Tracking Computer Vision Algorithm Based on Fusion Twin Network

Xin Wang
Software Institute, Hunan College of Information, Changsha, 410016, China

***Abstract*—Deep learning technology has promoted the rapid development of visual object tracking, among which algorithms based on twin networks are a hot research direction. Although this method has broad application prospects, its performance is often greatly reduced when encountering target occlusion or similar objects in the background. In response to this issue, a method is proposed to integrate channel and spatial dimension attention mechanisms into the backbone architecture of twin networks, to optimize the algorithm's recognition accuracy for tracking targets and its stability in changing environments. Then, a region recommendation network based on adaptive anchor box generation is adopted, combined with twin networks to enhance the network's modeling ability for complex situations. Finally, a new visual tracking algorithm is designed. Through comparative experiments, the success rate of the former increased by 0.6% and 0.9% respectively on the two datasets, and its accuracy also increased by 1.2% and 1.8% accordingly. The success rate of the latter increased by 1.5% and 1.2% respectively in the two datasets, and the accuracy also increased by 1.2% and 0.6% respectively. From this, the improved algorithm can improve the performance of target tracking and has certain application potential in visual target tracking.**

Keywords—Visual tracking; twin network; integration; attention mechanism; self-adaption

I. Introduction

Fatigue driving has become an important factor in causing traffic accidents, posing a serious threat to social safety and public health. Currently, although there are various fatigue driving detection models based on computer algorithms, they still face many challenges in practical applications, such as poor comfort, susceptibility to external factors such as lighting, masks, sunglasses, low detection accuracy, and poor real-time performance [1]. With the successful application of deep learning in object detection, some studies have begun to use detection techniques to guide the development of object tracking technology [2]. Xin et al. proposed a new Siamese adaptive learning network for visual tracking to address the manually adjusting parameters. The designed method took spatial alignment and model learning state as criteria for anchor quality evaluation, and employed Gaussian mixture distribution for adaptive allocation instead of IoU-based anchor allocation. The experimental results showed that the tracker had superiority on benchmarks GOT-10k, and LaSOT [3]. In order to address the issue of retaining much unfavorable background information in association operations, Jun W et al. proposed an effective feature recognizer that included channel and spatial attention modules to focus on key information. The representation capability was optimized. Experiments on six benchmark tests showed that the designed tracker outperformed other trackers. Especially, it achieved 80.4% AUC on TrackingNet and 68.4% AUC on GOT-10k during real-time operation [4]. Due to the insufficient control accuracy and stability of current virtual vision tracking technology, Jianbin et al. designed a new intelligent algorithm on the basis of human-computer interaction and virtual vision tracking technology to improve the overall performance of virtual vision tracking. The designed model was efficient, with tracking accuracy more than 10% higher than traditional methods [5]. Di et al. found that the discriminative model for predicting object tracking models was susceptible to interference from similar objects and required much-labeled data for training during actual use. Therefore, two methods were proposed to enhance the robustness of target tracking against interference from similar objects: multi-scale region search and response map processing based on Gaussian convolution. A large number of experiments showed that the enhancement function implemented in the tracking framework enhanced its robustness. The tracker based on self-supervised training had excellent tracking performance [6]. The existing Siamese trackers are insufficient to effectively distinguish between targets and the fluctuation interference embedded in the two branches of information, resulting in inaccurate classification and localization. Therefore, Liu et al. proposed two novel sub-network spaces for spatial feature embedding to optimize the discriminative ability of trackers in the embedding space and their adaptability to complex tracking scenarios. Compared with the most advanced trackers, the proposed tracker had competitive tracking performance in background clutter and similar object attributes, verifying the effectiveness of the method [7]. To solve the tracking drift caused by inaccurate initial positions in most existing trackers, Han et al. proposed a deep learning method that could generate accurate positions of objects given their rough positions. The proposed method was applied in the tracking process to improve the accuracy. A large number of experiments in object tracking benchmark testing verified its effectiveness [8]. Due to the lack of attention on the channel dimension in current trackers, their potential tracking capabilities are hindered. Therefore, Shaochuan et al. used a novel spatial channel converter that integrated information conveyed by features along both spatial and channel directions. To quantify temporal smoothness, a jitter metric that measured the cross-frame variation of predicted bounding boxes was proposed as a function of parameters like center displacement, area, and aspect ratio. Several well-known benchmark datasets demonstrated its robustness [9].

In the field of computer vision tracking, although some progress has been made in research, fatigue driving detection technology still faces multiple challenges. Firstly, existing models have shortcomings in terms of comfort, which may cause unnecessary interference to drivers. Secondly, sensitivity to external factors such as lighting, masks, sunglasses, etc. affects the accuracy and robustness of detection. In addition, detection accuracy and real-time performance are also key areas that need to be improved in current technology. For example, although the Siamese adaptive learning network proposed by Xin et al. performed well on GOT-10k and LaSOT benchmark tests, further optimization is still needed to improve tracking accuracy and stability. The effective feature recognizer proposed by the military and others has optimized its representation ability, but there is still room for improvement in real-time performance. The new spatial feature embedding subnetwork proposed by Liu et al. has optimized the recognition ability of trackers, but its adaptability in complex tracking scenarios still needs further verification. Although the deep learning method proposed by Han et al. can generate accurate positions of objects, there is still room for improvement in dealing with tracking drift caused by inaccurate initial positions.

To address these drawbacks, this study proposes a visual tracking algorithm based on traditional dual networks, combining attention mechanism and adaptive anchor box generation to improve the algorithm performance. The attention mechanism helps the algorithm to focus on key information and improve the efficiency of target feature extraction, while the adaptive anchor box generation is able to evaluate the anchor quality more accurately, using a Gaussian mixture distribution for adaptive allocation, thus improving the target tracking performance in complex environments.

II. Computer Vision Algorithm Based on Twin Network Tracking

A. Twin Network Visual Tracking Based on Attention Mechanism

In fatigue driving detection, the characteristic of computer vision algorithms is that they can analyze the driver's behavior and physiological signals in real time, such as blink frequency, head posture, etc., to determine whether they are in a fatigue state. However, existing algorithms often have low detection accuracy due to factors such as background interference and lighting changes [10]. To this end, an improved algorithm is proposed by introducing an attention mechanism to extract useful features. The designed structure utilizes multi-layer linear fusion to integrate different levels of feature information, thereby enhancing the recognition accuracy and robustness for fatigue states. Its structure is shown in Fig. 1.

