effectiveness of our algorithm. Finally, the tracker is assessed with state-of-the-art algorithms on multiple publicly available datasets.

4.1 Settings and Datasets

Our tracker is implemented in Python3.7 with PyTorch framework on a PC equipped with an Intel i7-10700CPU (2.90 GHz), 16 GB RAM, and an NVIDIA GeForce GTX 1650 GPU. The target search region is cropped to 255×255. A binary mask output is obtained by thresholding the predicted segmentation at 0.15, the size of which is changed to 15×15 and then stored as the value in the Memory Bank. We evaluate our approach on benchmarks: VOT2016 [54], OTB-100 [55], TC128 [56], and UAV123 [57].

4.2 Ablation Analysis

The ablation experiment is analyzed to demonstrate the validity of the mask propagation-based attention module (MPAM) in joint with the mask fusion mechanism (MFM) and the attention refinement pathway (ARP) on tracker performance in the VOT2016 dataset. To evaluate the overall performance, we adopt the evaluation metrics of accuracy (Acc.), robustness (Rob.), and expected average overlap (EAO). The arrows next to the evaluation indicator indicate that the larger (↑) or the smaller (↑) represents better performance.

Table 1 indicates the ablation experiment results at baseline, the tracker with ARP, and Our tracker. The baseline method means the SiamMask [5]. 'Baseline+ ARP' means that the ARP is used to refine the segmentation profile and further improve the accuracy of the tracking bounding box. 'Baseline +ARP+(MPAM+MFM)' represents our algorithm including all the components proposed in this article. 'MPAM+MFM' is jointly utilized to retrieve mask information from historical frames and fuse two masks to generate the final mask.

Table 1. The results of ablation experiments.

Tracker	**Acc. ↑**	**Rob. ↓**	**EAO ↑**
Baseline	0.622	0.214	0.436
Baseline+ ARP	0.631	0.214	0.443
Baseline +ARP+(MPAM+MFM)	0.635	0.200	0.445

Compared with the Baseline method, 'Baseline + ARP' improves accuracy rates and EAO performance due to mining the low-level features that are sensitive to edge detail while enhancing target characterization. It achieves high Accuracy and EAO scores of 63.1% and 44.3% on the VOT2016 dataset. The performance of the tracker is further boosted after adding the MPAM and MFM in accuracy, robustness, and EAO, in particular, the robustness is significantly enhanced. The results demonstrate that the application of historical mask information and background constraints has a positive effect on tracker performance.

4.3 Quantitative Evaluation

We employ widely recognized standards to evaluate our results and enable comparison with other methods. The one-pass evaluation (OPE) is applied to the following datasets: OTB100, VOT2016, TC128, and UAV123.

4.3.1 Evaluation of the OTB Dataset

OTB100 dataset, a widely applied benchmark for evaluating the performance of trackers, contains 100 short-term video sequences with a total of 11,000 frames. It covers a diverse range of challenges such as background clutter, scale variation, and fast motion. We report the results on OTB100. Here we first compare our tracker with 11 recent state-of-the-art methods: DaSiamRPN [27], TADT [58], SESiamFCTracker [59], SiamMask [5], GradNet [13], DeepSRDCF [60], SiamRPN [17], SiamDWfc [61], SRDCF [62], SiamFC [26] and Staple [63]. **Fig. 5** shows the results of success plots and precision plots for OPE for the 11 trackers on OTB100. DaSiamRPN outperforms our tracking algorithm in terms of precision rate by the treatment of the influence of distractors and the strategy of a local-to-global strategy to redetect the target for long-term tracking. Our method ranks second in terms of precision rate and achieves optimal performance for the success rate.

In the OTB100 dataset, there are different challenges, such as out-of-view, background clutters, in-plane rotation, fast motion, illumination variation, and scale variation. **Fig. 6** displays the performance of success rates and precision rates for our method and other advanced methods on six different challenges. In general, our tracker outperforms the majority of those in comparison. It should be noted that our method performs best when facing background clutter, fast motion, and scale variation. The results show that DaSiamRPN achieves a higher precision rate when dealing with background clutters through employing a distractor-aware module and data augmentation, yet we also achieved an excellent performance through our approach. In our algorithm, we propose a mask propagation-based target attention module to constrain the background clutter in the features extracted from the backbone network. Furthermore, our method outperforms other approaches in the case of fast motion and scale variations. The reason for this is that we combine historical mask information with the current frame features.