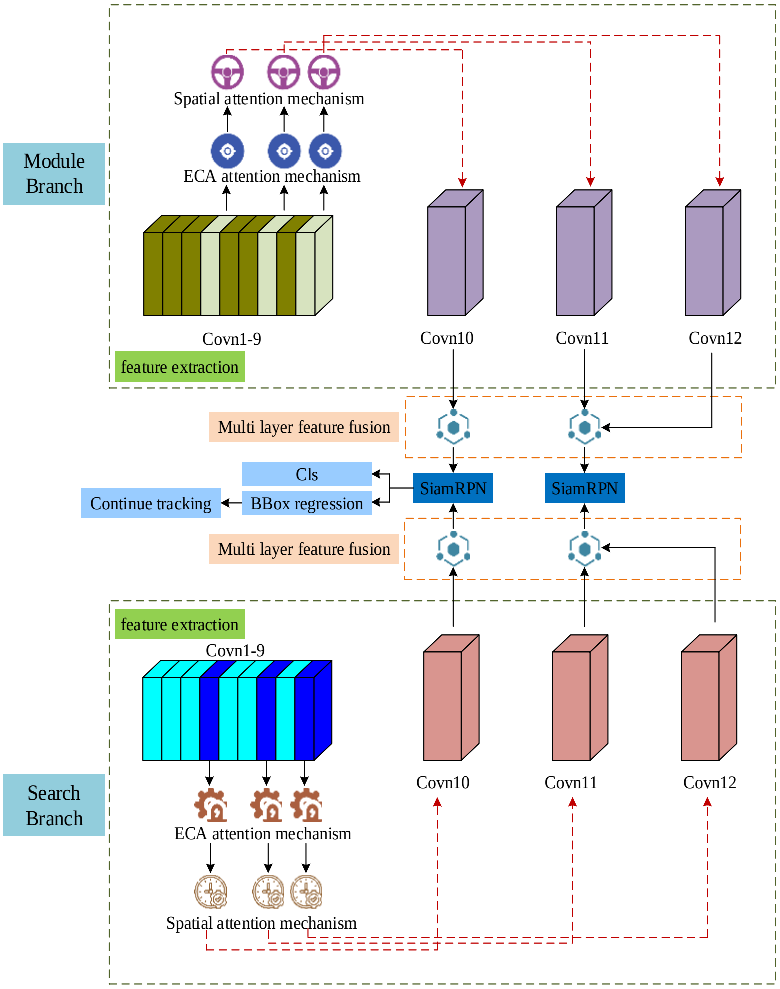


Fig. 1. Structure diagram of twin network target tracking algorithm incorporating attention mechanism.

From Fig. 1, the algorithm optimizes the feature extraction process in target tracking tasks by integrating channel and spatial attention mechanisms, thereby improving the performance of the tracking algorithm. Through analysis, it is found that striking a balance between accuracy and efficiency in target tracking is crucial [11]. Specifically, the research adopts the SiamRPN++ framework and makes improvements and adjustments based on it. Firstly, for the optimization of feature extraction, this method is similar to a fine filter. It can filter out feature information that is not conducive to target tracking, and only retain key information that helps improve tracking performance. This ensures the significant differences in feature responses across channels and spaces, providing a more accurate basis for subsequent similarity calculations. Secondly, in order to improve algorithm performance, effective fusion of feature information can be achieved through improvements to the backbone network. It includes two main parts: module branch and search branch. Both branches use Covin1-9 as the initial feature extraction layer and apply ECA attention mechanism and spatial attention mechanism to optimize feature representation capability. In the module branch, the features after passing through Covin1-9 are sent to the SiamRPN module for target tracking, and then the BBox regression is used to further accurately locate the position of the target [12]. In addition, the module branch also performs multi-layer feature fusion operations, combining features of different depths to obtain richer information. In the search branch, multi-layer feature fusion operations are also performed, but the fusion is performed after Covn1-9, indicating that this branch may be more focused on extracting and processing low-level features. This design fully utilizes the advantages of attention mechanism, enabling the network to output more accurate feature information, as shown in Fig. 2.

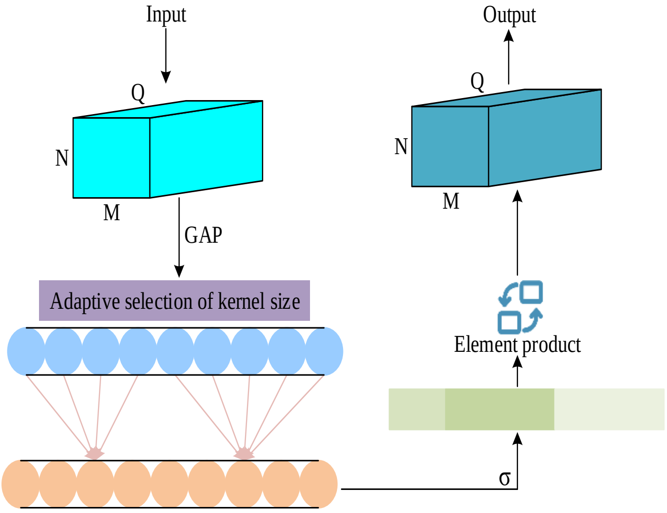


Fig. 2. Structure diagram of channel attention mechanism module.

From Fig. 2, the main purpose of the channel attention mechanism is to capture key feature information. It significantly enhances the network's performance in feature extraction by assigning higher weights to the target feature channel [13]. Then, the weights of each channel are obtained through the activation function, as shown in Eq. (1).

$$f(z)=\frac{1}{1+e^{-z}} \tag{1}$$

In Eq. (1), $f(z)$ represents the activation function. z represents the initial weight. The channel attention related feature map is shown in Eq. (2).

$$F_{ECA,i}=\omega_i \times F_i \tag{2}$$

In Eq. (2), $F_{ECA,i}$ is the attention feature map. ω_i is the weight on channel i. F_i is the original feature map on channel i. The output ω_i of each channel is weighted and summed all the features y_i^j within its receptive field, and then generated by the activation function σ, as shown in Eq. (3).

$$\omega_i=\sigma\left(\sum_{j=1}^{k}\alpha_i^j y_i^j\right), y_i^j \in \Omega_i^k \tag{3}$$

In Eq. (3), y_i^j represents the j-th element in Ω_i^k. α_i^j signifies the weight of the j-th element in the i-th channel. These weights are obtained through learning and are used for weighted combinations of different features. Then the corresponding weight of the size is generated, as shown in Eq. (4).

$$\omega=\sigma\left(CID_k\left(y\right)\right) \tag{4}$$

In Eq. (4), CID is a one-dimensional convolution. The size of k determines the coverage of interaction, and its relationship with the number of channels is shown in Eq. (5).

$$D\approx \exp\left(\gamma\times k-b\right) \tag{5}$$

In Eq. (5), D signifies the channels. γ and b represent empirical values. k can be determined by the number of channels D, as shown in Eq. (6).

$$k=\left|\frac{\log_2\left(D\right)}{\gamma}+\frac{b}{\gamma}\right| \tag{6}$$

To reduce the impact of background noise on target tracking and make the network more focused on the target itself. The study introduces the spatial attention mechanism, which weights different spatial positions of the feature map to focus on the key areas of the target object, while suppressing interference from background or irrelevant areas [14-15]. This helps the network automatically learn and focus on the key spatial positions of the target object, such as edges, corners, etc., thereby improving the recognition and tracking accuracy of the target. Its structure is shown in Fig. 3.

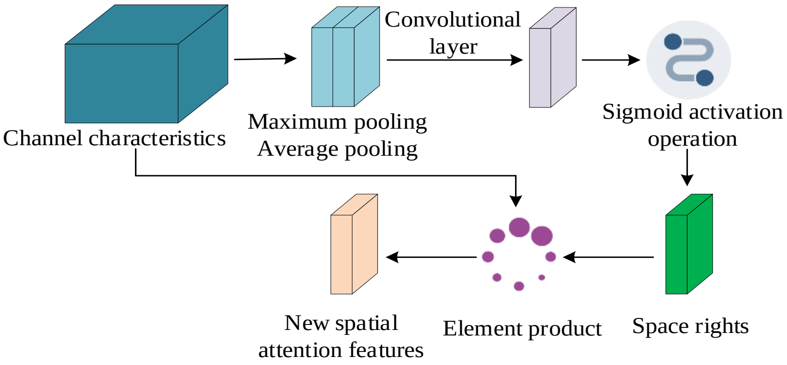


Fig. 3. Structure diagram of spatial attention mechanism module.

From Fig. 3, the input feature map first extracts deep features through convolutional layers, and then normalizes the output using the Sigmoid activation function, laying the foundation for subsequent spatial attention operations. The spatial attention mechanism multiplies the Sigmoid activated output with the original feature map through element multiplication operation. This operation achieves spatial position weighting of the feature map, focusing on key areas of the target (such as edges and corners) while reducing the interference of background noise. The weighted feature map is further processed through convolutional layers and combined with max pooling and average pooling techniques to extract global feature information. After these features are fused, they are refined again through convolutional layers. The generated spatial feature map significantly enhances the representation ability of the target area.

Finally, the spatial attention features is shown in Eq. (7).

$$M_S(F)=\sigma\left(f^{7*7}\left(\left[F_{avg}^{s};F_{\max}^{s}\right]\right)\right) \tag{7}$$

In Eq. (7), σ is the activation operation. F_{avg}^{s} is the average pooling feature. $F_{\max}^{s}$ is the maximum pooling feature. Then, multi-layer feature linear fusion is performed, and the fusion process is shown in Eq. (8).

$$\begin{cases}\hat{F}_4=\sum\left(F_4,F_5\right)\\ \hat{F}_3=\sum\left(F_3,\hat{F}_4\right)\end{cases} \tag{8}$$

In Eq. (8), $\hat{F}_4$ represents the fused mid-level feature map. F_4 represents the mid-level feature map. F_5 represents the deep feature map. $\hat{F}_3$ represents the shallow feature map after fusion. F_3 represents the shallow feature map.

B. *Twin Network Visual Tracking Algorithm Integrating Adaptive Anchor Box Generation*

Although introducing attention mechanisms and multi-layer linear fusion can optimize the extraction accuracy and robustness of algorithms, there are still some potential drawbacks. Firstly, the attention mechanism may increase the computational burden of the model, especially when dealing with rich channel and spatial position weights. Secondly, the model may overfit the training data, resulting in candidate boxes that cannot cover specific targets, thereby reducing its generalization ability on unseen data. Therefore, this study introduces a new adaptive anchor box to reduce manual intervention, enabling the model to automatically learn and generate candidate boxes based on the semantic information of the image itself, while reducing the computational burden of the attention mechanism. Its structure is displayed in Fig. 4.

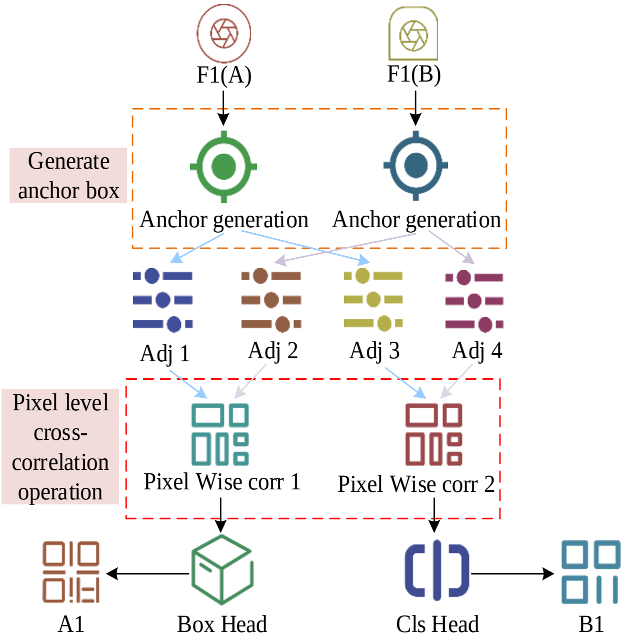


Fig. 4. Network structure diagram of adaptive generation anchor box region recommendation.

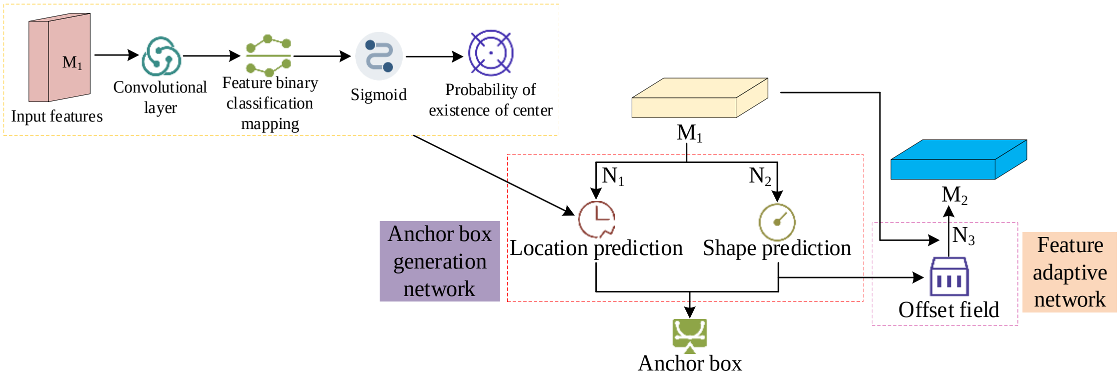


Fig. 5. Structure diagram of learnable generation module.

From Fig. 4, the input image generates feature maps through two convolutional layers F1 (A) and F1 (B), which are used for anchor box generation. Anchor boxes are generated at multiple scales (Adj 1 to Adj 4) through preset proportions and sizes to accommodate targets of different sizes and shapes. Subsequently, pixel level cross-correlation operations are performed between feature maps to generate Pixel Wise corr 1 and Pixel Wise corr 2 feature maps by calculating the correlation of corresponding pixel points, enhancing the spatial position information and correlation between features. These processed feature maps are further fed into the Box Head and Cls Head modules. Box Head is responsible for accurately regressing the coordinates of the target bounding box, while Cls Head performs classification tasks and predicts the target category for each anchor box. By optimizing the position and size of the anchor box, a detection box is ultimately generated to complete object detection. Based on the finer grained cross-correlation operations, the generated feature boundaries are made clearer [16]. This improves the success rate and accuracy of target tracking, thereby achieving more precise target tracking results. The core structure of generating anchor boxes is shown in Fig. 5.