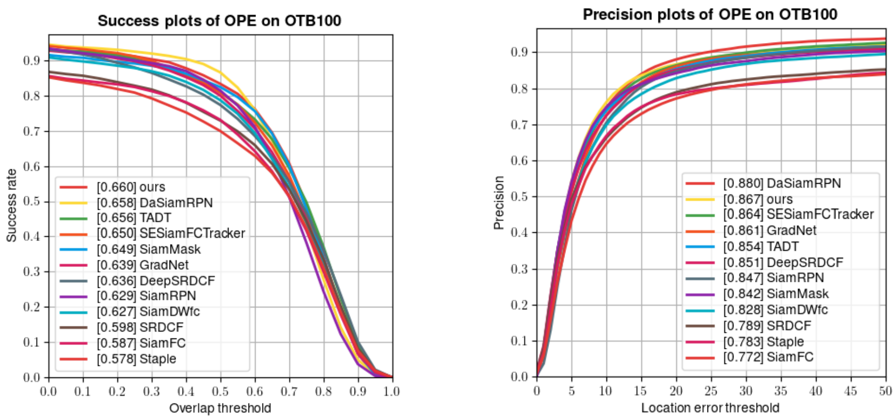


Fig. 5. Success and precision plots of OPE on the OTB100 dataset

To further demonstrate the effectiveness of our method, we select five typical sequences from the OTB dataset, which contain most of the challenges. Fig. 7 shows a qualitative comparison of our algorithm and baseline on the challenging videos. As can be seen from the figure, our algorithm works well in the face of background clutter, deformation, fast motion, and occlusion. For example, in the DragonBaby and Ironman sequence, the targets are deformed due to different appearance changes, but our algorithm can adapt to this situation. For background clutter, our algorithm can accurately perform target segmentation and localization in Ironman and ClifBar. Besides, for the case of fast motion, the Matrix sequence shows that our algorithm achieves very effective results. When the target is occluded, as shown in Freeman4, our method can accurately localize the target compared to the baseline tracker.