From Fig. 5, the module has two parts: one is the anchor box generation network, which is used to generate unique anchor boxes, and the other is the feature adaptive network. The input features are convolved to obtain a binary classification map, and then the probability of center existence is obtained through the Sigmoid activation function; Next, this probability value is input into the anchor box generation network to generate anchor boxes. At the same time, the input features are fed into the feature adaptive network to obtain three features: N1, N2, and N3. Finally, these three features are input together with the anchor box into the localization prediction and shape prediction to obtain the final offset field. This method can infer the position and shape of the anchor box, as displayed in Eq. (9).

$$p\left(x,y,w,h\,|\,I\right)=p\left(x,y\,|\,I\right)p\left(w,h\,|\,x,y,I\right) \tag{9}$$

In Eq. (9), $p\left(x,y,w,h\,|\,I\right)$ represents the position and shape of the target. $p\left(x,y\,|\,I\right)$ represents that anchor boxes have different probabilities of appearing at different positions [17]. Convolutional neural networks automatically learn feature representations of images through a series of convolutional layers. The study employs a two channel convolutional sub network, and these features are subsequently used for classification tasks. Two channels independently predict the width and height of the corresponding anchor box, allowing the predicted box to more accurately fit the actual size of the target object, as displayed in Eq. (10).

$$\begin{cases} w=\sigma * s * e^{dw} \\ h=\sigma * s * e^{dh} \end{cases} \tag{10}$$

In Eq. (10), w is the width. h is high. σ is the empirical value. s is the step size. This study conducts deformation convolution with offset to obtain adaptive feature maps based on anchor box shapes at different positions, as shown in Eq. (11).

$$f_i^{'}=N_3\left(f_i,w_i,h_i\right) \tag{11}$$

In Eq. (11), $f_i^{'}$ is the convolved feature. N_3 is the deformable convolution. f_i is the original feature. $\left(w_i,h_i\right)$ is the shape of the anchor box corresponding to the position. In the application of adaptive anchor box region recommendation network, the loss function is one of the key factors in the training process, which is applied to measure the difference between the predicted and the true situation [18-19]. Therefore, the study combines learning position and shape branches to form an adaptive anchor box loss function. By minimizing this comprehensive loss, the model can learn more accurate object detection and tracking capabilities [20]. Its expression is shown in Eq. (12).

$$L=\lambda_1 L_{loc}+\lambda_2 L_{shape}+L_{cls}+L_{reg} \tag{12}$$

In Eq. (12), L_{loc} is the position loss, which is used to measure the difference in position (i.e. the coordinates of the bounding box) between the predicted box and the real box. L_{shape} signifies the shape loss, used to measure the difference in shape (i.e. the aspect ratio of the bounding box) between the predicted box and the real box. By optimizing this loss, the model can better learn the shape features of the target object. L_{cls} is the classification loss, and L_{reg} is the regression loss, used in object tracking to measure the distance between the predicted and the true boxes. λ_1 and λ_2 are used to balance the relative importance of position loss and shape loss in the total loss. In the position prediction branch, the loss function is displayed in Eq. (13).

$$L_{loc} = -\alpha(1-p)^{\gamma}\log(p) \quad (13)$$

In Eq. (13), α and γ represent the empirical values of the parameters. p signifies the probability of positive samples. In the shape prediction branch, the loss function can be used to optimize shape prediction without calculating the objective, as shown in Eq. (14).

$$L_{shape} = L_1\left(1-\min\left(\frac{w}{w_g},\frac{w_g}{w}\right)\right)+L_1\left(1-\min\left(\frac{h}{h_g},\frac{h_g}{h}\right)\right) \quad (14)$$

In Eq. (14), L_1 represents the smoothing loss. w_g signifies the real frame width, and h_g signifies the height. w signifies the candidate box width, and h signifies the height. After generating adaptive anchor boxes, the study will also use pixel level cross-correlation operations to represent features with higher quality, as shown in Fig. 6.

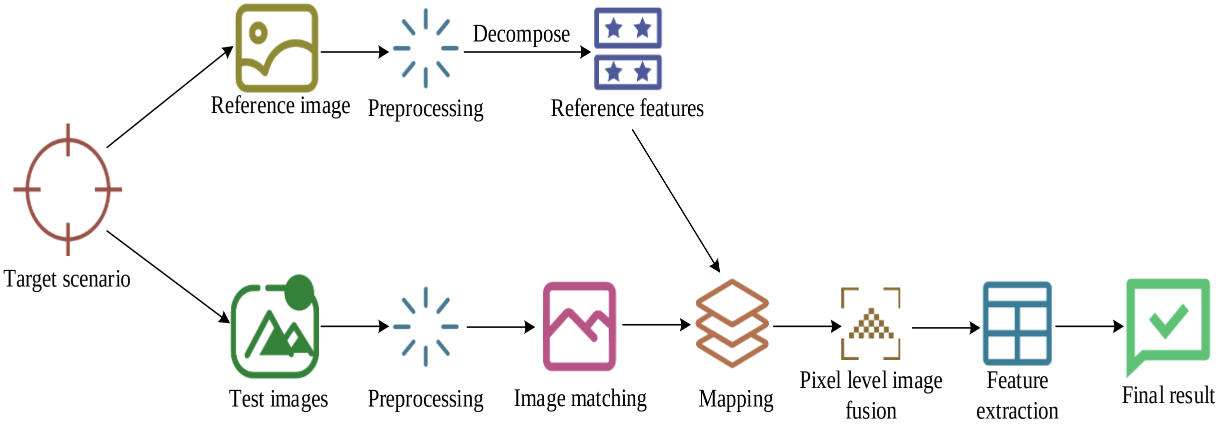


Fig. 6. Pixel level cross-correlation specific process.

From Fig. 6, in the object detection process, comparing the reference features with the test features generated after pre-processing of the test image, each part acts as a kernel to achieve deep interaction at the pixel level. After pixel level cross-correlation operation, high-quality features are sent to the mapping module for further processing. After optimization by the feature extraction module, detection results containing the precise position and category of the target are generated to avoid the target feature blurring. The process of generating fused images is shown in Eq. (15).

$$A = \{A_i \mid A_i = Z_i * X\}_{i\in\{1,\ldots,n\}} \quad (15)$$

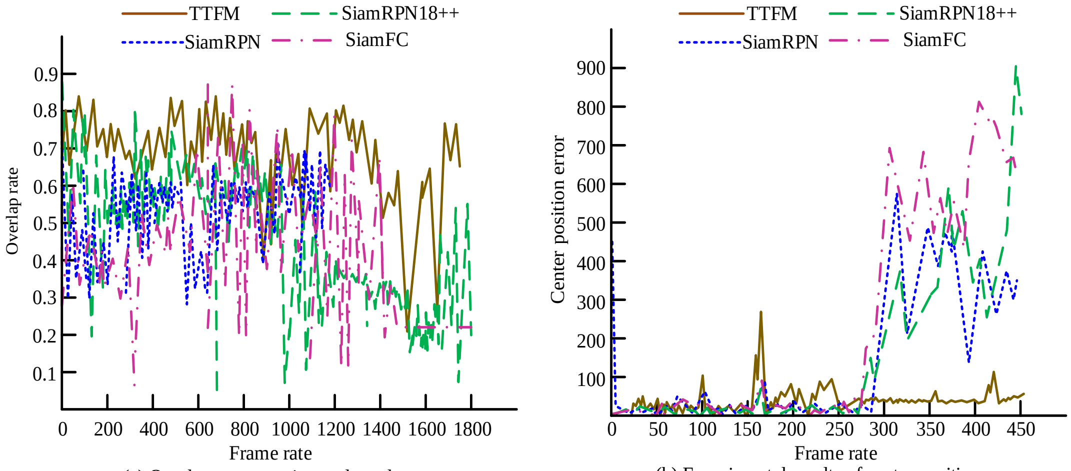
In Eq. (15), Z_i is the decomposed convolution kernel. $*$ signifies the convolution operation. X signifies the feature of the search area.