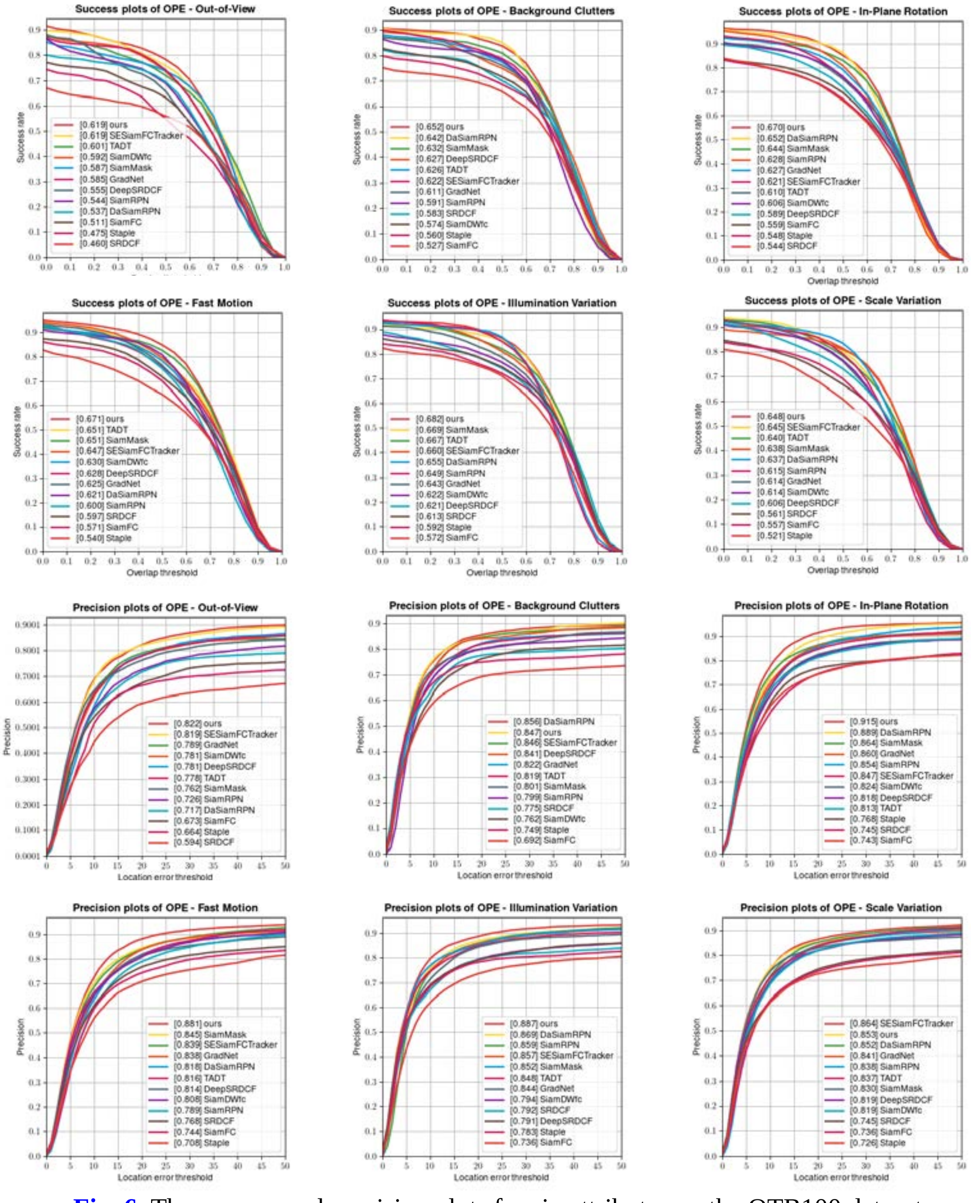


Fig. 6. The success and precision plots for six attributes on the OTB100 dataset

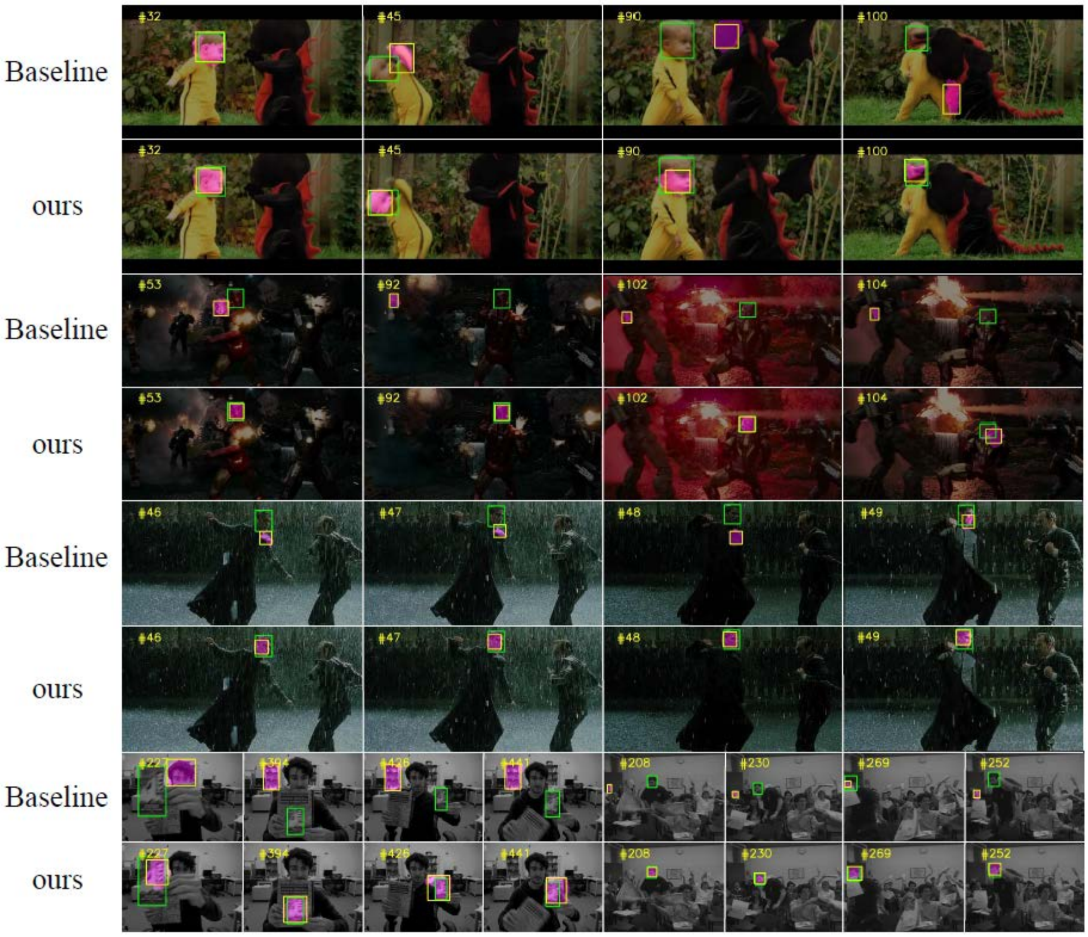


Fig. 7. The results of our method and baseline on five challenging sequences. (DragonBaby, Ironman, Matrix, ClifBar, Freeman4) The odd-numbered rows represent our algorithm and the even-numbered rows are the baseline trackers. The green bounding box is the ground truth, and the segmentation and tracking results are shown with the pink mask and the yellow bounding box.

4.3.2 Evaluation of the VOT2016 Dataset

Compared with OTB100, the VOT2016 dataset comprises 60 sequences, which provide higher resolution and have no grayscale data. Its evaluation metrics include three aspects, which are accuracy (Acc.: the higher the better.), robustness (Rob.: the lower is better.), and expected average overlap (EAO: the higher is better.). **Table 2** shows the performance of our tracker compared with 8 competitive trackers, including ADMT [64], RMIT [65], MemTrack [12], SiamMask [5], SiamRPN [17], SCSAtt [66], SCS-Siam [67], SiamLM [68]. It can be seen that our tracker ranks first in accuracy and EAO scores. Our tracker reaches 0.2, the same as the advanced tracker ADMT, and ranks third under the robustness criterion. RMIT and SCSAtt have better robustness, but their accuracy and EAO scores are much lower than ours. Overall, the results show that our algorithm can effectively enhance the target representation.

Table 2. The comparison of experimental results on the VOT2016 dataset. Red, green, and blue highlights indicate the best, second-best, and third-best values.

Tracker	Acc. ↑	Rob. ↓	EAO ↑
Ours	0.635	0.200	0.445
ADMT	0.560	0.200	0.368
RMIT	0.535	0.166	0.289
MemTrack	0.531	0.373	0.272
SiamMask	0.622	0.214	0.436
SiamRPN	0.560	0.302	0.334
SCSAtt	0.550	0.193	0.302
SCS-Siam	0.550	0.210	0.280
SiamLM	0.450	0.327	0.280

4.3.3 Evaluation of the TC128 Dataset

Unlike OTB, 128 color video sequences in the TC-128 dataset present more complex and challenging tracking tasks. To prove the universality of our algorithm, we further conducted tests on the TC-128 dataset. In **Table 3**, we perform quantitative comparisons between our trackers and multiple state-of-the-art trackers, including SiamMask [5], RMIT [65], HCFT [69], HCFTstar [70], CREST [71], and ECO [72]. As seen from the Table, our approach ranks first in success rate and second in precision, with a precision score of 74.3% and a success rate AUC score of 55.7%, respectively. In comparison with the baseline method SiamMask, we have a success rate of 17% and a precision rate of 18% higher, respectively.

Our tracker is better in success rate but slightly less accurate compared to the memory tracker RMIT. RMIT also mines target appearance information, using the memory residual branch to provide memory content about the target. Our approach differs from it in that the content of the memory is different. We memorize the mask information related to the history frames and use the method of mask propagation as well as the attention mechanism to tap the robust target expression. Thus our algorithm and RMIT both have better performance than other trackers.

Table 3. The comparison of experimental results on the TC-128 dataset. Red, green, and blue highlights indicate the best, second-best, and third-best values.

Tracker	Success rate	Precision
Ours	0.557	0.743
SiamMask	0.540	0.725
RMIT	0.551	0.761
HCFT	0.495	0.692
HCFTstar	0.488	0.695
CREST	0.533	0.708
ECO	0.555	0.740

4.3.4 Evaluation of the UAV123 Dataset

The UAV123 dataset, which consists of 123 challenging video sequences captured by UAVs from low altitudes, has an average length of 915 frames per video. It presents significant challenges for trackers due to the instability of the UAV view and frequent changes in distance to the target, resulting in low resolutions for many objects. **Fig. 8** shows the tracking results compared to other 7 Siamese trackers including SiamMask [5], SiamBAN [6], SiamRPN++ [7], SiamCAR [9], SiamRPN [17], SiamDW [61], ADMT [64]. Our tracker achieves a success score of 0.616 and a precision score of 0.811, which still outperforms recent competitive

Siamese trackers. It is worth noting that our algorithm can memorize variations in target appearance to improve the robustness of the tracker to challenges such as aspect ratio change, scale variation, and partial occlusion.

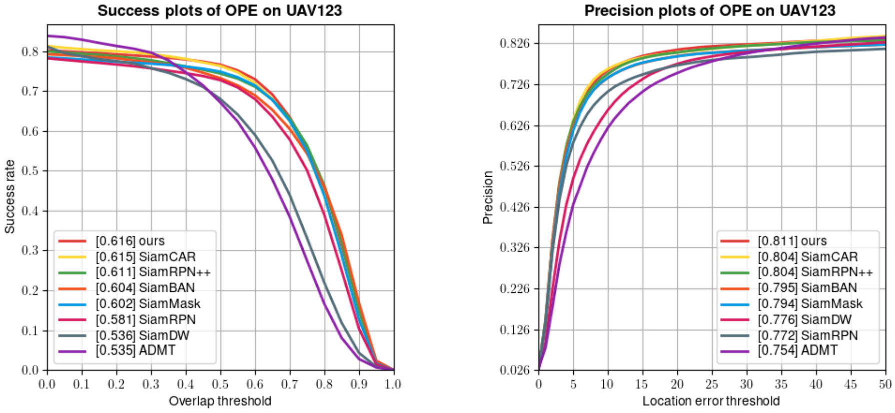


Fig. 8. Success plots and precision plots of OPE on the UAV123 dataset

5. Conclusion

In this work, we propose a mask propagation-based attention module that leverages rich historical mask information about the object and prevents the background noise of the backbone features. By learning mask information of the historical frame based on feature similarity calculation, target attention maps are produced to highlight the target region on backbone features, thus suppressing the adverse effects of background interference. To further refine the segmentation result, we introduce a lightweight attention module to the upsampling pathway, which focuses on low-level features that are sensitive to edge details, resulting in segmentation with clear edges and improved tracking performance. Additionally, we utilize a neural network-based method for mask quality assessment to obtain quality scores, which are used to assign weights and fuse masks to obtain a more robust mask. Extensive experiments on multiple benchmark datasets demonstrate the effectiveness of our proposed tracker when compared to advanced algorithms.