III. Results and Discussion

A. Analysis of Twin Network Visual Tracking Based on Attention Mechanism

In order to solve the low accuracy and poor robustness in traditional fatigue driving recognition, a twin network visual tracking algorithm on the basis of attention mechanism is proposed. To verify the performance of the proposed TTFM in target tracking in complex environments, particularly its robustness against interference factors such as lighting changes, occlusion, and size variations, the study first conducts multiple comparative experiments to assess the advantages of TTFM algorithm in tracking accuracy and efficiency. Fig. 7 displays the results, taking the center position error of overlap rate as the evaluation index.

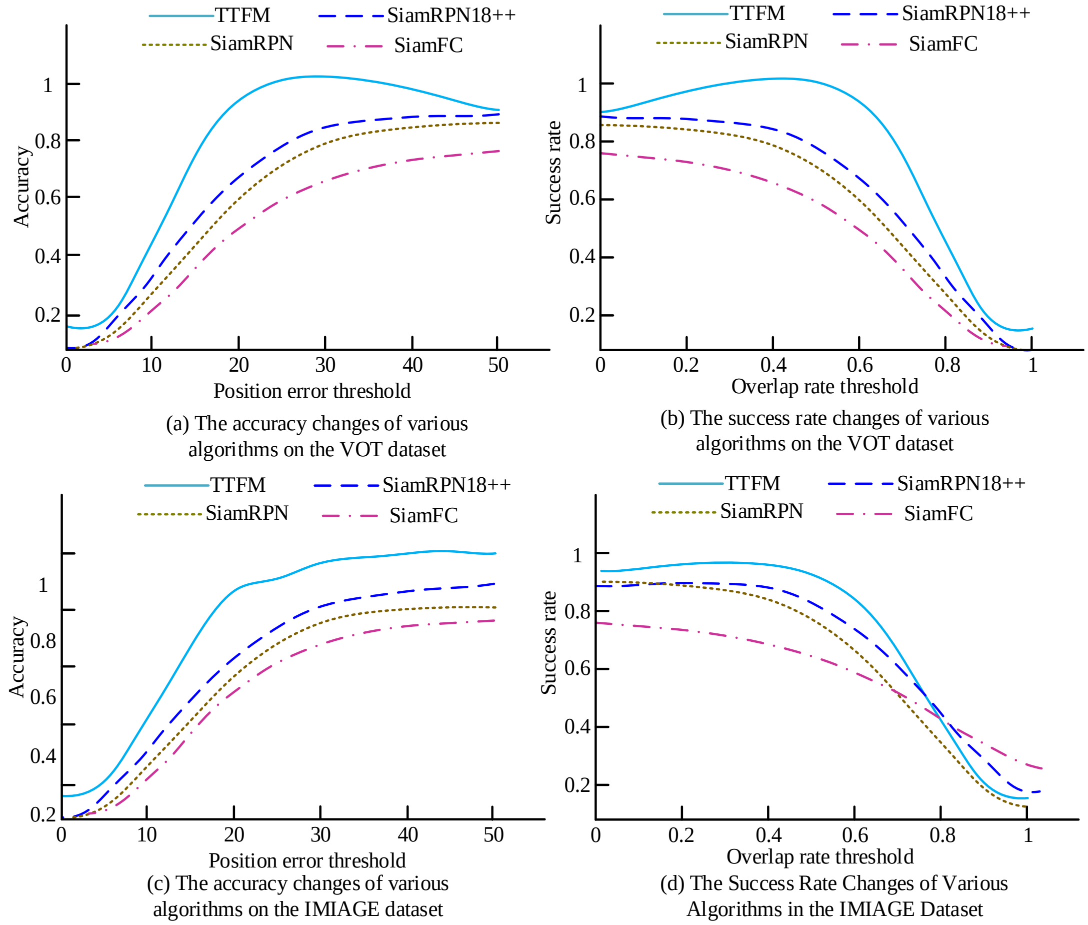
From Fig. 7, the TTFM algorithm had the best overlap rate of 0.75, indicating the best performance. Compared with other algorithms, the optimized TTFM algorithm increased the average overlap rate by 0.156%, effectively optimizing the efficiency of feature expression and target feature extraction. This lays the foundation for improving target tracking performance in complex environments. The TTFM algorithm performed well in tracking error at the center position of the sequence, with the smallest error. Compared with other basic algorithms, the TTFM algorithm reduced the center position error by 5.84, 30.63, and 43.42 pixels, respectively. When the target was subject to complex background interference such as changes in lighting, occlusion, and size, the TTFM algorithm significantly reduced the center position error. To further validate the effectiveness and performance improvement of the TTFM algorithm in single target tracking tasks. This study tested the commonly used VOT dataset and IMIAGE dataset, which contain visual tracking data under different adverse conditions such as lighting, to evaluate the improvement of TTFM algorithm and its generalization ability. The results are shown in Fig. 8.



(a) Overlap rate experimental results

(b) Experimental results of center position error

Fig. 7. Comparison analysis curve of overlap rate and center position.



(a) The accuracy changes of various algorithms on the VOT dataset

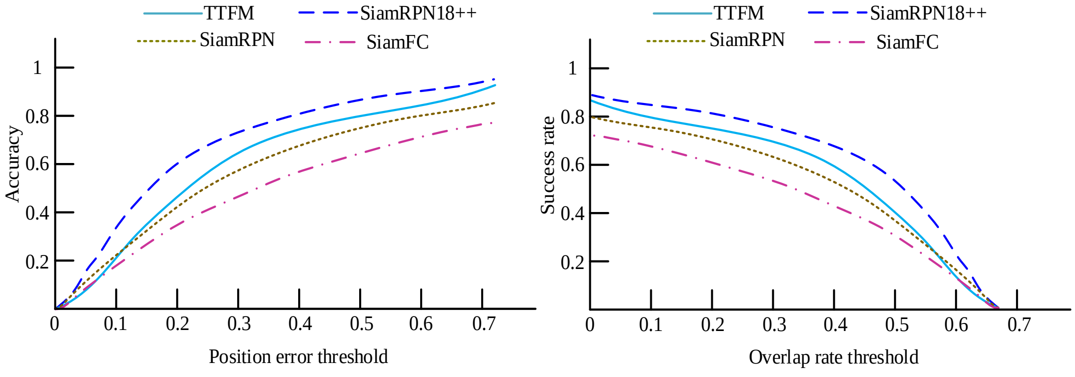
(b) The success rate changes of various algorithms on the VOT dataset

(c) The accuracy changes of various algorithms on the IMIAGE dataset

(d) The Success Rate Changes of Various Algorithms in the IMIAGE Dataset

Fig. 8. Success rate and accuracy on the VOT dataset.