Acknowledgment

The authors gratefully acknowledge the anonymous editors and reviewers for viewing this paper and valuable comments. Besides, this work is supported by the National Natural Science Foundation of China under Grant (62272423, 62102373), The Distinguished Youth Science Foundation of Henan Province of China under Grant (242300421055), Henan Province Key *R&D* Project (241111210400).

References

[1] Liu, L., Xing, J., Ai, H., Ruan, X., "Hand posture recognition using finger geometric feature," in *Proc. of the 21st International Conference on Pattern Recognition*, pp.565-568, Nov. 2012. Article (CrossRef Link)

[2] Adam, M.S., Anisi, M.H., Ali, I., "Object tracking sensor networks in smart cities: Taxonomy, architecture, applications, research challenges and future directions," *Future Generation Computer Systems*, vol.107, pp.909-923, Jun. 2020. Article (CrossRef Link)

[3] Liu, L., Xing, J., Ai, H., "Multi-view vehicle detection and tracking in crossroads," in *Proc. of the First Asian Conference on Pattern Recognition*, pp.608-612, 2011. Article (CrossRef Link)

[4] Emami, A., Dadgostar, F., Bigdeli, A., Lovell, B.C., "Role of Spatiotemporal Oriented Energy Features for Robust Visual Tracking in Video Surveillance," in *Proc. of 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pp.349-354. 2012. Article (CrossRef Link)

[5] Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P. H.S., "Fast Online Object Tracking and Segmentation: A Unifying Approach," in *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.1328-1338, 2019. Article (CrossRef Link)

[6] Chen, Z., Zhong, B., Li, G., Zhang, S., Ji, R., "Siamese Box Adaptive Network for Visual Tracking," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.6667-6676, 2020. Article (CrossRef Link)

[7] Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J., "SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks," in *Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4277-4286, 2019. Article (CrossRef Link)

[8] Xu, Y., Wang, Z., Li, Z., Yuan, Y., Yu, G., "SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol.34, no.7, pp.12549-12556, 2020. Article (CrossRef Link)

[9] Guo, D., Wang, J., Cui, Y., Wang, Z., Chen, S., "SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.6268-6276, 2020. Article (CrossRef Link)

[10] Bhat, G., Danelljan, M., Gool, L.V., Timofte, R., "Learning Discriminative Model Prediction for Tracking," in *Proc. of the 2019 IEEE/CVF International Conference on Computer Vision*, pp.6181-6190, 2019. Article (CrossRef Link)

[11] Danelljan, M., Gool, L.V., Timofte, R., "Probabilistic Regression for Visual Tracking," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7181-7190, 2020. Article (CrossRef Link)

[12] Yang, T., Chan, A.B., "Learning Dynamic Memory Networks for Object Tracking," in *Proc. of the 15th European Conference on Computer Vision – ECCV 2018*, vol.11213, pp.153-169, Oct. 2018. Article (CrossRef Link)

[13] Li, P., Chen, B., Ouyang, W., Wang, D., Yang, X., Lu, H., "GradNet: Gradient-Guided Network for Visual Object Tracking," in *Proc. of the 2019 IEEE/CVF International Conference on Computer Vision*, pp.6161-6170, 2019. Article (CrossRef Link)

[14] Yu, Y., Xiong, Y., Huang, W., Scott, M. R., "Deformable Siamese Attention Networks for Visual Object Tracking," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.6727-6736, 2020. Article (CrossRef Link)

[15] Liu, T., Wang, G., Yang, Q., "Real-time part-based visual tracking via adaptive correlation filters," in *Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp.4902-4912, 2015. Article (CrossRef Link)

[16] Henriques, J. F., Caseiro, R., Martins, P., Batista, J., "High-Speed Tracking with Kernelized Correlation Filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.37, no.3, pp.583-596, Mar. 2015. Article (CrossRef Link)

[17] Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X., "High Performance Visual Tracking with Siamese Region Proposal Network," in *Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.8971-8980, 2018. Article (CrossRef Link)

[18] Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., Torr, P. H. S., "End-to-End Representation Learning for Correlation Filter Based Tracking," in *Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp.5000-5008, 2017. Article (CrossRef Link)

[19] Zhou, W., Wen, L., Zhang, L., Du, D., Luo, T., Wu, Y., "SiamMan: Siamese Motion-aware Network for Visual Tracking," *arXiv preprint arXiv:1912.05515*, 2019. Article (CrossRef Link)

[20] Zhu, J., Chen, T., Cao, J., "Siamese Network Using Adaptive Background Superposition Initialization for Real-Time Object Tracking," *IEEE Access*, vol.7, pp.119454-119464, 2019. Article (CrossRef Link)

[21] Yao, R., Lin, G., Shen, C., Zhang, Y., Shi, Q., "Semantics-Aware Visual Object Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.29, no.6, pp.1687-1700, 2019. Article (CrossRef Link)

[22] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Proc. of the 15th European Conference on Computer Vision – ECCV 2018, Lecture Notes in Computer Science*, vol.11211, pp.833-851, 2018. Article (CrossRef Link)

[23] Voigtlaender, P., Luiten, J., Torr, P. H.S., Leibe, B., "Siam R-CNN: Visual Tracking by Re-Detection," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.6577-6587, 2020. Article (CrossRef Link)

[24] Lukežič, A., Matas, J., Kristan, M., "D3S – A Discriminative Single Shot Segmentation Tracker," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7131-7140, 2020. Article (CrossRef Link)

[25] Marvasti-Zadeh, S. M., Cheng, L., Ghanei-Yakhdan, H., Kasaei, S., "Deep Learning for Visual Tracking: A Comprehensive Survey," *IEEE Transactions on Intelligent Transportation Systems*, vol.23, no.5, pp.3943-3968, May 2022. Article (CrossRef Link)

[26] Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., Torr, P. H. S, "Fully-Convolutional Siamese Networks for Object Tracking," in *Proc. of Computer Vision – ECCV 2016 Workshops*, pp.850-865, Oct. 2016. Article (CrossRef Link)

[27] Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W., "Distractor-Aware Siamese Networks for Visual Object Tracking," in *Proc. of the 15th European Conference on Computer Vision – ECCV 2018*, pp.103-119, 2018. Article (CrossRef Link)

[28] Lu, X., Wang, W., Danelljan, M., Zhou, T., Shen, J., Van Gool, L., "Video Object Segmentation with Episodic Graph Memory Networks," in *Proc. of 16th European Conference on Computer Vision – ECCV 2020*, vol.12348, pp.661-679, 2020. Article (CrossRef Link)

[29] Yang, Z., Wei, Y., Yang, Y., "Collaborative Video Object Segmentation by Foreground-Background Integration," in *Proc. of 16th European Conference on Computer Vision – ECCV 2020*, vol.12350, pp.332-348, 2020. Article (CrossRef Link)

[30] Yang, Z., Wei, Y., Yang, Y., "Associating Objects with Transformers for Video Object Segmentation," in *Proc. of 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, vol.34, pp.2491-2502, 2021. Article (CrossRef Link)

[31] Zeng, X., Liao, R., Gu, L., Xiong, Y., Fidler, S., Urtasun, R., "DMM-Net: Differentiable Mask-Matching Network for Video Object Segmentation," in *Proc. of the 2019 IEEE/CVF International Conference on Computer Vision*, pp.3928-3937, 2019. Article (CrossRef Link)

[32] Zhang, L., Lin, Z., Zhang, J., Lu, H., He, Y., "Fast Video Object Segmentation via Dynamic Targeting Network," in *Proc. of the 2019 IEEE/CVF International Conference on Computer Vision*, pp.5581-5590, 2019. Article (CrossRef Link)

[33] Chen, X., Li, Z., Yuan, Y., Yu, G., Shen, J., Qi, D., "State-Aware Tracker for Real-Time Video Object Segmentation," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.9381-9390, 2020. Article (CrossRef Link)

[34] Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L.-C., "FEELVOS: Fast End-To-End Embedding Learning for Video Object Segmentation," in *Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.9473-9482, 2019. Article (CrossRef Link)

[35] Oh, S.W., Lee, J.-Y., Xu, N., Kim, S.J., "Video Object Segmentation Using Space-Time Memory Networks," in *Proc. of the 2019 IEEE/CVF International Conference on Computer Vision*, pp.9225-9234, 2019. Article (CrossRef Link)

[36] Zhang, Y., Wu, Z., Peng, H., Lin, S., "A Transductive Approach for Video Object Segmentation," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.6947-6956, 2020. Article (CrossRef Link)

[37] Zhang, Z., Hua, Y., Song, T., Xue, Z., Ma, R., Robertson, N., Guan, H., "Tracking-assisted Weakly Supervised Online Visual Object Segmentation in Unconstrained Videos," in *Proc. of the 26th ACM International Conference on Multimedia*, pp.941-949, Oct. 2018. Article (CrossRef Link)

[38] Itti, L., Koch, C., Niebur, E., "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.20, no.11, pp.1254-1259, Nov. 1998. Article (CrossRef Link)