From Fig. 8, the TTFM algorithm had excellent performance on the VOT dataset. Compared with the other three methods, the TTFM achieved obvious improvements in overall success rate and accuracy. Specifically, on the VOT dataset, the overall success rate of the TTFM algorithm reached 0.504, with an increase of 0.6% compared with the average success rate of other algorithms. Similarly, its accuracy also reached 0.642, with an increase of 0.9% compared with the average accuracy of other algorithms. However, the four algorithms do not change much in the IMIAGE dataset dataset, indicating that these four algorithms have good generalization ability. The outstanding performance of the TTFM algorithm on the VOT dataset is mainly attributed to the feature extraction and fusion strategy proposed in the research. The algorithm significantly enhances the ability to extract useful features by introducing attention mechanisms, which is particularly important in complex scenarios. Secondly, the algorithm utilizes multi-layer linear fusion technology to integrate feature information from different levels, which not only enhances the richness of features, but also improves the accuracy and robustness in identifying fatigue states. To further validate the generalization ability of these algorithms, additional tests are conducted on the TrackingNet dataset, keeping hyper-parameters such as learning rate unchanged. Fig. 9 presents the experimental results.



(a) Accuracy

(b) Success rate

Fig. 9. Success rate and accuracy on the TrackingNet dataset.

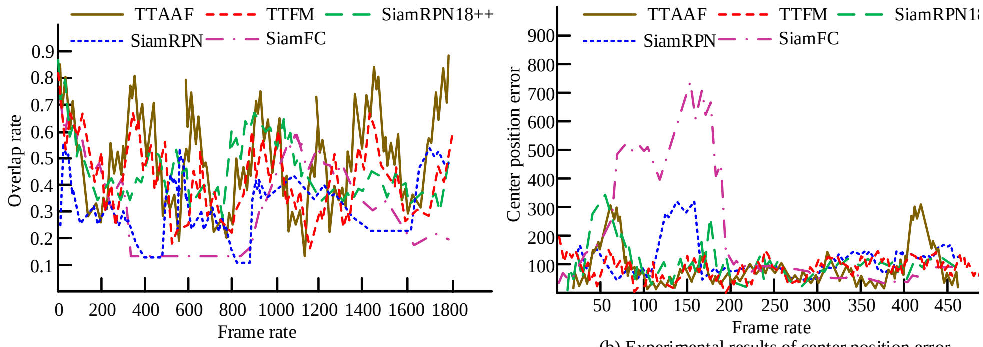
From Fig. 9, the TTFM performed better than the other three methods. For accuracy, the TTFM algorithm performed best when the position error threshold was low, ultimately converging to 0.830, while SiamFC performed better at higher position error thresholds. In the success rate, the TTFM algorithm performed well when the overlap rate threshold was high, while SiamFC performed well at lower overlap rate thresholds. The success rate and accuracy of the TTFM algorithm were 0.542 and 0.763, which were not significantly different from their performance in the VOT dataset. Overall, the TTFM algorithm performed the best on the TrackingNet dataset, especially for data containing occlusion attributes, averaging 0.4% and 3.3% higher than other algorithms. Compared with the VOT dataset, the performance of these three methods varied significantly. TTFM demonstrates superior tracking performance in complex environments. Therefore, the TTFM has better tracking performance and generalization ability, which is suitable for various complex tracking scenarios.

B. Analysis of Twin Network Visual Tracking Algorithm Integrating Adaptive Anchor Box Generation

From the analysis of the twin network visual tracking on the basis of attention mechanism, the tracking algorithm for preliminary components has good performance. Based on this, the study further optimizes it by introducing adaptive anchor box generation. To verify the repeatability and center position error of the twin network visual tracking algorithm that integrates adaptive anchor box generation, comparative analysis experiments are carried out on the VOT dataset, as presented in Fig. 10.

From Fig. 10, the improved algorithm had the best overlap rate. The improved algorithm performed well in terms of overlap rate, with an average overlap rate increase of 0.126 and an average center position error reduction of 23.450 pixels. The excellent performance of the improved algorithm in terms of overlap rate is mainly attributed to its innovative structure and mechanism. Firstly, the algorithm enhances the extraction of key behavioral and physiological signal features, such as blink frequency and head posture, by introducing attention mechanisms. These features are crucial for determining whether the driver is fatigued. Secondly, the application of multi-layer linear fusion effectively integrates feature information from different levels, enhancing the richness of features and the robustness of algorithms. To verify the tracking performance, comparative experiments are performed on the VOT dataset, as shown in Fig. 11.

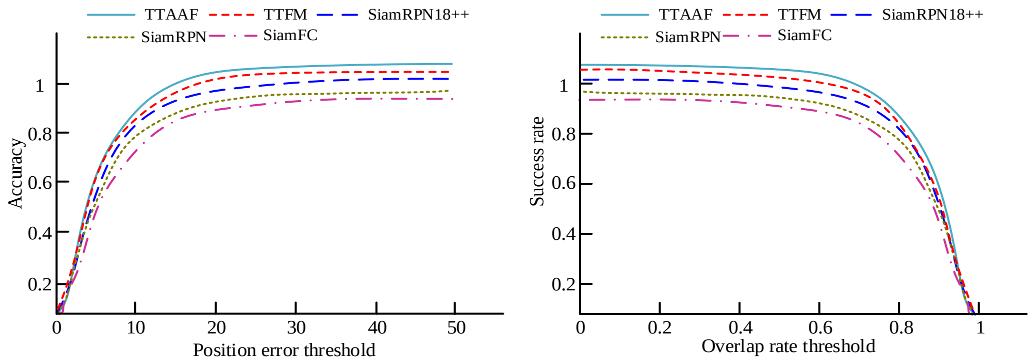
From Fig. 11, the TTAAF algorithm performed best when the position error threshold was low, while SiamFC performed better when the position error threshold was high. In the success rate, the TTAAF algorithm performed well when the overlap rate threshold was high, while SiamFC performd well at lower overlap rate thresholds. Overall, the TTAAF algorithm performed the best on the VOT dataset, with an overall success rate and accuracy of 0.831 and 0.862, respectively, which were on average 0.3% and 0.5% higher than other algorithms. Especially for data containing occlusion attributes, the success rate and accuracy of the TTAAF algorithm were 0.642 and 0.753, which were on average 0.2% and 0.4% higher than other algorithms. Therefore, the TTAAF has better tracking performance and generalization ability, which is suitable for various complex tracking scenarios. In addition, the results of the four algorithms on the TrackingNet dataset are shown in Fig. 12.