[39] Hu, J., Shen, L., Sun, G., "Squeeze-and-Excitation Networks," in *Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7132-7141, 2018. Article (CrossRef Link)

[40] Fan, Q., Zhuo, W., Tang, C.-K., Tai, Y.-W., "Few-Shot Object Detection With Attention-RPN and Multi-Relation Detector," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4012-4021, 2020. Article (CrossRef Link)

[41] Yang, K., He, Z., Zhou, Z., Fan, N., "Siamatt: Siamese attention network for visual tracking," *Knowledge-Based Systems*, vol.203, Sep. 2020. Article (CrossRef Link)

[42] Ding, L., Tang, H., Bruzzone, L., "LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol.59, no.1, pp.426-435, Jan. 2021. Article (CrossRef Link)

[43] Woo, S., Park, J., Lee, J.-Y., Kweon, I. S., "CBAM: Convolutional Block Attention Module," in *Proc. of the 15th European Conference on Computer Vision – ECCV 2018*, vol.11211, pp.3-19, 2018. Article (CrossRef Link)

[44] Wang, X., Girshick, R., Gupta, A., He, K., "Non-local Neural Networks," in *Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7794-7803, 2018. Article (CrossRef Link)

[45] Xiao, D., Tan, K., Wei, Z. et al., "Siamese block attention network for online update object tracking," *Applied Intelligence*, vol.53, no.3, pp.3459-3471, 2023. Article (CrossRef Link)

[46] Yuan, Y., Huang, L., Guo, J., Zhang, C., Chen, X., Wang, J., "OCNet: Object Context for Semantic Segmentation," *International Journal of Computer Vision*, vol.129, pp.2375-2398, 2021. Article (CrossRef Link)

[47] Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H., "Expectation-Maximization Attention Networks for Semantic Segmentation," in *Proc. of the 2019 IEEE/CVF International Conference on Computer Vision*, pp.9166-9175, 2019. Article (CrossRef Link)

[48] Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., Smola, A., "ResNeSt: Split-Attention Networks," in *Proc. of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp.2735-2745, 2022. Article (CrossRef Link)

[49] Choi, S., Kim, J. T., Choo, J., "Cars Can't Fly up in the Sky: Improving Urban-Scene Segmentation via Height-driven Attention Networks," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.9373-9383, 2020. Article (CrossRef Link)

[50] Pinheiro, P. O., Lin, T.-Y., Collobert, R., Dollár, P., "Learning to Refine Object Segments," in *Proc. of 14th European Conference on Computer Vision – ECCV 2016*, vol.9905, pp.75-91, 2016. Article (CrossRef Link)

[51] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11531-11539, 2020. Article (CrossRef Link)

[52] Garg, S., Goel, V., "Mask Selection and Propagation for Unsupervised Video Object Segmentation," in *Proc. of the 2021 IEEE Winter Conference on Applications of Computer Vision*, pp.1679-1689, 2021. Article (CrossRef Link)

[53] He, K., Zhang, X., Ren, S., Sun, J., "Deep Residual Learning for Image Recognition," in *Proc. of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016. Article (CrossRef Link)

[54] Kristan, M. et al., "The Visual Object Tracking VOT2016 Challenge Results," in *Proc. of Computer Vision – ECCV 2016 Workshops*, vol.9914, pp.777-823, 2016. Article (CrossRef Link)

[55] Wu, Y., Lim, J., Yang, M.-H., "Online Object Tracking: A Benchmark," in *Proc. of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp.2411-2418, 2013. Article (CrossRef Link)

[56] Liang, P., Blasch, E., Ling, H., "Encoding Color Information for Visual Tracking: Algorithms and Benchmark," *IEEE Transactions on Image Processing*, vol.24, no.12, pp.5630-5644, Dec. 2015. Article (CrossRef Link)

[57] Mueller, M., Smith, N., Ghanem, B., "A Benchmark and Simulator for UAV Tracking," in *Proc. of 14th European Conference on Computer Vision – ECCV 2016*, vol.9905, pp.445-461, Oct. 2016. Article (CrossRef Link)

[58] Li, X., Ma, C., Wu, B., He, Z., Yang, M.-H., "Target-Aware Deep Tracking," in *Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.1369-1378, 2019. Article (CrossRef Link)

[59] Sosnovik, I., Moskalev, A., Smeulders, A., "Scale Equivariance Improves Siamese Tracking," in *Proc. of the 2021 IEEE Winter Conference on Applications of Computer Vision*, pp.2764-2773, 2021. Article (CrossRef Link)

[60] Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M., "Convolutional Features for Correlation Filter Based Visual Tracking," in *Proc. of the 2015 IEEE International Conference on Computer Vision Workshop*, pp.621-629, 2015. Article (CrossRef Link)