(a) Overlap rate experimental results

(b) Experimental results of center position error

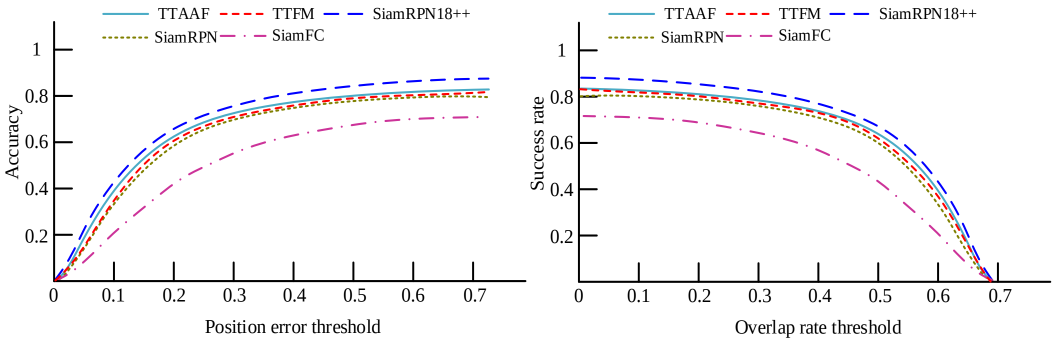
Fig. 10. Comparison analysis curve of overlap rate and center position.



(a) Accuracy

(b) Success rate

Fig. 11. Success rate and accuracy on the VOT dataset.



(a) Accuracy

(b) Success rate

Fig. 12. Success rate and accuracy graph on the TrackingNet dataset.

Fig. 12 is a comparison chart of the accuracy and success rates of four algorithms on the tracking network dataset. The horizontal axis represents the position error threshold and overlap threshold, and the vertical axis represents the accuracy and success rate. The blue dashed line represents TTAAE, the red solid line represents TTFM, the green dashed line represents SiamRPN++, and the purple dashed line represents SiamFC. From Fig. 12 (a), it can be observed that as the position error threshold gradually increases, the four curves gradually decrease. Within a smaller range of position error thresholds, the TTAAF algorithm has the highest accuracy, followed by TTFM, then SiamRPN++, and finally SiamFC; When the position error threshold is greater than 0.3, the accuracy of SiamFC exceeds SiamRPN++. Fig. 12(b) shows the trend of success rate with respect to the overlap rate threshold. It can be seen that as the overlap rate threshold increases, the four curves show an upward and then downward trend. Within a larger range of overlap rate thresholds, the TTAAF algorithm has the highest success rate, followed by TTFM, then SiamRPN++, and finally SiamFC; when the overlap rate threshold is less than 0.3, the success rate of SiamFC exceeds that of SiamRPN++. Combining the two graphs, it can be concluded that the TTAAF algorithm has the best overall performance on the tracking network dataset, with an accuracy and success rate of 0.932 and 0.962, respectively, which are 0.3% and 0.5% higher than other algorithms. Especially when dealing with occlusion attribute data, the performance of TTAAF algorithm is particularly outstanding, with accuracy and success rates of 0.642 and 0.753, respectively, which are 0.2% and 0.4% higher than other algorithms. This indicates that the TTAAF algorithm has good robustness in handling occlusion attribute data. Finally, it was applied to the actual target tracking process, and the experimental results are shown in Fig. 13.

From the experimental results in Fig. 13, it can be seen that there are differences in the performance of the target tracking algorithm in different scenarios. In scenes marked as' Normal ', the algorithm can successfully track targets, even in situations where there is occlusion or complex background between targets, as shown in the images in the upper and lower left corners, the algorithm can still accurately identify and track targets. However, in the scenario marked as' Tracking failed ', the tracking ability of the algorithm is challenged. The image in the upper right corner shows that when the target is partially obscured by the sofa, the algorithm cannot continue tracking the target, which may be due to confusion between the target features and background features, resulting in tracking loss. In the image in the lower right corner, the target also failed to track after entering the room due to changes in lighting and increased background complexity. This indicates that the algorithm may have limitations when dealing with lighting changes and complex backgrounds. In addition, the image on the right side of the middle shows that the algorithm can maintain tracking even when the target is moving rapidly, indicating that the algorithm has a certain robustness to dynamic scenes.

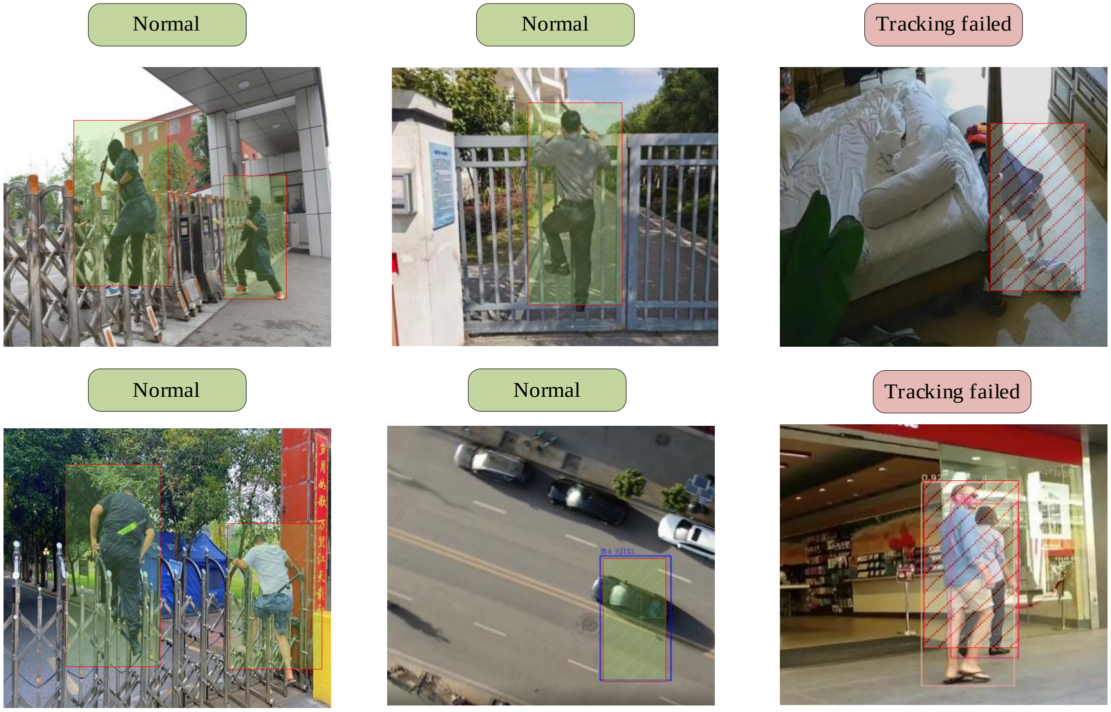


Fig. 13. Algorithm application effect interface diagram.

C. *Analysis of Twin Network Visual Tracking Algorithm Integrating Adaptive Anchor Box Generation*

A dual network visual tracking algorithm based on attention mechanism has been proposed to solve the problems in traditional fatigue driving detection, such as poor comfort, sensitivity to external factors, low detection accuracy, and poor real-time performance. By introducing attention mechanisms, algorithms can more effectively extract key features and improve target tracking performance in complex environments.