[61] Zhang, Z., Peng, H., "Deeper and Wider Siamese Networks for Real-Time Visual Tracking," in *Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4586-4595, 2019. Article (CrossRef Link)

[62] Danelljan, M., Häger, G., Shahbaz Khan, F., Felsberg, M., "Learning Spatially Regularized Correlation Filters for Visual Tracking," in *Proc. of the 2015 IEEE International Conference on Computer Vision*, pp.4310-4318, 2015. Article (CrossRef Link)

[63] Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P. H. S., "Staple: Complementary Learners for Real-Time Tracking," in *Proc. of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp.1401-1409, 2016. Article (CrossRef Link)

[64] Zhang, H., Liang, J., Zhang, J., Zhang, T., Lin, Y., Wang, Y., "Attention-Driven Memory Network for Online Visual Tracking," *IEEE Transactions on Neural Networks and Learning Systems*, pp.1-14, 2023. Article (CrossRef Link)

[65] Zhang, H., Zhang, J., Nie, G., Hu, J., Zhang, W.J., "Residual memory inference network for regression tracking with weighted gradient harmonized loss," *Information Sciences*, vol.597, pp.105-124, Jun. 2022. Article (CrossRef Link)

[66] Rahman, M. M., Fiaz, M., Jung, S. K., "Efficient Visual Tracking With Stacked Channel-Spatial Attention Learning," *IEEE Access*, vol.8, pp.100857-100869, 2020. Article (CrossRef Link)

[67] Fiaz, M., Mahmood, A., Jung, S. K., "Learning Soft Mask Based Feature Fusion with Channel and Spatial Attention for Robust Visual Object Tracking," *Sensors*, vol.20, no.14, 2020. Article (CrossRef Link)

[68] Zhang, H., Chen, J., Nie, G., Lin, Y., Yang, G., Zhang, W.J., "Light regression memory and multi-perspective object special proposals for abrupt motion tracking," *Knowledge-Based Systems*, vol.226, Aug. 2021. Article (CrossRef Link)

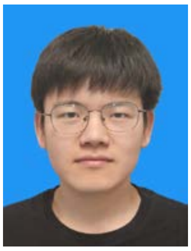
[69] Ma, C., Huang, J.-B., Yang, X., Yang, M.-H., "Hierarchical Convolutional Features for Visual Tracking," in *Proc. of the 2015 IEEE International Conference on Computer Vision*, pp.3074-3082, 2015. Article (CrossRef Link)

[70] Ma, C., Huang, J.-B., Yang, X., Yang, M.-H., "Robust Visual Tracking via Hierarchical Convolutional Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.41, no.11, pp.2709-2723, Nov. 2019. Article (CrossRef Link)

[71] Song, Y., Ma, C., Gong, L., Zhang, J., Lau, R. W.H., Yang, M.-H., "CREST: Convolutional Residual Learning for Visual Tracking," in *Proc. of the 2017 IEEE International Conference on Computer Vision*, pp.2574-2583, 2017. Article (CrossRef Link)

[72] Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M., "ECO: Efficient Convolution Operators for Tracking," in *Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp.6931-6939, 2017. Article (CrossRef Link)

Huanlong Zhang from the School of Aeronautics and Astronautics, Shanghai Jiao Tong University, China, in 2015. He is currently an Associate Professor with the College of Electric and Information Engineering, Zhengzhou University of Light Industry, Henan, Zhengzhou, China. He has published more than 40 technical articles in refereed journals and conference proceedings. His research interests include pattern recognition, machine learning, image processing, computer vision, and intelligent human-machine systems.



Weiqiang Fu was born in Nanyang, Henan, China, in 2000. He is currently pursuing his master's degree at Zhengzhou Institute of Light Industry, China. His research interests include pattern recognition, machine learning, image processing, computer vision, and intelligent human-machine systems.

Bin Zhou was born in Nanyang, Henan, China, in 1996. She is currently pursuing her M.A. degree with the Zhengzhou University of Light Industry, Zhengzhou, China. Her research interests include pattern recognition, machine learning, image processing, computer vision, and intelligent human-machine systems.

Keyan Zhou was born in Luoyang, Henan, China, in 1999. She is currently pursuing her M.A. degree with the Zhengzhou University of Light Industry, Zhengzhou, China. Her research interests include pattern recognition, machine learning, image processing, computer vision, and intelligent human-machine systems.

Xiangbo Yang was born in Pingdingshan, Henan, China, in 1999. He is currently pursuing his M.A. degree with the Zhengzhou University of Light Industry, Zhengzhou, China. His research interests include pattern recognition, machine learning, image processing, computer vision, and intelligent human-machine systems.

Shanfeng Liu was born in December 1986 and received his Ph.D. degree from the University of Chinese Academy of Sciences. He is currently a senior engineer at State Grid Henan Electric Power Research Institute. His research interests include artificial intelligence and disaster prevention and mitigation.