In the research of fatigue driving detection technology, this study proposes a visual tracking algorithm based on dual networks, which combines attention mechanism and adaptive anchor box generation. Compared with other methods in literature, it shows significant performance improvement. For example, the twin adaptive learning network proposed by Xin et al. performed well in GOT-10k and LaSOT benchmark tests, but the method proposed in this study can significantly reduce center position errors and improve tracking accuracy and robustness when dealing with complex background interference such as lighting changes, occlusion, and size changes [2]. The effective feature recognizer proposed by Jun W et al. optimized the representation ability, but there is still room for improvement in real-time performance [4]. In contrast, the overall success rate and accuracy of the algorithm in this study on the VOT dataset reached 0.831 and 0.862, respectively, which were 0.3% and 0.5% higher than other algorithms on average. Especially when processing occlusion attribute data, the success rate and accuracy were 0.2% and 0.4% higher than other algorithms on average, demonstrating better tracking performance and generalization ability.

The reasons and mechanisms for these results are as follows: the algorithm introduces attention mechanism, which can effectively filter out feature information that is not conducive to target tracking, retain key information, and improve feature extraction efficiency. Meanwhile, multi-layer linear fusion integrates feature information from different levels, enhancing feature richness and algorithm robustness. In addition, adaptive anchor box generation reduces manual intervention and improves target tracking performance.

In summary, the dual network visual tracking algorithm based on attention mechanism proposed in this study has made significant progress in improving the accuracy and robustness of fatigue driving detection, providing an effective technical means for fatigue driving recognition in practical applications. Future research can further optimize algorithms to improve their adaptability and real-time performance in different driving scenarios, in order to better serve road safety and public health.

IV. CONCLUSION

Aiming at the challenges in fatigue driving detection, a visual tracking algorithm based on twin networks was proposed. Firstly, the TTFM algorithm was proposed. Then the TTAAF algorithm was optimized on the basis of the TTFM. The two algorithms proposed in the study performed well on multiple benchmark datasets, effectively improving the accuracy and robustness of target tracking. Specifically, the TTFM optimized the feature extraction process by introducing channel and spatial attention mechanisms, significantly enhancing the algorithm's ability to extract useful features. On the VOT dataset, the overall success rate and accuracy of the TTFM reached 0.504 and 0.642, with an average improvement of 0.6% and 0.9% compared with other algorithms. On the TrackingNet dataset, the success rate and accuracy of the TTFM algorithm were 0.542 and 0.763, especially for data containing occlusion attributes, which were on average 0.4% and 3.3% higher than other algorithms. Compared with TTFM, the TTAAF algorithm reduced manual intervention by adaptively generating anchor boxes, allowing the model to automatically learn and generate candidate boxes based on the semantic information of the image itself. On the VOT dataset, the overall success rate and accuracy of the TTAAF algorithm were 0.831 and 0.862, which were on average 0.3% and 0.5% higher than other algorithms. On the TrackingNet dataset, the TTAAF algorithm performed equally well, with an overall success rate and accuracy of 0.932 and 0.962, respectively, which were on average 0.3% and 0.5% higher than other algorithms. In summary, both algorithms proposed in the study have effectively improved the performance of target tracking, especially robustness in complex environments. This provides strong technical support for practical applications such as fatigue driving detection. Although significant progress has been made in this study, there are still some shortcomings. For example, the algorithms had high computational complexity and requires further optimization to improve real-time performance. Meanwhile, the robustness of the algorithm still needs to be improved for target tracking under extreme lighting conditions.

REFERENCES

[1] Xiaofeng B, Chenggang G. SiamMaskAttn: inverted residual attention block fusing multi-scale feature information for multitask visual object tracking networks. Signal, Image and Video Processing, 2023, 18 (2): 1305-1316.

[2] Bin P, Ke X, Ze'an L. Siamese refine polar mask prediction network for visual tracking. Signal, Image and Video Processing, 2023, 18 (1): 923-933.

[3] Xin L, Fusheng L, Wanqi Y. Siamada: visual tracking based on Siamese adaptive learning network. Neural Computing and Applications, 2024, 36 (14): 7639-7656.

[4] Jun W, Peng Y, Wenhui Y. Exploiting multi-scale hierarchical feature representation for visual tracking. Complex & Intelligent Systems, 2024, 10 (3): 3617-3632.

[5] Jianbin D. Design of smart operating table based on HCI and virtual visual tracking technology. International Journal on Interactive Design and Manufacturing (IJIDeM), 2023, 18 (2): 1019-1031.

[6] Di Y, Gu G, Xiu S. Self-supervised discriminative model prediction for visual tracking. Neural Computing and Applications, 2023, 36 (10): 5153-5164.

[7] Liu K, Liu L, Yang Sl. Spatial feature embedding for robust visual object tracking. IET Computer Vision, 2023, 18 (4): 540-556.

[8] Han W, Bo Z, Guizhong L. Refiner: a general object position refinement algorithm for visual tracking. Neural Computing and Applications, 2023, 36 (8): 3967-3981.

[9] Shaochuan Z, Tianyang X, Jun X W. A Spatio-Temporal Robust Tracker with Spatial-Channel Transformer and Jitter Suppression. International Journal of Computer Vision, 2023, 132 (5): 1645-1658.

[10] Chloe C, Isaac A M. Cognitive-perceptual traits associated with autism and schizotypy influence use of physics during predictive visual tracking. The European journal of neuroscience, 2023, 58 (10): 4236-4254.

[11] Zhang J, Yang X, Wang W. Cross-entropy-based adaptive fuzzy control for visual tracking of road cracks with unmanned mobile robot. Computer-Aided Civil and Infrastructure Engineering, 2023, 39 (6): 891-910.

[12] Yijin Y, Xiaodong G. Learning rich feature representation and aggregation for accurate visual tracking. Applied Intelligence, 2023, 53 (23): 28114-28132.

[13] Mengquan L, Xuedong W, Siming T. Visual tracking via confidence template updating spatial-temporal regularized correlation filters. Multimedia Tools and Applications, 2023, 83 (12): 37053-37072.

[14] Yuping Z, Zepeng Y, Bo M. Structural-appearance information fusion for visual tracking. The Visual Computer, 2023, 40 (5): 3103-3117.

[15] Liu Y, Jinchun D, Jun S. A new passive vision weld seam tracking method for FSW based on K-means. The International Journal of Advanced Manufacturing Technology, 2023, 128 (7-8): 3283-3295.

[16] An Z, Yi Z. Evota: an enhanced visual object tracking network with attention mechanism. Multimedia Tools and Applications, 2023, 83 (8): 24939-24960.

[17] Ning D, Kazuya T, Wenhui Jl. Estimation of control area in badminton doubles with pose information from top and back view drone videos. Multimedia Tools and Applications, 2023, 83 (8): 24777-24793.

[18] Amin S N, Shivakumara P, Jun T X, et al. An Augmented Reality-Based Approach for Designing Interactive Food Menu of Restaurant Using Android[C]//Artificial Intelligence and Applications. 2023, 1(1): 26-34.

[19] Nsugbe E. Toward a Self-Supervised Architecture for Semen Quality Prediction Using Environmental and Lifestyle Factors. Artificial Intelligence and Applications. 2023, 1(1): 35-42.

[20] Choudhuri S, Adeniye S, Sen A. Distribution Alignment Using Complement Entropy Objective and Adaptive Consensus-Based Label Refinement for Partial Domain Adaptation[C]//Artificial Intelligence and Applications. 2023, 1(1): 43-51.